Statistical learning in high dimensions : a rigorous statistical physics approach

Cédric Gerbelot

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Statistical learning in high dimensions: a rigorous statistical physics approach

Soutenue par
Cédric GERBELOT
Le 25 Août 2022

École doctorale n°564
Physique en Île-de-France

Spécialité
Physique Théorique

Composition du jury :

Laurent MASSOULIÉ INRIA Paris	*Président*
Arian MALEKI Columbia University	*Rapporteur*
Afonso BANDEIRA ETH Zürich	*Rapporteur*
Gérard BEN AROUS New York University	*Examinateur*
Lenka ZDEBOROVA EPFL	*Examinateur*
Marc LELARGE ENS & INRIA Paris	*Co-directeur*
Florent KRZAKALA EPFL	*Directeur de thèse*

ENS | PSL

He deals the cards as a meditation
And those he plays never suspect
He doesn't play for the money he wins
He don't play for respect

He deals the cards to find the answer
The sacred geometry of chance
The hidden law of a probable outcome
The numbers lead a dance

Shape of my heart, Sting - 1993

Remerciements

Je tiens tout d'abord à remercier les rapporteurs Arian Maleki et Afonso Bandeira d'avoir relu et commenté mon manuscrit, Laurent Massoulié pour avoir présidé le jury ainsi que les autres membres du jury Gérard Ben Arous, Lenka Zdeborovà, Florent Krzakala et Marc Lelarge pour leur attention et leurs questions.

Mes remerciements se dirigent ensuite évidemment vers mon directeur de thèse Florent, dont je me suis rapproché sans le savoir lors d'un stage d'été à la fin de ma première année à l'ESPCI. J'avais demandé à Elie Raphaël, alors directeur de l'unité de recherche *Gulliver*, si il pouvait m'accueillir au sein de son équipe pour quelques semaines, ayant apprécié son cours de méthodes mathématiques. Ce stage, bien que bref, a été pour moi une expérience déterminante qui m'a motivé à poursuivre dans la voie de la recherche. J'ai été inspiré par l'enthousiasme et la compétence de mes encadrants, Elie Raphaël et Thomas Salez, que je remercie chaleureusement. Devenir chercheur en mathématiques ou en physique théorique me paraissait jusqu'alors impossible. Au moment de postuler en thèse plusieurs années plus tard, Elie m'a recommandé le groupe de Florent Krzakala, un ancien collègue et ami travaillant dans les domaines qui m'intéressaient. J'ai retrouvé dans ce groupe le mélange de compétence et d'enthousiasme qui m'avait impressionné au *Gulliver*. Je garde d'excellents souvenirs de discussions aussi bien sur des aspects techniques que génériques, qui ont considérablement développé ma compréhension des systèmes désordonnés en grandes dimensions et des outils probabilistes reliés. Florent, j'ai été inspiré par ta capacité à réduire des concepts parfois très techniques et nécessitant des calculs parfois très lourds à des raisonnement simples et intuitifs. Il me tiendra à coeur de préserver cette approche dans mes travaux futurs. Je te remercie tout autant pour ton humour, ta bonne humeur à toute épreuve et ton soutien moral infaillible, dignes d'un entraîneur olympique (tu m'as dis toi-même qu'une carrière académique compétitive, c'est comme un parcours d'athlète de haut niveau !). Merci à Lenka de m'avoir montré un exemple de chercheure intrépide et déterminée, et de m'avoir presque co-encadré durant la deuxième partie de ma thèse. Enfin, merci à Marc Lelarge de m'avoir accueilli dans les locaux de l'équipe DYOGENE à l'INRIA durant ma dernière année, ainsi qu'à Jean Barbier pour m'avoir invité à Trieste malgré un pandémie mondiale qui limitait fortement les possibilités de déplacement.

Je tiens à présent à remercier tous les amis et collègues que j'ai eu le plaisir de cotoyer durant ces trois années. Je suis particulièrement reconnaissant à Alia qui m'a accompagné dans mes premiers pas pour prouver la formule de Kabashima, avec de nombreuses sessions de calculs au tableau et de discussions sur des thématiques allant de la psychologie jusqu'à la géopolitique. Un grand merci pour ta patience et ta bienveillance. Merci à Francesca, Maria et Ruben avec qui j'ai partagé de nombreux déjeuners, cafés et moments au laboratoire. Merci aussi aux autres membres de l'équipe SPHINX qui étaient présents lors de mon arrivée : Marylou, Benjamin, Antoine et Sebastian. Un merci tout particulier à Bruno et Gabriele, virtuoses de la méthode des répliques, pour des collabo-

rations fructueuses lors desquelles j'ai beaucoup appris, à Paris puis à Lausanne. Merci aussi à tous les membres des groupes IdePHICS et SPOC à l'EPFL, pour la bonne ambiance générale et des discussions intéressantes autour d'excellents cafés savamment préparés par Bruno : Hugo, Damien, Luca, ... et tous les autres.

Je ne peux évidemment pas oublier trois compères de l'ESPCI qui m'ont accompagné pendant toutes ces années : Benoît, Hugues et Thomas. Les amis, merci infiniment pour vos encouragements, vos bons conseils pendant les moments difficiles et pour de nombreuses discussions métaphysiques dans des états d'ébriété plus ou moins avancés. Un grand merci aussi à Romain, Victor, Juliette et les petits jeunes (rires) pour avoir supporté mon humeur parfois difficile et pour de nombreuses bières partagées dans la bonne humeur. Merci enfin à Guillaume, un ami du collège que j'ai eu le plaisir de revoir quand les mesures de confinement se sont allégées.

Je remercie enfin ma famille, que j'aime beaucoup : mon frère Jean-Marc, mes soeurs Auriane et Raphaëlle, et mes parents Nathalie et Jean-Christophe, qui ont posé les bases des mes connaissances en mathématiques et joué un rôle déterminant dans mon parcours.

Résumé en français

Ce manuscrit contient le travail que j'ai effectué pendant mon doctorat à l'Ecole Normale Supérieure de Paris, principalement sous la direction du Pr.Florent Krzakala. Le coeur de ce texte est constitué d'une introduction et de trois parties, qui proposent une approche analytique rigoureuse à la théorie de l'apprentissage automatique supervisé en grande dimension sous l'hypothèse de données aléatoires. Ce résumé, dont la version en anglais, plus complète, peut être trouvée après la table des matières sous la forme d'un avant-propos, suppose que le lecteur possède des notions de probabilités en grandes dimensions, de théorie des verres de spins ainsi que de l'apprentissage statistique supervisé. Le lecteur ne possédant pas ces notions peut se référer à l'introduction (Chapitre 1, en anglais), puis revenir à ce résumé.

Organisation du manuscrit et aperçu des contributions

Le chapitre 1 propose une introduction courte à l'apprentissage automatique ainsi qu'à la théorie de l'apprentissage statistique, qui permet de mieux motiver le besoin d'approches basées sur les probabilités en grandes dimensions et la physique mathématique, ainsi que de proposer un point de vue cohérent pour ces thèmes. Nous donnons ensuite un aperçu de la physique statistique des milieux désordonnés ainsi que des outils analytiques non rigoureux qui sont utilisés dans ce domaine, tel que la méthode de la cavité et la méthode des répliques, ou bien des relaxations asymptotiques de l'algorithme de propagation de convictions. Ceci nous amène naturellement aux pendants rigoureux de ces méthodes, qui peuvent être globalement comprises comme des procédures de découplage de mesures de probabilités compliquées, de manière à les décomposer en des produits de mesures plus simples pour lesquelles les résultats de concentration sont plus faciles à établir et, d'un point de vue pratique, qui peuvent être simulées en un temps et avec des ressources raisonnables. Après avoir fourni une description brève des résultats existants sur des modèles de données i.i.d. Gaussiennes, nous soulignons les difficultés principales qui apparaissent lorsque l'on tente de pousser la théorie plus proche des scénarios réalistes, des algorithmes qui constituent l'état de l'art, et des résultats correspondant venant de la physique statistique :

- les données structurées mènent naturellement à des problèmes non séparables, là où de nombreuses preuves existentes ne sont valables que pour des problèmes séparables,
- les algorithmes d'agggrégation de prédicteurs, machines à comités et problèmes multiclasses nécessitent des méthodes de preuves qui donnent les distributions asymptotiques jointes d'un nombre fini d'estimateurs, plutôt que d'un seul,
- tous les problèmes sont à température zéro, aus ens de la physique statistique, ce qui empêche l'utilisation d'identités simplificatrices issues de la Bayes-optimalité tel que l'identité de Nishimori,

- les prédictions existantes issues de la physique statistique montrent que les résultats d'asymptotiques exactes pour les algorithmes de passage de messages approximés peuvent être obtenus pour des modèles bien plus complexes que les modèles linéaires généralisés, en particulier pour des modèles multicouches à poids aléatoires ou d'a priori génératifs,

- sous réserve que l'on puisse obtenir des prédictions asymptotiquement exactes sur les modèles présentant des données structurées, à quel point ces résultats peuvent être utilisés sur des données réelles ?

La section 1.7 présente ensuite un aperçu des outils mathématiques principaux qui seront utilisés dans ce manuscrit, notamment les inégalités de comparaison Gaussienne et les méthodes de conditionnement itératif Gaussien dans le contexte de l'étude des algorithmes de passage de message approximé (AMP). Nous illustrons aussi ces techniques sur des problèmes simples, de manière à fournir une intuition claire sur les résultats qui sont présentés pour des modèles plus complexes dans les chapitres qui suivent. Les raisons principales qui sous-tendent le succès des approches proposées en vue des objectifs présentés ci-avant sont les suivantes :

- les modèles non-séparables peuvent être traités en utilisant des inégalités de comparaison Gaussiennes dans le cas convexe ainsi qu'une décomposition du prolème appropriée à l'aide de multilicateurs de Lagrange. Cette approche échoue, en revanche, pour les ensembles d'estimateurs,

- les itérations AMP peuvent être étudiées rigoureusement avec à la fois des effets non-séparables et des estimateurs matriciels, mais pour caractériser une solution précise, il faut réaliser un contrôle de la trajectoire de l'itération vers cette solution,

- les itérations AMP peuvent être construites et leurs trajectoires contrôlées précisément dans le cas convexe de manière systématique,

- en ce qui concerne les problèmes de dynamique, le schéma de conditionnement itératif au coeur des preuves reliées aux algorithmes AMP peut être étendus aux cas multicouches et aux problèmes composites impliquant plusieurs matrices aléatoires, des perturbations de rang faibles, entre autres,

- des modèles de référence exactement solvables (au sens de la physique statistique, voir le chapitre 1) dont les courbes d'apprentissage correspondent exactement à des scenarios réalistes peuvent être définis à partir de données Gaussiennes corrélées.

Ce manuscrit s'articule autour de ces idées, commençant par les résultats les plus généraux, avant de les utiliser dans des cas plus spécifiques correspondant à une famille de problèmes convexes qui définissent des estimateurs utilisés en apprentissage supervisé.

A cet égard, la Partie I est focalisée sur la dynamique en grandes dimensions des algorithmes AMP pour une classe de modèles large ainsi que sur l'application des idées de conditionnement Gaussien itératif pour l'étude des algorithmes de descentes de gradients stochastiques. Nous commençons, dans les chapitres 2 et 3, avec des résultats publiés dans l'article

[110] C. Gerbelot and R. Berthier, *Graph-based approximate message passing iterations*, arXiv preprint arXiv:2109.11905, (2021)

actuellement en revue. Ce travail étend les preuves d'équations d'évolution d'état (*state evolution (SE) equations*) de [41, 28, 42] à des itérations AMP composites en les indexant sur un graphe orienté pouvant être étendu arbitrairement pourvu que des conditions structurelles simples soient vérifiées. Nous prouvons que toute itération AMP pouvant être indéxée sur un tel graphe admet des équations SE rigoureuses, et nous donnons la forme de ces équations. Le graphe orienté peut être composé arbitrairement pour fournir de nouvelles itérations AMP ainsi que leurs équations SE, atteignant une flexibilité proche de celle des approches heuristiques basées sur des équations de type Thouless-Anderson-Palmer (TAP) pour les problèmes multicouhces, notamment [194, 188, 13], qui sont rendues rigoureuses par notre résultat. Nous montrons aussi comment des extensions rencontrées souvent dans les problèmes d'inférence, comme les modèles plantés, des matrices spikées ou encore du couplage spatial, peuvent être inclues dans notre approche.

Une première application de ces résultats est proposée dans les chapitres 4 et 5, où nous étudions les dynamiques d'algorithmes AMP multicouches (MLAMP), initialement proposés dans [188], lorsque les matrices Gaussiennes denses de mélange sont remplacées par des matrices de convolutions aléatoires. Ces chapitres sont basés sur la publication, acceptée dans *Advances in Neural Information Processing Systems (NeurIPS) 2022*,

[70] M. Daniels, C. Gerbelot, F. Krzakala, and L. Zdeborová, *Multi-layer state evolution under random convolutional design*, arXiv preprint arXiv:2205.13503, (2022)

La méthode de preuve repose sur l'incorporation de l'itération AMP avec les matrices convolutionnelles au sein d'une itération plus large possédant des matrices denses pour laquelle la preuve rigoureuse des équations d'évolution d'états peut être conduite. La structure convolutionnelle est conservée en l'encodant dans des non-linéaritées circulantes de l'itération plus large, maintenant définie avec des variables à valeurs matricielles.

Dans le chapitre 6, nous continuons la discussion démarrée dans la section 1.7 de l'introduction qui présente la dynamique en grandes dimensions des méthodes de descente de gradient. Nous montrons que le conditionnement itératif Gaussien utilisé pour les preuves d'AMP de notre contribution [110] peut être utilisé pour prouver les équations de théorie dynamique à champ moyen (*dynamical mean field theory (DMFT)*), adaptées à la descente de gradient stochastique dans [198], et récemment prouvées dans un cadre plus restreint en utilisant une itération AMP à mémoire dans [56]. La contribution principale de ce travail est de montrer que l'incorporation implicite de la descente de gradient stochastique dans une itération de type AMP peut être évitée, fournissant ainsi une preuve complètement explicite dans laquelle l'apparition des noyaux de corrélations à deux temps de la dynamique DMFT de fait en suivant un raisonnement de récurrence. Nos résultats bénéficient aussi de la généralité des lemmes intermédiaires prouvés dans notre contribution précédente [110]. Ce chapitre est basé sur la publication suivante, actuellement en revue,

[111] C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborova, *Rigorous dynamical mean field theory for stochastic gradient descent methods*, arXiv preprint arXiv:2210.06591, (2022)

Nous avançons alors vers la Partie II qui concerne des modèles exactement solvables pour l'apprentissage supervisé avec des transformations de prédicteurs réalistes ainsi que des modèles de données structurées. Nous commençons par l'analyse d'un modèle convexe linéaire généralisé avec une matrice de design présentant une structure corrélée par blocs, dans les chapitres 7 et 8, basés sur les résultats proposés dans la publication

[176] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová, *Learning curves of generic features maps for realistic datasets with a teacher-student model*, Advances in Neural Information Processing Systems, 34 (2021), pp. 18137–18151

La structure corrélée par blocs de la matrice de données représente des transformations de prédicteurs différentes pour le modèle génératif planté et le modèle d'apprentissage. La méthode de preuve est basée sur le cadre des inégalités de comparaisons Gaussiennes proposé dans [281, 204, 57], et obtient des formules qui correspondent aux prédictions effectuées à l'aide de la méthode des répliques. Nous montrons alors empiriquement que, pour une classe large de transformation de prédicteurs, le modèle Gaussien synthétique dont les matrices de covariance sont les mêmes que les matrices de covariance empiriques du jeu de données réel capture exactement les courbes d'apprentissage réelles pour les tâches de regressions, ce qui nous amène à la conjecture dite "d'équivalence Gaussienne" (*Gaussian equivalence conjecture*) pour ces modèles. La conjecture ne semble pas tenir aussi bien pour les problèmes de classification, ce qui motive le besoin d'un modèle de référence supplémentaire.

Nous nous tournons donc vers l'étude de problèmes de classification multiclasse dans les chapitres 9 and 10, que nous modélisons par l'apprentissage d'un nombre fini d'hyperplans séparateurs d'une mixture de Gaussiennes arbitraire en utilisant un modèle linéaire généralisé convexe. Ces résultats ont été publiés dans l'article

[178] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, *Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions*, Advances in Neural Information Processing Systems, 34 (2021), pp. 10144–10157

La méthode de preuve utilise une trajectoire convergente [29, 82] d'une itération AMP construite spécifiquement pour la résolution de ce problème. Cette construction repose sur une représentation de la classification de la mixture de Gaussiennes corrélées comme un problème d'optimisation couplé en espace [154, 135] sur une variable matricielle, et présentant des effets non-séparables. Les équations rigoureuses d'évolution d'état de cette itération AMP sont établis avec les résultats de note contribution précédente [110]. Les résultats rigoureux sont une fois de plus en accord avec les prédictions obtenues par des calculs de répliques. Les simulations montrent alors que, pour des jeux de données simples comme MNIST ou Fashion-MNIST, les courbes d'apprentissages exactes pour des tâches de classification peuvent être obtenues exactement en utilisant un modèle synthétique de mixture de Gaussiennes dont les moyennes et les covariances sont estimées empiriquement à partir du jeu de données réel. Pour des données plus structurées ou des tâches plus complexes, le nombre de composants de la mixture de Gaussienne peut être augmenté pour amener la prédiction proposée par les formules obtenues pour le modèle synthétique plus proche de la courbe réelle.

Motivés par l'importance des méthodes d'agrégation d'estimateurs en apprentissage automatique ainsi que des informations que ces méthodes peuvent donner sur les réseaux de neurones [72], nous

nous tournons aux chapitres 11 and 12 vers l'apprentissage d'ensembles de prédicteurs, chacun desquels est défini par un modèle convex linéaire généralisé avec un modèle de données Gaussiennes corrélées par blocs similaire à celui proposé précédemment dans notre contribution [176]. Ces résultats sont basés sur la publication

[177] B. Loureiro, C. Gerbelot, M. Refinetti, G. Sicuro, and F. Krzakala, *Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension*, International Conference on Machine Learning (ICML), (2022)

La preuve repose sur une itération AMP à variables matricielles et à non-linéarités non-séparables pour laquelle nous utilisons le même contrôle de trajectoire que dans nos études précédentes des problèmes de classification multitâches [178], et où la validité des équations d'évolution d'état est guarantie par les résultats de notre contribution [110]. Ici encore, nous observons que les prédictions obtenues par les méthodes de répliques sont correctes. Nous utilisons ces formules pour étudier les effets de l'agrégation de prédicteurs, notamment en terme de réduction de variance et de régularisation implicite, sur des tâches usuelles comme la régression logistique ou l'apprentissage avec des caractéristiques aléatoires [238], ainsi que l'alignement des prédicteurs.

Enfin, la partie III présente des résultats publiés dans les articles

[108] C. Gerbelot, A. Abbara, and F. Krzakala, *Asymptotic errors for high-dimensional convex penalized linear regression beyond gaussian matrices*, in Conference on Learning Theory, PMLR, 2020, pp. 1682–1713
[109] C. Gerbelot, A. Abbara, and F. Krzakala, *Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula)*, arXiv preprint arXiv:2006.06581, (2020)

L'apparition du second dans *IEEE Transactions on Information Theory* est prévue. Ces résultats sont des preuves de formules de répliques qui ont été obtenues par Y. Kabashima [138, 140, 277] dans le cas de modèles convexes linéaires généralisés pour lesquels la matrices de données est invariante par rotations à gauche et à droite, et dont les valeurs singulières sont issues i.i.d. d'une distribution arbitraire à support compact. Le résultat du deuxième article [109] est plus général que celui du premier [108], qui n'est donc pas reproduit dans cette thèse. Le lecteur intéressé peut néanmoins consulter l'article [108] pour des formules plus simples ainsi que des exemples d'applications supplémentaires, notamment concernant l'acquisition compressée. La méthode de preuve est basée sur la construction de trajectoires convergentes de l'algorithme de passage de message approximé vectoriel à deux couches (2-MLVAMP) [242, 97], qui propose des équations d'évolution d'état rigoureuses pour des itérations optimisant des modèles linéaires généralisés convexes dont les matrices de données sont invariantes par rotations à gauche et à droite, et dont les valeurs singulières sont issues i.i.d. d'une distribution arbitraire à support compact. Etant donnée la structure des algorithmes de passage de message approximé vectoriel, l'étude des trajectoires est différente de celles menées précedemment pour des itération d'AMP classiques (à matrices denses Gaussiennes ou sub-Gaussiennes) : nous reformulons l'algorithme de passage de message approximé vectoriel à deux couches en un système dynamique, pour lequel nous déterminons une fonction de Lyapunov adaptée au problème, en utilisant des résultats de théorie du controle optimal, et plus particulièrement des système dynamiques sous contraintes intégrales quadratiques [166]. Nos résultats prouvent des guaranties de convergence

algorithmique pour des problèmes suffisemment fortement convexes, et ces guaranties ne dépendent pas de la haute dimensionalité du problème. Sous une hypothèse de concentration, nous montrons qu'un prolongement analytique du résultat peut être mené afin d'étendre la validité de la formule de réplique à tout problème convexe. Nous proposons des simulations pour la formule de répliques prouvées sur une grande variété de problèmes ainsi que pour les guaranties de convergence algorithmiques de 2-MLVAMP.

Nous proposons finalement une conclusion ainsi que des persepctives de travaux futurs, notamment en ce qui concerne les problèmes non-convexes, les résultats d'universalités ainsi que les taux de convergence de taille finies dans le chapitre IV.

Some notations and abbreviations

$x, \mathbf{x}, \mathbf{X}$	scalar, vector, matrix
$\langle ., . \rangle$	inner product
$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	real numbers
$\mathbb{S}_d^+$	set of semi positive definite matrices
$\mathbb{S}_d^{++}$	set of positive definite matrices
$\|.\|_p$ l_p	norm
$\|.\|_F$	Frobenius norm
$\|.\|_{op}$	operator norm
$a.s.$	almost surely
$w.h.p.$	with high probability
∂f	subdifferential operator of a (convex) function f
$\text{zer}(A)$	zeroes of an operator A
$\ker(\mathbf{M})$	the null-space of a matrix $\mathbf{M}$
$\text{Tr}(\mathbf{M})$	trace of the matrix $\mathbf{M}$
$\mathbb{E}$	mathematical expectation
$\mathbb{E}_X$	expectation with respect to a single random variable X
$\mathbb{E}[X\|Y]$	conditional expectation of X given Y
$X_{\|Y}$	conditional distribution of X given Y
$\sigma(X^1, X^2, ..., X^t)$	sigma-algebra generated by the random variables $X^1, X^2, ..., X^t$
$\mu, \boldsymbol{\mu}$	Lebesgue measure, mean of a random vector
$x \sim p$	the random variable x is distributed according to p
$x \overset{d}{=} y$	the random variable x has the same distribution as y
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{a.s.}$	almost sure convergence
$\xrightarrow{Plk}$	convergence in the Plk sense (will be used for informal statements)
$X_n \overset{\text{P}}{\simeq} Y_n$	for two sequences of random variables X_n, Y_n, $X_n - Y_n \xrightarrow{P} 0$
$\mathbf{I}_d$	identity matrix of dimension d
$\mathcal{F}$	will usually denote a Hilbert space
$\mathcal{X}$	input space of a given function or operator, almost always euclidian
$\text{prox}_{\gamma f}$	proximal operator of a convex function f with parameter γ
$\mathcal{M}_{\gamma f}$	Moreau envelope of a convex function f with parameter γ
$\text{span}(\mathbf{M})$	the subspace spanned by the columns of $\mathbf{M}$
$\mathbf{P}_{\mathbf{M}}, \mathbf{P}_{\mathbf{M}}^{\perp}$	the orthogonal projector on span($\mathbf{M}$) and the orthogonal projector on its complement
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

$\mathbf{f}^t, \mathbf{F}^t$	vector valued, matrix valued functions
$\text{div}(f)$	divergence of a differentiable function f
$\mathcal{J}_f$	Jacobian of a differentiable function f
$\mathcal{H}_f$	Hessian of a twice differentiable function f
SE	state evolution
AMP	approximate message passing
GLM	generalized linear model
$MLAMP$	multilayer approximate message passing
$VAMP$	vector approximate message passing

Contents

Foreword **17**

1 Introduction **22**
1.1 Artificial intelligence . . . 22
1.2 Supervised learning . . . 23
1.2.1 Empirical risk minimization . . . 23
1.2.2 Choosing the candidate functional space . . . 24
1.2.3 Loss functions and optimization . . . 26
1.3 Statistical learning theory . . . 27
1.4 Statistical physics of disordered systems . . . 28
1.5 Statistical physics of supervised learning . . . 31
1.6 Goal of the present work and technical challenges . . . 33
1.7 Overview of the technical tools . . . 35
1.7.1 Elements of concentration of measure . . . 35
1.7.2 Elements of convex analysis . . . 38
1.7.3 Gaussian comparison inequalities . . . 39
1.7.4 Iterative Gaussian conditioning . . . 43
1.7.5 Approximate message-passing . . . 47

I High-dimensional dynamics : graph-based AMP iterations and first order methods **52**

2 Graph-based AMP iterations **53**
2.1 Graph-based AMP iterations . . . 56
2.2 State evolution for graph-based AMP iterations . . . 60
2.2.1 Main theorem . . . 60
2.2.2 Reduction of graph-based AMP iterations to the matrix-valued, non-separable symmetric case . . . 62
2.2.3 Useful extensions . . . 65
2.3 Applications to inference problems . . . 68
2.3.1 A building block: AMP for generalized linear models . . . 69
2.3.2 Multilayer generalized linear estimation . . . 69
2.3.3 Spiked matrix with generative prior . . . 70
2.3.4 An example with matrix-valued variables . . . 70
2.3.5 An example with structured random matrices . . . 70

2.3.6 An example of spatial coupling with non-separable non-linearities 71
2.4 Perspectives 71

3 Proofs for the Graph-based AMP iterations 73
3.1 Changing time indices 73
3.2 Matrix-valued symmetric AMP iterations with non-separable non-linearities 74
3.2.1 State evolution description 74
3.2.2 Application: proof of Theorem 4 76
3.3 Proof of Theorem 5 76
3.3.1 Proof outline and intermediate lemmas 77
3.3.2 Proof of Lemma 5 and Theorem 5 80
3.3.3 Proof of intermediate lemmas 82
3.4 Low-rank perturbations and projections 94
3.4.1 Additive low-rank perturbation 94
3.4.2 Dependence on an additional linear observation 97
3.5 Useful definitions and probability lemmas 104

4 Multi-layer State Evolution Under Random Convolutional Design 109
4.1 Definition of the problem 111
4.1.1 Multi-channel Convolutional Matrices 111
4.1.2 Multi-layer AMP 112
4.2 Main result 113
4.2.1 Proof Sketch 114
4.3 Numerical Experiments 116

5 Proofs for the multi-layer random convolutional model 118
5.1 Proof of the main theorem 118
5.1.1 State evolution for generic multilayer AMP iterations with matrix valued variables and dense Gaussian matrices 118
5.1.2 State evolution for multilayer AMP iterations with random convolutional matrices 122
5.1.3 Bayes-optimal MLAMP with random convolutional matrices 132

6 Asymptotics of stochastic gradient descent 135
6.1 Introduction 135
6.2 Related works 137
6.3 Main result 138
6.3.1 Examples of algorithms belonging to the considered family 138
6.3.2 Statement of the main theorem 140
6.4 Proof 142
6.4.1 A first example: gradient descent with sample splitting 142
6.4.2 The general case 144
6.5 Useful definitions and probability results 151
6.6 Proof of Theorem 9 151
6.6.1 Relaxing the non-degeneracy assumption 154
6.7 Detailed mapping for Nesterov acceleration 155

II Exact asymptotics for convex models : feature maps, ensembling and multiclass problems **157**

7 Learning curves of generic features maps for realistic datasets with a Gaussian covariate model **158**
7.1 Introduction . . . 158
7.2 Main technical result . . . 161
7.2.1 Random kitchen sink with Gaussian data . . . 164
7.2.2 Kernel methods with Gaussian data . . . 165
7.2.3 GAN-generated data and learned teachers . . . 165
7.2.4 Learning from real data sets . . . 167

8 Proofs for the Gaussian covariate model **170**
8.1 Necessary assumptions . . . 171
8.2 Main theorem . . . 172
8.2.1 Theoretical toolbox . . . 173
8.2.2 Determining a candidate primary problem, auxiliary problem and its solution. 179
8.2.3 Study of the scalar equivalent problem : geometry and asymptotics. . . . 188
8.2.4 Back to the original problem : proof of Theorem 14 and 15 . . . 194
8.2.5 Relaxing the deterministic teacher assumption . . . 197
8.2.6 The 'vanilla' teacher-student scenario . . . 198
8.3 Equivalence with the replica prediction . . . 198
8.3.1 Solution for separable loss and ridge regularization . . . 199
8.3.2 Matching with Replica equations . . . 200

9 Learning Gaussian mixtures with convex generalized linear models **202**
9.1 Introduction . . . 202
9.2 Technical results . . . 205
9.3 Results on synthetic and real datasets . . . 208
9.3.1 Correlated sparse mixtures . . . 208
9.3.2 Separability transition for the cross-entropy loss . . . 209
9.3.3 Binary classification with real data . . . 211

10 Proofs for the Gaussian mixture **214**
10.1 Required background . . . 214
10.2 Reformulation of the problem . . . 218
10.3 Finding the AMP sequence . . . 219
10.4 Proof of Theorem 17 using the AMP sequence . . . 223
10.5 Proof of Lemma 45 . . . 228

11 Fluctuations, Bias, Variance & Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension **231**
11.1 Introduction . . . 231
11.1.1 Setting . . . 232
11.2 Learning with an ensemble of random features . . . 235
11.3 Applications . . . 237
11.3.1 Ridge regression . . . 237

11.3.2 Binary classification . . . 238
11.4 The case of general loss and regularisation . . . 240
11.4.1 The random feature case . . . 243

12 Proofs for the ensembling 245
12.1 Proof of the main theorem . . . 245
12.1.1 The learning problem . . . 245
12.1.2 Asymptotics for the strongly convex problem . . . 246
12.1.3 Relaxing the strong convexity constraint . . . 257
12.1.4 A comment on non-pseudo-Lipschitz subgradients . . . 258
12.1.5 Toolbox . . . 258

III Convex GLMs with left and right orthogonally invariant matrices 259

13 How to prove Kabashima's replica formula 260
13.1 Introduction . . . 260
13.1.1 Background and motivation . . . 260
13.1.2 Main contributions . . . 261
13.1.3 Related work . . . 262
13.2 Background on MLVAMP . . . 263
13.2.1 Link with variable splitting and proximal descent . . . 263
13.2.2 2-layer MLVAMP and its state evolution . . . 263
13.3 Main result . . . 265
13.4 Numerical results . . . 270
13.4.1 Validity of the replica prediction . . . 270
13.4.2 Sparse logistic regression . . . 271
13.5 Sketch of proof of Theorem 22 . . . 274
13.6 Convergence analysis of 2-layer MLVAMP . . . 277
13.6.1 2-layer MLVAMP as a dynamical system : sketch of proof of Lemma 3 . . . 278
13.6.2 Numerical experiments for Lemma 54 . . . 281

14 Proofs for the Kabashima formula 283
14.1 Convergence of vector sequences . . . 283
14.2 Convex analysis and properties of proximal operators . . . 284
14.3 From replica potentials to Moreau envelopes . . . 286
14.4 Fixed point of multilayer vector approximate message passing . . . 287
14.5 State evolution equations . . . 288
14.5.1 Heuristic state evolution equations . . . 288
14.5.2 Necessary assumptions for the rigorous state evolution equations . . . 292
14.5.3 Rigorous state evolution formalism . . . 293
14.5.4 Scalar equivalent model of state evolution . . . 294
14.5.5 Direct matching of the state evolution fixed point equations . . . 296
14.6 Numerical implementation details . . . 299
14.6.1 Regularization : elastic net . . . 299
14.6.2 Loss functions . . . 300
14.7 Proof of Lemma 54: Convergence analysis of 2-layer MLVAMP . . . 301

14.7.1 Proof of Proposition 9 . . . 302
14.7.2 Bounds on $\hat{Q}_{1x}^{(t+1)}, \hat{Q}_{1z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)}$ . . . 302
14.7.3 Operator norms and Lipschitz constants . . . 303
14.7.4 Dynamical system convergence analysis . . . 306
14.8 Analytic continuation . . . 311
14.8.1 Real analyticity of the left hand side of Eq.(14.181) . . . 312
14.8.2 Analytic continuation to $(\tilde{\lambda}_2, \lambda_2) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$ . . . 315
14.8.3 Real analytic approximation of strongly convex problems . . . 315
14.8.4 Continuous extension to $\tilde{\lambda}_2 = 0$ . . . 316
14.8.5 Continuous extension to $\lambda_2 = 0$ . . . 316
14.8.6 Real analytic approximation of usual cost functions with fast decaying higher-order derivatives . . . 317

IV Future directions and bibliography 318
14.9 Future directions . . . 319

Foreword

This manuscript contains the work I did during my PhD at Ecole Normale Supérieure de Paris, mainly under the supervision of Pr. Florent Krzakala. The main body of the text consists in an introduction and three parts, which propose a rigorous analytical approach to the theory of high-dimensional supervised machine learning with random data. This summary assumes the reader is familiar with the field of high-dimensional probability, spin glass theory and supervised machine learning. The reader unfamiliar with these topics may go through the introduction (Chapter 1), and come back to this summary.

Organization of the manuscript and overview of contributions

Chapter 1 proposes a short introduction to machine learning and statistical learning theory, in order to better motivate the need for approaches rooted in high-dimensional probability and mathematical physics, as well as put them into perspective. We then give an overview of statistical physics of disordered systems along with a few of the non-rigorous tools used in this field, such as the cavity and replica method, or asymptotic relaxations of the belief-propagation algorithm. This naturally leads us to the rigorous counterparts of those methods, which can be broadly understood as decoupling procedures for complex probability measures, in order to decompose them into simple product measures for which concentration results are easier to establish and, on a more practical side, numerical evaluation becomes tractable and efficient. After providing a brief description of existing results with i.i.d. Gaussian data, we highlight the main difficulties of bringing theory closer to realistic scenarios, state of the art algorithms and the predictions from statistical physics :

- structured data naturally leads to non-separable problems, whereas a number of existing proof methods dealt with separable ones,
- ensembling algorithms, committee machines and multiclass problems require proofs that give the joint asymptotic distribution of finitely many estimators, rather than single ones,
- all problems are, in the statistical physics sense, at zero temperature, which impedes simplifications given by Bayes-optimality such as the Nishimori identity,
- existing predictions in statistical physics show that exact asymptotics of approximate message passing algorithms may be obtained far beyond generalized linear models, in particular for multilayer problems with random weights or generative priors
- provided one can obtain the exact asymptotics for models with structured data, can we quantify how realistic they are ?

Section 1.7 then provides a glimpse of the main mathematical tools that will be used in this manuscript, namely convex Gaussian comparison inequalities and most importantly, iterative Gaussian conditioning in the context of approximate message passing algorithms. We also illustrate those techniques on simple problems, in order to provide intuition on the results that will be obtained on more complex models. The main reasons for which the goals listed above may be reached can be summarized as follows :

- non-separable models can be handled using convex Gaussian comparison inequalities and appropriate problem decompositions, but they break down for matrix-valued estimators,
- AMP iterations can be rigorously studied with both non-separable effects and matrix-valued iterates, but to study a given estimator one needs to design an iteration converging to this estimator,
- appropriate design and control of the trajectories of AMP iterations may be achieved systematically in the convex case,
- regarding dynamics, the iterative conditioning scheme at the heart of AMP proofs can be extended to multilayer or composite problems involving several random matrices, low-rank perturbations and more,
- benchmark, exactly solvable models that exactly match learning curves obtained on realistic scenarios can be designed with synthetic correlated Gaussian data.

The manuscript is articulated around those ideas, starting with the most general results, before specializing them to the family of convex problems defining estimators found in supervised learning.

In this regard, Part I focuses on the high-dimensional dynamics of AMP iterations for a wide range of models and application of iterative Gaussian conditioning ideas to the study of stochastic gradient descent. We start, in Chapter 2 and Chapter 3, with results that were published in the preprint

[110] C. Gerbelot and R. Berthier, *Graph-based approximate message passing iterations*, arXiv preprint arXiv:2109.11905, (2021)

currently under review. This work extends the proofs of state evolution (SE) equations from [41, 28, 42] to composite AMP iterations by indexing them on an oriented graph and proving that any AMP iteration supported by such a graph admits rigorous SE equations. The graph may be composed arbitrarily to provide new AMP iterations and their SE equations, matching the flexibility of heuristic approaches based on TAP equations for multilayer problems, e.g. [194, 188, 13], which are made rigorous by our result. We show how many of the refinements often encountered in inference problems, such as planted models, spiked matrices or spatial coupling, can be accounted for in our framework.
A first application of those results is proposed in Chapter 4 and 5, where we study the dynamics of multilayer approximate message passing (MLAMP) [188] when the random, dense Gaussian matrices are replaced with random convolutional ones. It is based on the preprint, currently under review,

[70] M. Daniels, C. Gerbelot, F. Krzakala, and L. Zdeborová, *Multi-layer state evolution under random convolutional design*, arXiv preprint arXiv:2205.13503, (2022)

The proof method relies on an embedding of the AMP iteration with convolutional matrices into a larger, matrix-valued one with dense Gaussian matrices where the convolutions are accounted for by designing appropriate circulant non-linearities. In Chapter 6, we continue the discussion started in section 1.7 of the introduction regarding the high-dimensional dynamics of gradient descent methods. We show that the Gaussian iterative conditioning ideas used for the AMP proof in [110] can be used to prove dynamical mean field theory (DMFT) equations, adapted to gradient descent in [198], and recently proven under a more restrictive setup using AMP iterations with memory in [56]. The main contribution is to show that the implicit embedding of gradient descent into an AMP iteration of [56] may be avoided, providing a completely explicit proof were memory kernels of the DMFT prediction build up along the induction. Our result also benefits from the generality of the intermediate lemmas of [110]. This Chapter is based on the following work accepted at *Advances in Neural Information Processing Systems (NeurIPS) 2022*,

[111] C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborova, *Rigorous dynamical mean field theory for stochastic gradient descent methods*, arXiv preprint arXiv:2210.06591, (2022)

We then move to Part II that is concerned with exactly solvable models for supervised learning with realistic feature maps and data models. We start with the analysis of a Gaussian covariate convex generalized linear model, in Chapter 7 and 8 proposed in the published paper

[176] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová, *Learning curves of generic features maps for realistic datasets with a teacher-student model*, Advances in Neural Information Processing Systems, 34 (2021), pp. 18137–18151

where the design matrix has a block covariance structure, representing different feature maps for the teacher and student model. The proof method is based on the convex Gaussian comparison inequalities framework of [281, 204, 57], and matches the replica prediction performed by coauthors. We empirically show that, for a wide range of feature maps, the synthetic Gaussian model with matching covariances exactly captures realistic learning curves for regression tasks, leading to the so called Gaussian equivalent conjecture for those models. The conjecture does not seem to hold as well for classification tasks, prompting the need for another benchmark model.
We thus turn to the study of a multiclass classification problem in Chapter 9 and 10, modelled by the task of learning a finite number of separating hyperplanes of a Gaussian mixture using a matrix-valued convex generalized linear model. The results have been publihsed in the paper

[178] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, *Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions*, Advances in Neural Information Processing Systems, 34 (2021), pp. 10144–10157

The proof method uses a converging trajectory [29, 82] of a carefully designed AMP iteration, involving a representation of the correlated Gaussian mixture as a matrix-valued, spatially coupled [154, 135] problem with non-separable effects. The rigorous state evolution equations are established using our previous results from [110]. The proof result matches the replica computation performed

by coauthors. Simulations then show that, for simple datasets such as MNIST or Fashion-MNIST, the exact learning curves of classification tasks may be predicted using a synthetic Gaussian mixture model where the means and covariances of each cluster is estimated from the data. For more structured tasks, augmenting the number of clusters makes the prediction more accurate.
Motivated by the importance of ensembling methods in machine learning and the insight they provide for neural networks [72], we turn in Chapter 11 and 12 to learning an ensemble of predictors, each of which is defined according to a Gaussian covariate model similar to the one of [176]. The results are based on the published paper

[177] B. Loureiro, C. Gerbelot, M. Refinetti, G. Sicuro, and F. Krzakala, *Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension*, International Conference on Machine Learning (ICML), (2022)

The proof is based on a non-separable, matrix-valued AMP iteration for which we use the same trajectory control as in our previous study [178], and [110] for the rigorous state evolution equations. Once again, the proof matches the replica prediction performed by coauthors. We use the formulas to study the effect of ensembling on usual tasks such as logistic regression, random feature learning and the alignment of different learners.

Finally, Part III presents results published in the papers

[108] C. Gerbelot, A. Abbara, and F. Krzakala, *Asymptotic errors for high-dimensional convex penalized linear regression beyond gaussian matrices*, in Conference on Learning Theory, PMLR, 2020, pp. 1682–1713
[109] C. Gerbelot, A. Abbara, and F. Krzakala, *Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula)*, arXiv preprint arXiv:2006.06581, (2020)

the second of which is currently in review. These results are proofs of replica formulas that were obtained by Y. Kabashima [138, 140, 277], for the specific case of convex generalized linear models, where the design matrix is left- and right-rotationally invariant with a spectrum sampled i.i.d. from an arbitrary distribution with compact support. The result of the second paper [109] is more general than the first one [108], thus the latter is not reproduced here. The reader may nevertheless consult the paper [108] for simpler formulas and more examples of applications. The proof method is based on the construction of converging trajectories of the 2-layer vector approximate message passing (VAMP) algorithm [242, 97] which proposes rigorous state evolution equations for iterations solving generalized linear models with rotationally invariant matrices. Due to the structure of VAMP algorithms, the study of trajectories is different from those of the AMP sequences discussed before : we reformulate 2-layer VAMP as a dynamical system, for which we find an appropriate Lyapunov function, using results from control theory [166]. Our result provides algorithmic convergence guarantees for sufficiently strongly convex problems that do not depend on the high-dimensional nature of the problem. We provide numerical simulations for both the proven replica formula on a variety of generalized linear models and the algorithmic convergence of 2-layer VAMP.

We conclude with a brief discussion on future directions and the bibliography in Chapter IV.

Remark For the papers [108, 109, 110], the author was the main contributor to all aspects of the work (writing, simulations, figures, technical statements and proofs). For the papers [176, 178, 177, 70] the author's contributions were the technical statements/results and related rigorous proofs of formulas obtained using the replica method, which were part of the litterature for [70], and were performed by collaborators B. Loureiro, G. Sicuro and H. Cui for [176, 178, 177]. The author also contributed to the writing of non-technical statements and discussions of those papers. The proofs of the technical results are constructive in that the replica results are not required in advance. Nevertheless, having access to the replica prediction was helpful to understand how the problems could be decomposed to fit the rigorous frameworks of convex Gaussian comparison inequalities and approximate message passing iterations. The replica computations are not reproduced in this thesis and can be found in the original papers, along with simplifications that were used in specific cases for examples and figures. In [70], M. Daniels presented discussions and related works for deep generative models that are not reproduced in this thesis. The interested reader may consult the original paper for further details.

Work not included in this thesis We did not include the following preprint under review

[65] E. Cornacchia, F. Mignacco, R. Veiga, C. Gerbelot, B. Loureiro, and L. Zdeborová, *Learning curves for the multi-class teacher-student perceptron*, arXiv preprint, (2022)

which provides a proof of the replica formula for the convex multiclass perceptron. The author only provided the rigorous proof and did not contribute to the main body of the paper outside of the technical statement. The proof is also based on a converging trajectory of a matrix-valued AMP algorithm and constitutes a simpler instance of the proofs from the aforementioned works [178, 177]. It extends the Bayes-optimal results of [15] proven using the Guerra interpolation to the zero-temperature, convex case.

Chapter 1

Introduction

Although machine learning is now an established field with firm theoretical grounding in optimization, probability and statistics, the recent empirical success of deep learning often challenges the usual knowledge of statistical learning theory. From self-driving cars to numerical solvers for high-dimensional systems of partial differential equations, the possibilities offered by the variety of methods encompassed by artificial intelligence go well beyond problem-specific combinations of statistical estimators. This has prompted a surge of interest into new theoretical approaches to bridge the gap between the fast paced empirical progress and slower paced theoretical one. The goal of this chapter is to briefly present the core concepts in machine learning, statistical physics and probability that motivate the family of problems investigated in the present work as well as the theoretical approach that is chosen. Naturally, the presentation is far from exhaustive and pointers to appropriate references are provided throughout.

1.1 Artificial intelligence

One way to approach the field of artificial intelligence is through the formalization of physiological concepts. For instance learning to perform a given task from examples, defining notions of similarity to organize a set of unknown objects into groups or adapting a behaviour to an environment for an organism to thrive. The mathematical formulation of these notions leads to the three main methodologies of modern machine learning, see e.g. [206], respectively : supervised learning, unsupervised learning and reinforcement learning. In supervised learning, one seeks to reconstruct a function, or probability distribution, the output of which we observe through a given set of samples, the *training set*. Unsupervised learning consists in defining a notion of similarity in order to separate a given set of elements into groups where members of each group approximately have the same measure of similarity. Reinforcement learning relies on the optimization of a reward function with a sequence of decisions based on a time varying interaction with an unknown environment. Those three problems have close ties to existing fields with extensive litterature. Statistical inference [292] and signal processing [185] are both concerned with the reconstruction of quantities (codes, images, ...) based on available measurements, while kernel density estimation in nonparametric statistics [293] can be used on a non-labeled dataset to estimate the underlying density that generated the samples. Finally, the optimization of a desired outcome from a time-dependent process is at the core of control theory [146]. We may therefore wonder what makes machine learning different, and more precisely, given the variety of theoretical results in the existing fields discussed above, what technical challenges are brought by the practical goals of artificial intelligence.

A first difficulty is the absence or lack of knowledge about the ground truth operating behind either labels, data points or the environment. Indeed, while in control theory one seeks to optimize a strategy given a known system, reinforcement learning adds the process of discovering the environment. A second difficulty comes from the high-dimensional nature of the problem, brought by the increasing amount of available data for a number of tasks and large number of parameters in state-of-the-art models. Classical statistics result for instance, where the number of predictors is usually assumed to be much smaller than the number of data points, are known to break down when the dimension becomes comparable or larger than the number of samples [292]. Finally, machine learning aspires to be "intelligent" : not only do we want to solve the aforementioned problems, but we want the methods to adapt to whatever structure is present in each instance, without having to manually tailor them to those structures. For instance any high-dimensional problem intrinsically depending on a latent space of lower dimension should be identified as such by the algorithm, which would then learn an optimal approximation of the target function on this latent space.

1.2 Supervised learning

Let us now focus on supervised learning, which will be the motivation for the problems considered in this work. Our main reference for this part is [206].

1.2.1 Empirical risk minimization

Consider a given set of n points $(\mathbf{x}_1, ..., \mathbf{x}_n)$ in $\mathbb{R}^d$, labeled according to a hidden joint density $p^*(\mathbf{x}, y)$. The set $(\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y})$, where the vector $\mathbf{y} \in \mathbb{R}^n$ contains the available labels, is referred to as the *training set*. The goal is to find a candidate function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ belonging to a chosen candidate functional space $\mathcal{F}$ in order to best reproduce the joint density $p^*(\mathbf{x}, y)$. To do so, the usual approach is to minimize an error measure defined by a *cost function* $C : \mathbb{R}^2 \to \mathbb{R}$, leading to the following optimization problem over the *expected risk*

$$\hat{f} \in \inf_{f \in \mathcal{F}} \quad \mathbb{E}_{(\mathbf{x},y)\sim p^*} \left[C(f(\mathbf{x}), y) \right]. \tag{1.1}$$

However, since we only have access to a finite set of realisations of p^*, the expected risk is replaced by the *empirical risk*, leading to

$$\hat{f} \in \inf_{f \in \mathcal{F}} \quad \frac{1}{n} \sum_{i=1}^{n} C(f(\mathbf{x}_i), y_i) \tag{1.2}$$

which, assuming the samples are drawn independently, should be a logical proxy for problem (1.1) according to the law of large numbers. The difference in performance between the estimators obtained from the expected and empirical risk is the *generalization error*, i.e. the ability of a model trained on a finite number of samples to predict new labels reliably. The complexity of the probability distribution p^* and dimensionality of the problem will govern how well the empirical risk approximates the expected one for a given number of samples. We can thus expect that these quantities will directly appear in theoretical predictions for the performance of a given estimator. Then, the expressivity of the functional space $\mathcal{F}$, that is the variety of functions it can express, also plays a key role. A typical example of this is polynomial regression of a sinusoidal function in

one dimension, see e.g. [40]. On the one hand, if no limitations are placed on the degree of the polynomial, any finite set of pairs (x, y) sampled from the ground-truth can be interpolated by the corresponding Lagrange polynomial, which can vary greatly for different realisations of the dataset, even if it captures complex behavior on a single dataset : the estimator is *overfitting* the dataset. On the other hand, if we restrict the candidate functional space to linear or quadratic functions, the model will be too simple and present a high bias with respect to the ground truth. This dilemma is referred to as the *bias-variance* tradeoff in machine learning. The common approach is then to choose a fairly expressive set of functions and add a *regularization term* to the problem (1.2) by constraining the norm of f :

$$\hat{f} \in \inf_{f \in \mathcal{F}} \quad \frac{1}{n} \sum_{i=1}^{n} C(f(\mathbf{x}_i), y_i) + \lambda \|f\|_{\mathcal{F}}^2 \tag{1.3}$$

where λ is a positive scalar parameter. For a concrete example, if $\mathcal{F}$ is a Sobolev space, the regularization term will constrain the total variation of higher order derivatives and impose a degree of smoothness depending on the value of λ. The main practical challenges of supervised learning can thus be summarized as follows

- the choice of the candidate functional space $\mathcal{F}$
- the choice of the loss function (and regularisation)
- the choice of the optimization algorithm to solve problem (1.3)

The main theoretical challenge is to have mathematical justifications for these choices.

1.2.2 Choosing the candidate functional space

The appropriate transformation of data can lead to drastic simplification of a problem. For instance, consider a 2-dimensional task of separating datapoints distributed according to two noisy concentric circles. Parametrizing the boundary between the two sets for classification purposes can seem difficult when adopting a naive approach. A simple change of parametrization from the initial (x_1, x_2) to polar-like coordinates $(x_1^2, \sqrt{2}x_1x_2, x_2^2)$ leads to a linear boundary [265]. In this example however, the human eye spots the circular geometry of the data, which may be much harder to do in high-dimension, with structures that go well beyond concentric circles !

Linear models Following the statistics litterature, e.g. [292], the most common estimators are linear ones, parametrized by a weight vector denoted $\mathbf{w} \in \mathbb{R}^d$. Concatenating the samples $(\mathbf{x}_1, ..., \mathbf{x}_n)$ into a design matric $\mathbf{X} \in \mathbb{R}^{n \times d}$, the optimization problem defining a linear estimator then reads

$$\inf_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} C(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda r(\mathbf{w}) \tag{1.4}$$

where the the function $r : \mathbb{R}^d \to \mathbb{R}$ is typically a norm, and we consider any intercept as included in the dimension of the input space d, without loss of generality. Although linear models have weak expressive power, they are both simple to implement and to analyze theoretically. In the case of convex cost and regularization functions, they form the family of *generalized linear models* (GLM), the basis of many machine learning algorithms such aas least-squares regression or max-margin classification. Finding an appropriate functional space can then be seen as finding a mapping

$\phi : \mathbb{R}^d \to \mathbb{R}^p$, that should be tailored to each problem instance. Such mappings are often referred to as *feature maps* in the machine learning litterature, leading to the formulation :

$$\inf_{\mathbf{w}\in\mathbb{R}^p} \quad \frac{1}{n}\sum_{i=1}^{n} C(\mathbf{w}^\top \phi(\mathbf{x}_i), y_i) + \lambda r(\mathbf{w}) \tag{1.5}$$

where r is now defined on $\mathbb{R}^p$. Refining linear models then resides in finding the good feature map.

Kernel methods The originally predominent method to choose feature maps were kernel methods [258, 265], which is a form of non-parametric regression. The idea is to use a *reproducing kernel Hilbert space (RKHS)* [10], as the candidate functional space, and use its reproducing property to find a tractable form of the optimization problem now defined over a potentially infinite dimensional feature space. The target RKHS is defined by a *reproducing kernel*, i.e. a bilinear symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and is composed of all linear combinations of the functions $K(\mathbf{x}_i, .)$ supported by the points in $\mathcal{X}$ along with the pointwise limits of the corresponding Cauchy sequences. The reproducing property then states that for any function f in the RKHS, its value at any point $\mathbf{x}_i$ can be expressed through the inner product $f(\mathbf{x}_i) = \langle f, K_{\mathbf{x}_i}\rangle_{\mathcal{F}}$. Provided the cost function is increasing in $\|f\|_{\mathcal{F}}$, which is easily enforced with the regularisation, an orthogonal decomposition shows that the predictor can be expressed as a linear combination of the kernel functions supported by the points in the dataset, i.e. there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that $f = \sum_{j=1}^{n} \alpha_i K(\mathbf{x}_i, .)$. The optimization problem (1.3) can then be expressed as

$$\inf_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad \frac{1}{n}\sum_{i=1}^{n} C((\mathbf{K}\boldsymbol{\alpha})_i, y_i) + \lambda r(\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}) \tag{1.6}$$

effectively reducing the search to an $n-$dimensional linear regression, where the *kernel matrix* $\mathbf{K} \in \mathbb{R}^{n\times n}$ is defined by $\mathbf{K}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$ for any $1 \leqslant i, j \leqslant n$, and is positive definite, see e.g. [265]. This result is called the *representer theorem* and spawns a wide range of models which can be analyzed theoretically using functional analysis combined with the framework of linear models. Reproducing kernels can then be manually tailored depending on the different tasks at hand, ranging from polynomial kernels $K(\mathbf{x}, \mathbf{x}') = \left(1 + \mathbf{x}^\top \mathbf{x}'\right)^k$ for vector-valued data to the Fisher score of probabilistic models for strings (sentences, DNA sequences, etc ...). Despite their elegance, firm theoretical grounding and apparent limitless expressive power, kernel methods are not adaptative : each reproducing kernel has to be chosen manually and tuned for each problem, and linear combinations or products of usual kernels hardly solve this issue. Finding a correct basis to decompose a function on is also a long standing problem in harmonic analysis, with Fourier and wavelet decompositions [185] being the most widely used examples in statistics, signal processing and machine learning. As is the caase with kernel methods, Fourier or wavelet decompositions still rely on a fixed set of basis functions. Although adaptative methods using wavelet decompositions can achieve impressive performance on complex tasks such as image recognition [52], the family of feature maps now holding the state of the art in close to all applied fields are neural networks.

Neural networks Inspired by biological neurons, the perceptron was proposed by Rosenblatt [245] as a model of information storage in the brain. It is simply defined as a sigmoidal *activation function*, a hyperbolic tangent for instance, taking as input a scalar product.

$$f(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) \tag{1.7}$$

where $\mathbf{w} \in \mathbb{R}^d$ are the trainable parameters of the model. Compositions of linear combinations of perceptrons led to the the multilayer perceptron (MLP), the first deep learning model, along with its gradient-based optimization [249], with the notable application of document recognition [157]. Neural networks keep breaking benchmarks on tasks of increasing complexity in computer vision [150], natural language processing [201], etc ..., and routinely solve NP-hard problems for reasons that are still unclear. A neural network with L *layers* is thus a parametric model with parameters $\mathbf{W}_1, ..., \mathbf{W}_L$ where, for any $1 \leqslant i \leqslant L$, $\mathbf{W}_i \in \mathbb{R}^{n_i \times n_{i-1}}$ with $n_0 = d$, the input dimension.

$$f(\mathbf{x}) = \sigma_L \left(\mathbf{W}_L \sigma_{L-1} \left(\mathbf{W}_{L-1} .. \sigma_1 \left(\mathbf{W}_1 \mathbf{x} \right) \right) \right) \tag{1.8}$$

From an approximation point of view, multilayer perceptrons are known to be able to approximate any continuous functions, when sufficiently wide, under mild conditions [69, 20], while being completely parametric. The empirical risk minimization problem now reads

$$\inf_{\{\mathbf{W_i}\}_{i=1,...,L}} \quad \frac{1}{n} \sum_{i=1}^{n} C(\sigma_L \left(\mathbf{W}_L \sigma_{L-1} \left(\mathbf{W}_{L-1} .. \sigma_1 \left(\mathbf{W}_1 \mathbf{x}_i \right) \right) \right), y_i) + \lambda r \left(\mathbf{W}_1, ..., \mathbf{W}_2 \right), \tag{1.9}$$

which can be optimized explicitly using gradient based methods. Empirically, neural networks seem to adapt and learn automatically the appropriate representation from data, and therefore solve the problem of finding the appropriate basis change we are looking for. An entire bestiary of network architectures now exists [119], with a variety of practical tricks to improve generalization, trainability or interpretability. One can naively interpret the success of deep learning with the fact that neural networks are a heavily parametrized and completely tunable way to represent arbitrary functions. However, this intuition does not answer the questions of choosing the aactivation functions, the width and depth, how to regularize, etc ... Finding the appropriate functional analysis framework to describe neural networks and their adaptative properties is an active research topic, see e.g. [234, 275, 16, 179] and is beyond the scope of this thesis. As advocated by approaches inspired by statistical physics, we will focus on simpler models that capture some of the empirical behaviours of neural networks, and that can be studied exactly. But for now, let us continue with our description of supervised learning.

1.2.3 Loss functions and optimization

Supervised learning tasks are usually separated into two types : *regression* and *classification*. Regression aims at reconstructing a function with continuous output, while classification is concerned with finding a discrete valued function that best separates object into groups labeled by the output of the function. For regression, the square loss appears as a natural choice : the further away we are from the available output, the larger the cost. For classification however, the cost should be the same for all predictions falling into the wrong class, and zero for the correct ones. This prompts the use of the $0-1$ loss, whose discontinuity makes it difficult to optimize. The most widely used method is to use convex surrogates, such as the hinge or logistic losses, which approximate the $0-1$ behaviour in a smoother manner and benefit from the optimization guarantees of convexity [244]. Once a convex objective is formulated, a wide variety of optimization algorithms can provably reach the estimator of interest in polynomial time, such as gradient descent [212, 49] or proximal based methods [50, 224]. We will give more background on proximal operators later on, as they will play a key role in some of our results. As mentioned above, a kernel regression problem can be reduced to a linear model, thus, once a convex loss and regularisation are chosen, methods from convex optimization also apply to kernel methods. For neural networks however, the objective function is

highly non-convex and the number of parameters can be quite large. The methods of choice for deep learning are stochastic gradient descent (backpropagation in deep learning) [249] along with a variety of landscape and data adaptative variants [87, 144]. Theoretical guarantees for non-convex landscapes are much harder to obtain that for convex ones, and consitute an active research topic in optimization [134]. Controlling high-dimensional trajectories of a certain class of algorithms will turn out to be crucial in this thesis, and we will also study algorithmic convergence properties that do not depend on the dimensionality of the problem. While exact asymptotics for stochastic gradient descent methods that do not depend on convexity will also be proven, we will not study converging trajectories in non-convex settings.

1.3 Statistical learning theory

The goal of statistical learning theory [48, 206] is to provide robust bounds to estimate the performance of a given estimator $\hat{f}$ defined by (1.3) for a given task. Robustness is at the heart of the approach, in order for the predictions to hold in a wide range of practical cases which may involve complex underlying functions or data distributions. The aforementioned bias-variance tradeoff can be formalized by introducing the Bayes error, i.e. the minimum achievable error for a given cost function if we assume the distribution $p^*(\mathbf{x}, y)$ is known, leading to the *Bayesian decision* $f_{Bayes}(\mathbf{x}') = \inf_z \mathbb{E}\left[C(z, y)|\mathbf{x} = \mathbf{x}'\right]$. Defining the cost $\mathcal{R}_f = \mathbb{E}\left[C(f(\mathbf{x}, y))\right]$, the excess risk for an estimator $\hat{f}$ can then be decomposed as

$$\left|R_{\hat{f}} - R_{f_{Bayes}}\right| = \underbrace{\left|R_{\hat{f}} - \inf_{f\in\mathcal{F}} R_f\right|}_{E_1} + \underbrace{\left|\inf_{f\in\mathcal{F}} R_f - R_{f_{Bayes}}\right|}_{E_2}. \tag{1.10}$$

The term E_1 represents the error coming from the approximation of the expected risk by the empirical risk, and will become larger as overfitting becomes predominent. The term E_2 represents the approximation error, that is the ability of the candidate functional class $\mathcal{F}$ to approximate the Bayesian decision f_B. We thus recover the dilemma of expressivity described in section 1.2.1. Theoretical analysis of the approximation error often involves the decomposition of the target function on a suitable basis of the candidate functional space, for instance spherical harmonics if we assume the data points to have bounded norm, enabling direct comparisons of the coefficients. This usually gives rates of approximation mainly depending on regularity assumptions of the underlying truth (smoothness, etc ...). The litterature on function approximation is quite extensive, notably in numerical methods for partial differential equations, harmonic analysis and non-parametric statistics, and approximation error proofs are often based on related methods. Bounding the generalization error term E_1 is more characteristic of machine learning, and rests on the notion of uniform bounds, i.e. the convergence of the empirical risk to the expected one over all functions in the class. Such control may be achieved using the *Rademacher complexity*, which represents the ability of a given function class to fit random noise, and reads

$$\mathcal{R}_X(f) = \mathbb{E}_{\mathbf{x},\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} |\frac{2}{n}\sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i)\right] \tag{1.11}$$

where the σ_i's are i.i.d. Rademacher variables and $\mathbf{x}$ is the data distribution. We give an example from [48] for $d = 1$

Theorem 1. *For all $\delta > 0$, with probability at least $1 - \delta$*

$$\forall\, f \in \mathcal{F} \quad \mathbb{E}\left[f(x,y)\right] \leqslant \frac{1}{n}\sum_{i=1}^{n} f(x_i, y_i) + \sqrt{2\mathcal{R}_x(f) + \frac{\log(1/\delta)}{n}} \tag{1.12}$$

The game of bounding the generalization error then consists in accounting for higher dimensionality d, and explicit evaluation of the Rademacher complexity using structural assumptions on the functional class $\mathcal{F}$, see e.g. [22]. Uniform bounds in statistical learning theory can also be understood from the point of view of upper and lower bounds of random processes, which we will use for a family of Gaussian processes, which are presented in [288] in a common, succinct and elegant way in chapter 7 and 8 of [288]. Although these bounds are robust and adaptable to a wide range of machine learning problems, it can be difficult to control the tightness of the bound or more intuitively, how far the actual behaviour of an estimator actually is from the upper bound. Also, the bounds are agnostic to the data distribution and taking the supremum over the functional class, i.e. considering the worst possible function, may not be the most representative way of what happens on average. Furthermore, in modern applications both the number of samples and the dimension of the feature space are very large. Indeed, for a polynomial kernel of degree k on an original feature space of dimension d, the new feature space is of dimension $\binom{d+k}{k}$, while modern neural networks can have sevral thousands (even millions) of parameters [150, 201]. It may therefore be interesting to consider simpler benchmark problems, with explicit data distributions, ground truth and candidate functional spaces, where exact solutions can be obtained using stronger statements in concentration of measure and large deviation theory. This is precisely what has been done in statistical physics for over a century.

1.4 Statistical physics of disordered systems

This section is largely based on the lecture notes [155]. Long before machine learning, extracting meaningful quantities from a large number of interacting random variables has been at the heart of statistical physics for over a century. Models in statistical physics aim at understanding the behaviour of macroscopic physical systems composed of many microscopic particles through a reduced number of scalars, often called *order parameters*. The study of magnetism in solids [295] brought early versions of notions commonly used in machine learning such as the mean field approximation, for example. Considering the average number of particles in physical systems, typically Avogadro's number of $6.022e23$, the application of natural laws from classical, quantum or relativistic mechanics to each individual particle appears unrealistic. Particles are thus described by *ensembles*, i.e. probability distributions describing the likeliness for the system to be in a given state.

Equilibrium statistical physics In equilibrium statistical physics, the most commonly used description is the Boltzmann probability distribution defined over a set of n particles $(w_1, ..., w_n)$, where n will be assumed very large. The particles interact according to the potential, or *Hamiltonian* $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}$ at an inverse temperature β, leading to the joint distribution of particles

$$p_{\mathcal{H},\beta}\left(w_1, ..., w_n\right) = \frac{1}{Z_n(\beta)} \exp^{-\beta\mathcal{H}(\{w_i\}_{i=1,..,n})} \tag{1.13}$$

The *partition function* $Z_n(\beta) = \int_{\mathcal{X}^n} \exp^{-\beta\mathcal{H}(\{w_i\}_{i=1,..,n})} \prod_i dw_i$ plays a key role in statistical physics, in particular in the form of the *free energy* $\Phi_n = \frac{\log(Z_n)}{n}$, which is closely related to the moment

generating function of the Boltzmann measure. Note that, by taking the zero temperature limit in Eq.(1.13), the problem reduces to finding the ground state of the Hamiltonian. The Boltzmann measure formulation thus contains both the sampling and optimization approaches to estimation, depending on the chosen value of β. This will be discussed further in the next section. One of the simplest examples is the Curie-Weiss ferromagnet [295], where a systems of d random variables $(s_1, ..., s_n)$ (spins) taking values in $\{-1, +1\}^n$ interact according to the potential

$$\mathcal{H}_n(s_1, ..., s_n) = -\frac{1}{2n} \sum_{1 \leqslant i,j \leqslant n} s_i s_j - h \sum_i s_i \tag{1.14}$$

The goal in this problem is to find the asymptotic value of the average magnetisation $\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i$, when the number of particles diverges. A fully rigorous combinatorics argument then shows that, in the high-dimensional limit, the free energy converges to the optimal value of the one-dimensional optimization problem

$$\lim_{n \to \infty} \Phi_n = \sup_m \phi(m) \tag{1.15}$$

$$\text{where} \quad \phi(m) = H(m) + \frac{1}{2}\beta m^2 + \beta h m \tag{1.16}$$

$$\text{and} \quad H(m) = -\frac{1+m}{2} \log\left(\frac{1+m}{2}\right) - \left(\frac{1-m}{2}\right) \log\left(\frac{1-m}{2}\right) \tag{1.17}$$

whose zero-gradient condition reads

$$m = \tanh(\beta(m + h)) \tag{1.18}$$

. This leads to a large deviation principle for $\bar{s}$ which shows that, if equation (1.18) has a unique solution m^*, then $\bar{s}$ converges with high probability to m^*. This example illustrates the intuition at the heart of statistical physics : to understand the behaviour of a complex, high-dimensional system with an asymptotically exact relation involving only low dimensional quantities and simple functions. Models admitting asymptotic characterizations of this flavour are called *exactly solvable*, and the related low dimensional equations form the *mean field* description of these systems. An entire bestiary of exactly solvable models can be found in the statistical physics litterature, going well beyond the equilibirum Boltzmann measure, notably out-of-equilibrium problems and disordered systems, which we will now describe.

Disordered systems Disordered systems are sets of particles whose interactions are parametrized by additional random variables. A notable example are *spin glasses*, originally models to understand magnetism in solids. The simplest instance is the random field Ising model, for which the Hamiltonian reads

$$\mathcal{H}_n(s_1, ..., s_n) = -\frac{1}{2n} \sum_{1 \leqslant i,j \leqslant n} s_i s_j - \sum_i h_i s_i \tag{1.19}$$

where the $\mathbf{h}$ is a vector with i.i.d. $\mathcal{N}(0, \Delta)$ elements. In similar fashion to the Curie-Weiss model, the average magnetisation obeys a large deviation principle governed by the fixed point equation

$$m = \mathbb{E}\left[\tanh(\beta\,(h + m))\right] \tag{1.20}$$

Beyond the actual phenomenology of the model, introducing the disorder $\mathbf{h}$ leads to a key technical difference : the rigorous combinatorics argument leading to the solution of the Curie-Weiss model

does not go through for the random field Ising model. Various non-rigorous methods were developed in theoretical physics to tackle problems involving disordered Hamiltonians, notably the replica method [196]. Based on the identity $\log Z = \lim_{n\to 0} \frac{Z^n-1}{n}$, the replica method allows to compute the moments of a Boltzmann measure by decoupling the powers in the integral defining Z^n using field-theoretic arguments and heuristic central-limit like results. The final step of taking the limit $n \to 0$ is also heuristic. The replica method was famously used by recent Nobel Prize recipient Giorgio Parisi [225, 226] to study the landscape of the Sherrrington-Kirkpatrick Hamiltonian [266], defined by the optimization problem

$$\sup_{\mathbf{s}\in\{-1,+1\}^n} \mathbf{s}^\top \mathbf{A}\mathbf{s} \tag{1.21}$$

where $\mathbf{A}$ is an element of the Gaussian orthogonal ensemble $GOE(n)$. The variational principle governing the set of solutions to this problem is far more complicated than that of the Curie-Weiss or random field Ising model, and still motivates research to this day [279, 221, 207]. Replicas and other theoretical physics inspired methods, although originally meant for spin glasses, have been successfully applied to a variety of problems such as coding theory, combinatorial optimization and more recently, machine learning [215, 195, 196, 154]. The reader familiar with the probabilistic approach to machine learning will recognize some of the concepts inherent to statistical physics : approximating a distribution with a simpler one for optimization and tractability purposes is one of the main goals of variational inference [290], where the term mean field is often used as well. A common relative to those fields can be found in the *belief-propagation* [100, 227] algorithm, which is mainly known as an iterative marginalization procedure exploiting the conditional independence structure of probability distributions supported by graphical models. The intermediate, partially integrated marginals that are transmitted in the algorithm are often called *messages*. An early instance can be found in physics, once again in a model to study magnetism in solids, in the theory of superlattices [38]. A limitation of belief-propagation is the restriction of its exactness and convergence to tree graphical models, and generalizations of the algorithm to loopy graphs have been the subject of intense scrutiny both in statistical physics [195] and machine learning, see e.g. [299] and references therein for the machine learning part. For disordered systems, the asymptotic analysis of belief propagation and the approximation of complex probability distributions by locally tree-like graphs has led to the so-called cavity method and Thouless-Anderson-Palmer (TAP) equations [196], which reduce the problem of computing a complex partition function to solving a set of scalar, non-linear equations. The intuition underlying those methods is once again a heuristic form of concentration of measure : messages in the BP algorithm, or consistency conditions of individual marginals can lead to simple, asymptotically exact low-dimensional descriptions in the large system limit. Spin glasses have thus provided a true cornucopia [8] of methods to obtain large deviation principles and concentration results on a priori intractable problems. Since these results are heuristic but in most cases extensively verified through simulation and surprisingly robust, it appears quite natural to attempt to understand the mathematical reasons operating behind them.

The rigorous approach : high-dimensional probability The description of statistical physics given above highlights the value of its insights for probability theory : a wide range of a priori highly non-trivial probability distributions exhibit large deviations principles and concentration properties that enable to characterize new phenomena in probability theory and random geometry, for instance the asymptotic volume of the intersection of a discrete cube $\{-1,+1\}^d$ with a number $p = \alpha d$ of i.i.d. random half-spaces [102]. An entire branch of probability theory is therefore devoted to the rigorous mathematical study of spin-glass like systems [279, 221], and has given birth to an

extensive mathematical toolbox whose main purpose can be summarized in the following way : for a given probability measure involving a large number of interacting, high-dimensional particles, find a decomposition as a product measure of simple components, typically independent, parametrized by a finite set of low-dimensional parameters, that captures the exact asymptotic behaviour of the original measure. Once this decomposition is found, it becomes much easier to study concentration properties using existing results for independent random variables, see e.g. [47, 288]. Concentration of measure [160] and large deviations [286] are thus omnipresent in this field, with ties to extrema of random processes [161], random matrices [7] and applied mathematics, notably in optimization [147] and sampling [149]. The need to make predictions obtained with statistical physics methods rigorous and ground them in concrete mathematical concepts is particularly relevant for machine learning, where robustness holds a central place. The litterature on rigorous results inspired by statistical physics thus extends to the machine learning setup, joining high-dimensional statistics and applied probability. This thesis is a contribution to this field, and proves results in the context of the statistical physics approach to supervised learning, which we now describe.

1.5 Statistical physics of supervised learning

The benefit of the statistical physics methodology, along with the corresponding rigorous mathematics, is quite clear : obtaining an exact description in terms of simple distributions allows to compute all the quantities a statistician would be interested in : reconstruction error, confidence intervals, rates, etc ... The typical framework studied in this field is the teacher-student scenario, see, e.g., [300] where the performance of a given learning method (the student) is studied in the recovery of a given generative model (the teacher). The usual formulation is that of probabilistic inference : consider a ground truth vector $\mathbf{w}_0 \in \mathbb{R}^d$ ditributed according to a probability density $p_{0,w}(\mathbf{w}_0)$. We then observe an output of n observations $\mathbf{y} \in \mathbb{R}^n$ from a transition probability $p_{0,y} = p_{0,y}(\mathbf{y}|\mathbf{w}_0)$, which may include other sources of randomness such as noise. The goal is to reconstruct the ground truth vector $\mathbf{w}_0$ and transition probability $p_{0,y}$. The minimum mean squared error estimator (MMSE) then reads, using Bayes rule

$$\hat{\mathbf{w}} = \mathbb{E}\left[\mathbf{w}|\mathbf{y}\right] = \frac{1}{Z(\mathbf{y})}\int_{\mathbb{R}^d} \mathbf{w} p_{0,w}(\mathbf{w}) p_{0,y}(\mathbf{y}|\mathbf{w}) d\mu(\mathbf{w}) \tag{1.22}$$

where μ is the Lebesgue measure on $\mathbb{R}^d$. Here we assume that the probability distributions defining the ground truth are known, which means we may study the actual MMSE : this is the *Bayes-optimal* scenario. It is particularly relevant for signal processing, or to evaluate fundamental limits of inference such as recovery thresholds from noisy measurements. Indeed, for square integrable random variables, the conditional expectation represents the best possible approximation in ℓ_2 norm of a random variable given the sigma-algebra of the observed one. In the non-Bayes optimal scenario, the ground truth distributions are not available, and we postulate a model $p_{1,w}(\mathbf{w}), p_{1,y}(\mathbf{y}|\mathbf{w})$ to estimate $\mathbf{w}_0$ with

$$\hat{\mathbf{w}} = \mathbb{E}\left[\mathbf{w}|\mathbf{y}\right] = \frac{1}{Z(\mathbf{y})}\int_{\mathbb{R}^d} \mathbf{w} p_{1,w}(\mathbf{w}) p_{1,y}(\mathbf{y}|\mathbf{w}) d\mu(\mathbf{w}) \tag{1.23}$$

where $\mathbf{y}$ is observed. To recover the optimization problems usually found in supervised learning, consider the postulated densities

$$p_{1,\mathbf{w}} \propto \exp(-\beta r(\mathbf{w})) \quad p_{1,y}(\mathbf{y}|\mathbf{w}) \propto \exp(-\beta L(\mathbf{w}, \mathbf{y})) \tag{1.24}$$

where L, r are usually positive functions, and β a positive scalar parameter. We thus recover a Boltzmann measure

$$\hat{\mathbf{w}}_\beta = \mathbb{E}\left[\mathbf{w}|\mathbf{y}\right] = \frac{1}{Z(\mathbf{y})}\int_{\mathbb{R}^d} \mathbf{w}\exp(-\beta r(\mathbf{w}))\exp(-\beta L(\mathbf{w},\mathbf{y}))d\mu(\mathbf{w}) \tag{1.25}$$

As mentioned in section 1.2.1, the transition probability corresponding to supervised learning will depend on a design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ through the product $\mathbf{Xw}$ for linear models. Here, by linear model, we also mean with respect to a feature map such as a kernel or a learnt neural network. The estimator then reads

$$\hat{\mathbf{w}}_\beta = \mathbb{E}\left[\mathbf{w}|\mathbf{y}\right] = \frac{1}{Z(\mathbf{y})}\int_{\mathbb{R}^d} \mathbf{w}\exp(-\beta\left(L(\mathbf{Xw},\mathbf{y}) + r(\mathbf{w})\right))d\mu(\mathbf{w}) \tag{1.26}$$

In order for this model to be exactly solvable, an assumption on the design matrix should be made, the most classical one being i.i.d. normal elements with variance $\frac{1}{d}$. The Boltzmann density $\frac{1}{Z(\mathbf{y})}\mathbf{w}\exp(-\beta\left(L(\mathbf{Xw},\mathbf{y}) + r(\mathbf{w})\right))\mathbf{w}$ may then be studied using tools from disordered systems, in the proportional limit $n, d \to \infty$ with $n/d = \alpha$ for finite values of α, leading to asymptotically exact, closed form expressions for key quantities such as the average mean-squared error $\frac{1}{d}\|\hat{\mathbf{w}} - \mathbf{w}_0\|_2^2$ or the average test error between the output of the postulated model with respect to the ground truth on a fresh data sample. At strictly positive temperatures, i.e. finite β, the problem of estimating $\hat{\mathbf{w}}$ boils down to the evalutation of a posterior mean, for which the belief-propagation algorithm is particularly suited. Although the graph representing the Boltzmann distribution Eq.(1.26) is dense, the corresponding BP equations can be simplified in the high-dimensional limit, showing that the messages are asymptotically Gaussian in the case of independent elements (not necessarily identically distributed), with appropriately scaled variance, leading to the family of approximate message passing algorithms [280, 196]. These algorithms have then been sucessfully used in statistical inference with random design, notably starting with the LASSO [83, 84, 85]. AMP iterations and the related proofs will be one of the central subjects of the work that follows, thus we postpone further background to the next sections. Assuming the minimum of the cost $L(\mathbf{Xw},\mathbf{y}) + r(\mathbf{w})$ is well-defined, we may take the $\beta \to +\infty$ limit and use Laplace's approximation to recover the setup of empirical risk minimization :

$$\lim_{\beta\to+\infty} \hat{\mathbf{w}}_\beta \in \inf_{\mathbf{w}^d} \quad L(\mathbf{Xw},\mathbf{y}) + r(\mathbf{w}) \tag{1.27}$$

which is indeed the typical supervised learning setup for a linear model. Now that the link between statistical physics and empricial risk minimization has been provided, we will put aside the probabilistic formulation of supervised learning leading to Boltzmann-like measures and focus on the high-dimensional optimization problem :

$$\hat{\mathbf{w}} \in \inf_{\mathbf{w}\in\mathbb{R}^d} \quad L(\mathbf{Xw},\mathbf{y}) + r(\mathbf{w}) \tag{1.28}$$

$$\text{such that} \quad \mathbf{y} = f_0(\mathbf{Xw}_0), \tag{1.29}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}^n$ represents a label generating function, which may include additional sources of randomness such as noise, and is generally separable across lines. In the case where the functions L, r are convex, the minimization problem (1.28) represents the class of *convex generalized linear models*, the building block of modern machine learning. These estimators include the most basic and most widely used models in statistics and machine learning, notably the ridge regression,

logistic regression and the LASSO. One can also consider the ensembling of a finite number K of predictors, which represents the simplest instance of a neural network, for which the optimization problem becomes

$$\hat{\mathbf{W}} \in \inf_{\mathbf{W} \in \mathbb{R}^{d \times K}} \quad L(\mathbf{X}\mathbf{W}, \mathbf{y}) + r(\mathbf{W}) \tag{1.30}$$

$$\text{such that} \quad \mathbf{y} = f_0(\mathbf{X}\mathbf{W}_0) \tag{1.31}$$

where $\mathbf{W}^0$ is now in $\mathbb{R}^{d \times K}$, and f_0 is typically a function of the form $f_0(\mathbf{X}\mathbf{W}_0)_i = \phi(\frac{1}{K}\sum_{k=1}^{K} \mathbf{w}_k^\top \mathbf{x}_i)$ for some function $\phi : \mathbb{R} \to \mathbb{R}$ representing the action of f_0 on each sample. Numerous works characterized the asymptotic properties of such estimators for different instances of Eq.(1.28) in the case where $\mathbf{X}$ has i.i.d. $\mathcal{N}(0, \frac{1}{d})$ elements, see [29, 282, 82, 281, 204] for instance, and the related works sections of subsequent chapters for more references. Although these results led to a better understanding of some important building blocks, the restriction to i.i.d. Gaussian matrices drastically limits their practical usage, notably from the point of view of feature maps, which are fundamental to understand realistic machine learning scenarios.

1.6 Goal of the present work and technical challenges

How can we add realisic structure to models of empirical risk minimization while keeping exactly solvable problems ? A natural extension to the i.i.d. Gaussian design case is to add a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, which can represent the covariance operator of a given kernel, learnt features from a neural network or simply the original data. The simplest instance of exactly solvable empirical risk minimization is ridge regression with linear ground truth, which reads

$$\hat{\mathbf{w}} \in \inf_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\| + \frac{\lambda_2}{2}\|\mathbf{w}\|_2^2 \tag{1.32}$$

$$\text{where} \quad \mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\epsilon} \tag{1.33}$$

where the ground truth $\mathbf{w}_0$, and noise vector $\boldsymbol{\epsilon}$ have i.i.d. centered subgaussian coordinates with respective variance τ_0, Δ_0 and are mutually independent and independent from the design matrix $\mathbf{X}$. Further assume that the dimensions n, d go to infinity with a finite ratio α. In particular, the squared elements of $\mathbf{w}_0$ and $\boldsymbol{\epsilon}$ are subexponential and we may apply Bernstein's inequality [287] to obtain

$$\frac{1}{d}\|\mathbf{w}_0\|_2^2 \xrightarrow[d\to\infty]{a.s.} \tau_0 \quad \frac{1}{n}\|\boldsymbol{\epsilon}\|_2^2 \xrightarrow[n\to\infty]{a.s.} \Delta_0 \tag{1.34}$$

For strictly positive λ_2, the solution is unique and reads

$$\hat{\mathbf{w}} = (\mathbf{X}\top\mathbf{X} + \lambda_2 \mathbf{I}_d)^{-1} \mathbf{X}^\top \mathbf{y} \tag{1.35}$$

Let $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{X}$, where $\mathbf{U} \in \mathbb{O}_n, \mathbf{V} \in \mathbb{O}_d$ are orthogonal matrices, and $\mathbf{S} \in \mathbb{R}^{n \times d}$ contains the singular values of $\mathbf{X}$. Using the orthogonality of the matrices $\mathbf{U}, \mathbf{V}$ and the expression for the response vector $\mathbf{y}$, we may rewrite the solution as

$$\hat{\mathbf{w}} = \mathbf{V}\left(\left(\mathbf{S}^\top\mathbf{S} + \lambda_2 \mathbf{I}_d\right)^{-1} \mathbf{S}^\top\mathbf{S}\mathbf{V}^\top\mathbf{w}_0 + \left(\mathbf{S}^\top\mathbf{S} + \lambda_2 \mathbf{I}_d\right)^{-1} \mathbf{S}^\top\mathbf{U}^\top\boldsymbol{\epsilon}\right) \tag{1.36}$$

The average mean-squared error can then be written

$$\begin{aligned}&\frac{1}{d}\|\hat{\mathbf{w}}-\mathbf{w}_0\|_2^2=\\&\frac{1}{d}\left(\left(\left(\mathbf{S}^\top\mathbf{S}+\lambda_2\mathbf{I}_d\right)^{-1}\mathbf{S}^\top\mathbf{S}-\mathbf{I}_d\right)\mathbf{V}^\top\mathbf{w}_0+\left(\mathbf{S}^\top\mathbf{S}+\lambda_2\mathbf{I}_d\right)^{-1}\mathbf{S}^\top\mathbf{U}^\top\boldsymbol{\epsilon}\right)^\top\\&\left(\left(\left(\mathbf{S}^\top\mathbf{S}+\lambda_2\mathbf{I}_d\right)^{-1}\mathbf{S}^\top\mathbf{S}-\mathbf{I}_d\right)\mathbf{V}^\top\mathbf{w}_0+\left(\mathbf{S}^\top\mathbf{S}+\lambda_2\mathbf{I}_d\right)^{-1}\mathbf{S}^\top\mathbf{U}^\top\boldsymbol{\epsilon}\right)\end{aligned}\tag{1.37}$$

Assuming the eigenvalues and eigenvectors of $\mathbf{X}$ verify the required conditions for the quadratic forms to concentrate, see [78] and references therein, using the distributional assumptions on $\mathbf{w}_0$ and $\boldsymbol{\epsilon}$, we may expect a result of the form

$$\frac{1}{d}\|\hat{\mathbf{w}}-\mathbf{w}_0\|_2^2\xrightarrow[n,d\to\infty]{w.h.p.}\mathbb{E}\left[\frac{\Delta_0\lambda_{\mathbf{S}^\top\mathbf{S}}+\tau_0\lambda_2^2}{\left(\lambda_{\mathbf{S}^\top\mathbf{S}}+\lambda_2\right)^2}\right]\tag{1.38}$$

where $\lambda_{\mathbf{S}^\top\mathbf{S}}=\lambda_{\mathbf{X}^\top\mathbf{X}}$ is a random variable distributed according to the *limiting spectral density* $\lim_{d\to\infty}\frac{1}{d}\sum_{i=1}^d\delta(\lambda-\lambda_{\mathbf{X}^\top\mathbf{X},i})$. The study of sample covariance matrices $\mathbf{X}\top\mathbf{X}$ and more specifically their eigenvalue distributions is the core objective of random matrix theory, which is one way to study random design machine learning problems. In particular, the limiting spectral density of Gaussian covariate matrices of the form $\mathbf{Z}\Sigma\mathbf{Z}^\top$ where $\mathbf{Z}$ has i.i.d. Gaussian elements with variance $\frac{1}{d}$ and $\Sigma\in\mathbb{S}_d^{++}$ is positive definite, with a spectral density that converges to a distribution with compact support has been the subject of intense srcutiny since the seminal work of Marcenko and Pastur [189]. The concentration properties of related quadratic forms appearing in ridge regression problems have been studied in [123, 159, 78] among others. The problem of ridge regression presented above is solved using tools from random matrix theory in [78], where the MSE and average test error are expressed in terms of the Stieltjes transform of $\mathbf{X}$. Further derivations, using the replica method, and comparisons with real data scenarios are given in [45], showing that adding a covariance matrix to the initial i.i.d. Gaussian design is meaningful and gives insight into realistic scenarios. Note that here, we have only provided an expression for a single observable of the estimator $\hat{\mathbf{w}}$, rather than a complete description of its asymptotic distribution in terms of simpler, decoupled components. We will show how to do so on all models studied in the subsequent chapters, notably revisiting the present ridge-regression with arbitrary bounded covariance in part II and III.

The problem of moving beyond the ridge regression setting is that there is no closed form for the estimator $\hat{\mathbf{w}}$. Indeed, the optimality condition of problem (1.28) reads, for differentiable loss and regularisation,

$$\mathbf{X}^\top\nabla L(\mathbf{X}\hat{\mathbf{w}},\mathbf{y})+\nabla r(\hat{\mathbf{w}})=0.\tag{1.39}$$

Which does not seem, at first sight, solvable using tools from random matrix theory. One of the great benefits of the replica method is that non-linearities going beyond ridge regression can be treated straighforwardly, see e.g. [154, 13, 188] among other examples which will be given throughout this manuscript. In that sense, what are the corresponding rigorous mathematical tools that enable to study the asymptotic behaviour of optimization problems beyond ridge regression ? Let us briefly describe four of these methods. The first one is the Guerra-Toninelli interpolation which is based on building an interpolating path between the initial Hamiltonian and the decoupled one initially obtained from the replica prediction. Although this method is quite powerful and has

led to groundbreaking results on complex models such as the Sherrington-Kirkpatrick hamiltonian [279, 221], it appears restricted to the Bayes-optimal setting for inference problems [18]. Adaptations of this method have been applied to various inference problems but, to the best of our knowledge, no results for empirical risk minimization have been obtained. Another method is the cavity method, which has been described in section 1.4 from the theoretical physics viewpoint and can be made rigorous on various models. It rests on the comparison of a system with n particles to a system with $n+1$ particles, which leads to self-consistent equations in the large n limit. Here again this method has been applied sucessfully to Bayes-optimal problems [163] but also to convex generalized linear models [89]. Extending the results of this method to non-separable problems however, notably those obtained by introducing covariance matrices, is not always straightforward. This leads us to the two methods that will be discussed and used in this thesis. The first one is based on convex Gaussian comparison inequalities, in the form which appeared in the study of penalized linear regression [273, 282]. The second one is based on iterative Gaussian conditioning arguments, in the form that initially appeared in the context of the rigorous study of approximate message-passing algorithms [28, 42]. Now that enough context and motivation has been given, showing the importance of the high-dimensional asymptotics approach to machine learning along with the main technical challenges that the current endeavor brings, we dive into the mathematics that are necessary to move forward.

1.7 Overview of the technical tools

The purpose of this section is to provide insights into the core technical tools that underly the results presented in this thesis. We start with notions in concentration of measures and convex analysis that will be used repeatedly in the following chapters, before presenting convex Gaussian comparison inequalities, and illustrating their use on a simple example. We will then move to Gaussian iterative conditioning, which enables to obtain asymptotically exact decoupled models for optimization algorithms involving random matrices, and describe how they can be used to study convex generalized linear estimators. We stress that this section is not meant to be exhaustive : several notions will not be reminded (subdifferentials, conjugate of a convex function, subgaussian random variables, ...), and intermediate steps that do not carry significant importance will not be detailed. The material presented here is intended to provide the core objects and proof ideas that we will build upon, and why they ultimately allow us to reach our goals. Complete and fully rigorous proofs on more complex models will be given in the subsequent chapters.

1.7.1 Elements of concentration of measure

As is common in disordered systems and statistics, we will mainly consider Gaussian design matrices with different variations for their covariance structures. Most of the other quantities, such as noise or ground truth vectors will be assumed to have fast decaying tails, typically subGaussian random variables. The different functions involved such as loss, regularization or observables describing the performance of an estimator such as the mean-squared error, will be assumed to be *pseudo-Lipschitz* [28, 37] :

Definition 1 (Pseudo-Lipschitz function)**.** *For $k \in \mathbb{N}^*$ and any $N, m, q \in \mathbb{N}^*$ where k, q do not depend on N, m, a function $\Phi : \mathbb{R}^{N\times q} \to \mathbb{R}^{m\times q}$ is said to be* pseudo-Lipschitz of order k *if there exists a constant L, independent on N, m such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{N\times q}$,*

$$\frac{\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|_F}{\sqrt{m}} \leqslant L\left(1 + \left(\frac{\|\mathbf{x}\|_F}{\sqrt{N}}\right)^{k-1} + \left(\frac{\|\mathbf{y}\|_F}{\sqrt{N}}\right)^{k-1}\right)\frac{\|\mathbf{x}-\mathbf{y}\|_F}{\sqrt{N}} \tag{1.40}$$

For a scalar (or low-dimensional) valued observable of iterates of an algorithm or an estimator, we will typically have $m = 1$, arbitrary N (which will ultimately be taken to infinity) and $q < +\infty$, while an update function of an algorithm will usually have artbitrary $m = N$ and $q < \infty$. The parameter q is introduced such that our framework is fit to deal with the ensembling of a finite number of predictors or any embedding that requires a matrix valued variable, as will often be the case. The definition of pseudo-Lipschitz function originally proposed in [28], which studies the dynamics of a class of approximate message passing algorithms, does not include the scaling by $\sqrt{m}, \sqrt{N}$ of definition 1, where the property is defined for $q = 1$ as

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|_2 \leqslant L\left(1 + \|\mathbf{x}\|_2^{k-1} + \|\mathbf{y}\|_2^{k-1}\right)\|\mathbf{x} - \mathbf{y}\|_2, \tag{1.41}$$

All concentration statements in [28] are presented for separable functions, and the following proposition is a consequence of their lemma 5, whose proof is based on a truncature argument.

Proposition 1 (Concentration of separable, pseudo-Lipschitz function [28])**.** *Let $\mathbf{z} \in \mathbb{R}^n$ be a random vector with i.i.d. coordinates from a distribution p_z with bounded k-th moments. Then, for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$*

$$\lim_{n\to\infty} \frac{1}{d}\sum_{i=1}^{n} \psi(z_i) \overset{a.s.}{=} \mathbb{E}[\psi(z)] \tag{1.42}$$

However, since we will consider non-separable functions, we will follow the framework of [37] which includes the scaling in the definition for $q = 1$. Combined with the Gaussian-Poincaré inequality, definition 1 allows to prove the concentration of non-linear transforms for Gaussian random vectors quite straightforwardly. Let's look at a simple example to better understand the procedure.

Proposition 2 (Gaussian Poincaré inequality [47])**.** *Let $\mathbf{z} \in \mathbb{R}^n$ be a $\mathbf{N}(0, I_n)$ random vector. Then for any continuous, weakly differentiable φ:*

$$Var[\varphi(\mathbf{z})] \leqslant c\mathbb{E}\left[\|\nabla\varphi(\mathbf{z})\|_2^2\right] \tag{1.43}$$

We then have the following concentration result, a straightforward extension to $q > 1$ of lemma C.8 from [37]

Lemma 1. *Let $\mathbf{Z} \sim \mathbf{N}(0, \boldsymbol{\kappa}\otimes\mathbf{I}_N)$ where $\boldsymbol{\kappa} \in \mathcal{S}_q^+$ has bounded operator norm. Let $\Phi_N : \mathbb{R}^{N\times q} \to \mathbb{R}$ be a sequence of random functions, independent of $\mathbf{Z}$, such that $\mathbb{P}(\mathcal{E}_N) \to 1$ as $N \to \infty$, where $\mathcal{E}_N$ is the event that Φ_N is pseudo-Lipschitz of (deterministic) order k with (deterministic) pseudo-Lipschitz constant L. Then $\Phi_N(\mathbf{Z}) \overset{\mathrm{P}}{\simeq} \mathbb{E}[\Phi_N(\mathbf{Z})]$.*

Proof. First, it is straightforward to see that

$$\Phi_N(\mathbf{Z}) = \Phi_N(\tilde{\mathbf{Z}}\boldsymbol{\kappa}^{1/2}) = \tilde{\Phi}_N(\tilde{\mathbf{Z}}) \tag{1.44}$$

where $\tilde{\mathbf{Z}} \in \mathbb{R}^{N\times q}$ is an i.i.d. standard normal matrix, and $\tilde{\Phi}_N = \Phi_N(.\boldsymbol{\kappa}^{1/2})$. Since $\|\boldsymbol{\kappa}\|_{op}$ is bounded for all N, $\tilde{\Phi}$ is also pseudo-Lipschitz of order k, with constant $L\max(\|\boldsymbol{\kappa}\|_{op}^{1/2}, \|\boldsymbol{\kappa}\|_{op}^{k/2})$. Since q is finite and independent on N, m, $\tilde{\Phi}_N$ can be considered as a pseudo-Lipschitz function acting on

a vector of size Nq with i.i.d. standard normal components. Under $\mathcal{E}_N$, using the definition of pseudo-Lipschitz functions and proposition 2:

$$\mathbb{E}_{\mathbf{Z}}\left[\|\nabla\Phi_N(\mathbf{Z})\|_2^2\right] \leqslant \frac{L^2}{Nq}\mathbb{E}_{\mathbf{Z}}\left[\left(1+2\left(\frac{1}{\sqrt{Nq}}\|\mathbf{Z}\|_2\right)^{k-1}\right)^2\right] \leqslant \frac{L^2}{Nq}C(k) \tag{1.45}$$

for a constant $C(k)$ that only depends on k. Then for any $\epsilon > 0$, there exists a constant $c > 0$, independent of N, such that:

$$\begin{aligned}\mathbb{P}\{|\Phi_N(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\Phi_N(\mathbf{Z})]| > \epsilon\} &\leqslant \mathbb{E}\{\mathbb{P}\{|\Phi_N(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\Phi_N(\mathbf{Z})]| > \epsilon\}\mathbb{I}_{\mathcal{E}_N}\} + \mathbb{P}(\bar{\mathcal{E}}_N)\\ &\leqslant \frac{\mathrm{Var}\left[\Phi_N(\mathbf{Z})\right]}{\epsilon^2} + \mathbb{P}(\bar{\mathcal{E}}_N)\\ &\leqslant \frac{L^2C(k)}{Nq\epsilon^2} + \mathbb{P}(\bar{\mathcal{E}}_N)\end{aligned} \tag{1.46}$$

where the second and third line are obtained by applying Chebyshev's inequality and proposition 2 with the variance bound evaluated at Eq.(1.45). □

The cost of the generality of this result is a weak control over the rate at which the concentration happens : we will give little interest to finite size rates in this thesis, and will generally prefer asymptotic statements. For Lipschitz functions of i.i.d. Gaussian random vectors, (not necessarily separable), usual Gaussian concentration results give an exponential tail, see e.g. [47], while [251] also provides an exponential tail for separable, pseudo-Lipschitz functions of order 2 and subgaussian inputs. Note that the proof of lemma 1 is valid for any distribution verifying a log-Sobolev inequality. A benefit of including the scaling in the definition of the pseudo-Lipschitz function is that it does not require writing the dimension explicitly each time, which will be useful in tedious derivations. However, it may not be obvious to check this property each time. Machine learning losses and regularisations are usually pseudo-Lipschitz of order 2, with losses being separable. In this regard, it is useful to note that for a scalar, pseudo-Lipschitz of order 2 function $\psi : \mathbb{R} \to \mathbb{R}$, the function $\frac{1}{n}\sum_{i=1}^n \psi(.)$ is pseudo-Lipschitz of order 2 in the sense of definition 1. The mean-squared error $\frac{1}{n}\|\hat{\mathbf{w}} - \mathbf{w}_0\|_2^2$ of instance, is pseudo-Lipschitz of order 2. Further useful results about Lipschitz functions are contained in appendix Graph-AMP and Gordon.

For a given estimator or any related quantity on which we wish to prove statements regarding its asymptotic distribution, we will write it in terms of the concentration of pseudo-Lipschitz observables of this quantity. Owing to the definition of pseudo-Lipschitz functions, two random matrices $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{n\times q}$ will have the same behaviour in the Plk sense if we can control their higher order moments and the quantity $\frac{1}{\sqrt{N}}\|\mathbf{X} - \mathbf{Z}\|_F$ converges to zero with high probability. For comparison, proving the convergence of the empirical distribution of an estimator can be done by studying the convergence of the empirical mean of bounded continuous functions of this estimator towards an expectation over a mean-field model. Convergence in the pseudo-Lipschitz sense is thus a similar statement but can be adapted to non-separable functions and includes non bounded observables commonly used in machine learning such as the mean squared error. In informal statements, we will sometimes denote $\xrightarrow[n,d\to\infty]{Plk}$ the fact that two random variables asymptotically have the same behaviour.

1.7.2 Elements of convex analysis

In this paragraph, we introduce functions that appear in convex analysis and will be used repeatedly in all the proofs regarding convex empirical risk minimization. Indeed, we will see that the cost functions and estimators defined by convex generalized linear models can be expressed using well defined objects with convenient regularity properties, the Moreau envelope and proximal operator [25, 224]. All the results presented here can be found in [25].

Definition 2 (Moreau envelope and proximal operator)**.** *Consider a proper, closed, lower semicontinuous convex function $f : \mathbb{R}^d \to \mathbb{R}$. Its Moreau envelope is defined by the optimization problem*

$$\forall\, \tau > 0, \quad \mathcal{M}_{\tau f}(\mathbf{x}) = \min_{\mathbf{z}\in\mathbb{R}^d} \left\{ f(\mathbf{z}) + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{z}\|_2^2 \right\} \tag{1.47}$$

and its proximal operator

$$\forall\, \gamma > 0, \quad prox_{\gamma f}(\mathbf{x}) = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \left\{ f(\mathbf{z}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{z}\|_2^2 \right\} \tag{1.48}$$

Owing to the convexity of f, the strong convexity and supercoercivity of the quadratic term, the optimization problem defining the Moreau envelope has a unique solution [25]. Thus the proximal operator is the unique point realizing the minimum of the Moreau envelope:

$$\mathcal{M}_{\tau f}(\mathbf{x}) = f(\mathrm{prox}_{\tau f}(\mathbf{x})) + \frac{1}{2\tau}\left\|\mathbf{x} - \mathrm{prox}_{\tau f}(\mathbf{x})\right\|_2^2 \tag{1.49}$$

Moreau envelopes have the same set of minimizers as the original function and are continuously differentiable on their domain, with derivatives:

$$\nabla_{\mathbf{x}}\mathcal{M}_{\tau f}(\mathbf{x}) = \frac{1}{\tau}(\mathbf{x} - \mathrm{prox}_{\tau f}(\mathbf{x})) \tag{1.50}$$

$$\frac{\partial}{\partial\tau}\mathcal{M}_{\tau f}(\mathbf{x}) = -\frac{1}{2\tau^2}\left\|\mathbf{x} - \mathrm{prox}_{\tau f}(\mathbf{x})\right\|_2^2 \tag{1.51}$$

They can be understood as a smoothed version of the original function f, which may be non-differentiable, such as the $l1$ norm in machine learning, while the proximity operator can be understood as a projection on the level sets of the function f. Indeed, replacing f with the indicator function of an ensemble recovers the othogonal projector on this ensemble. The expression for the gradient Eq.(1.50) shows that the proximal operator with parameter τ is also equivalent to taking a gradient step with step-size τ on the Moreau envelope with parameter τ. Furthermore, the optimality condition of the optimization problem Eq.(1.48) gives the following alternate characterization of proximity operators

$$\mathrm{prox}_{\gamma f}(\mathbf{x}) = (Id + \gamma\partial f)^{-1}(\mathbf{x}) \tag{1.52}$$

where ∂f is the subdifferential of f. This formulation is the resolvent of the subdifferential operator of f. This shows, in turn, the following equivalence

$$\mathrm{prox}_{\gamma f}(\mathbf{x}) = \mathbf{x} \quad \Longleftrightarrow \quad \mathbf{x} \in \mathrm{zer}(\partial f) \tag{1.53}$$

Additionally, proximal operators are firmly non-expansive, i.e.

$$\forall\gamma > 0, \quad \forall\mathbf{x},\mathbf{y} \in \mathcal{X}, \quad \left\|\mathrm{prox}_{\gamma f}(\mathbf{x}) - \mathrm{prox}_{\gamma f}(\mathbf{y})\right\|_2^2 \leqslant \langle \mathbf{x}-\mathbf{y}, \mathrm{prox}_{\gamma f}(\mathbf{x}) - \mathrm{prox}_{\gamma f}(\mathbf{y})\rangle, \tag{1.54}$$

which is a useful property to control the trajectories of proximal based optimization algorithms. These properties motivate the use of these operators to optimize convex functions in stable and efficient fashion, and proximal algorithms are one of the cornerstones of convex optimization, see e.g. [50, 224], with the simplest instance being the proximal-point algorithm

$$\mathbf{x}^{t+1} = \text{prox}_{\gamma f}(\mathbf{x}^t) \tag{1.55}$$

We will further discuss related algorithms in part III. Thus, even if they are defined with optimization problems, we will consider a problem to be solved once we have reached expressions involving the Moreau envelopes and proximal operators of sums of independent random variables/vectors/low-rank matrices. For example, the proximity operators of quadratic form reads

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c \quad \mathbf{A} \in \mathbb{S}_d^+, \mathbf{b}, \mathbf{c} \in \mathbb{R}^d \\ \text{prox}_{\gamma f}(\mathbf{x}) &= (\gamma \mathbf{A} + \mathbf{I}_d)^{-1}(\mathbf{x} - \gamma \mathbf{b}) \end{aligned} \tag{1.56}$$

and the proximal operator for the ℓ_1 norm is an element-wise application of the soft-thresholding operator

$$\forall 1 \leqslant i \leqslant n \quad \left[\text{prox}_{\gamma\|.\|_1}(\mathbf{x})\right]_i = \text{sign}(x_i)\max\left(0, |x_i| - \gamma\right). \tag{1.57}$$

More generally, proximal operators of usual convex functions (logistic loss, log-barrier, hinge loss, ...) are straightforward to compute and stable to evaluate numerically.

1.7.3 Gaussian comparison inequalities

We now turn to the description of a first method that can be used to decouple the measure implied by equation (1.28). Recall that by decoupling, we mean replacing the random (with extensive dimensions) matrix $\mathbf{X}$ by simpler, independent objects. We start by introducing a comparison inequality for Gaussian random processes indexed on compact sets [121, 161]:

Proposition 3. *(Gordon's inequality [121, 161]) Let $D_{\mathbf{u}} \subset \mathbb{R}^n$ and $D_{\mathbf{v}} \subset \mathbb{R}^m$ be two compact sets. Let $(X(\mathbf{u},\mathbf{v}))_{(\mathbf{u},\mathbf{v})\in D_{\mathbf{u}}\times D_{\mathbf{v}}}$ and $(Y(\mathbf{u},\mathbf{v}))_{(\mathbf{u},\mathbf{v})\in D_{\mathbf{u}}\times D_{\mathbf{v}}}$ be two centered Gaussian processes with continuous sample paths. Assume that*

$$\begin{cases} \mathbb{E}\left[X(\mathbf{u},\mathbf{v})^2\right] = \mathbb{E}\left[Y(\mathbf{u},\mathbf{v})^2\right] \quad \textit{for all} \quad (\mathbf{u},\mathbf{v}) \in D_{\mathbf{u}} \times D_{\mathbf{v}} \\ \mathbb{E}\left[X(\mathbf{u},\mathbf{v})X(\mathbf{u},\mathbf{v}')\right] \geqslant \mathbb{E}\left[Y(\mathbf{u},\mathbf{v})Y(\mathbf{u},\mathbf{v}')\right] \quad \textit{for all} \quad \mathbf{u} \in D_{\mathbf{u}}, \mathbf{v},\mathbf{v}' \in D_{\mathbf{v}} \\ \mathbb{E}\left[X(\mathbf{u},\mathbf{v})X(\mathbf{u}',\mathbf{v}')\right] \leqslant \mathbb{E}\left[Y(\mathbf{u},\mathbf{v})Y(\mathbf{u}',\mathbf{v}')\right] \quad \textit{for all} \quad \mathbf{u},\mathbf{u}' \in D_{\mathbf{u}}, \mathbf{v},\mathbf{v}' \in D_{\mathbf{v}} \textit{ s.t. } \mathbf{u} \neq \mathbf{u}' \end{cases}$$

Then for all $t \in \mathbb{R}$

$$\mathbb{P}\left(\min_{\mathbf{u}\in D_{\mathbf{u}}} \max_{\mathbf{v}\in D_{\mathbf{v}}} Y(\mathbf{u},\mathbf{v}) \leqslant t\right) \leqslant \mathbb{P}\left(\min_{\mathbf{u}\in D_{\mathbf{u}}} \max_{\mathbf{v}\in D_{\mathbf{v}}} X(\mathbf{u},\mathbf{v}) \leqslant t\right) \tag{1.58}$$

The inequality is rather intuitive : the fluctuations of Gaussian processes are governed by their covariance functions, and comparing the covariances leads to comparisons on their maxima and minima. It can then be used to obtain tight inequalitites on convex-concave minmax problems [282, 281].

Corollary 1. *(Convex Gaussian minmax theorem [273, 282, 281]) Let $D_{\mathbf{u}} \subset \mathbb{R}^n$ and $D_{\mathbf{v}} \subset \mathbb{R}^m$ be two compact sets and let $Q : D_{\mathbf{u}} \times D_{\mathbf{v}} \to \mathbb{R}$ denote a continuous function. Let $\mathbf{G} \in \mathbb{R}^{n\times m} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_m)$ be independent standard Gaussian vectors. Define the functions*

$$\begin{cases} C^*(\mathbf{G}) = \min_{\mathbf{u}\in D_{\mathbf{u}}} \max_{\mathbf{v}\in D_{\mathbf{v}}} \mathbf{v}^T\mathbf{G}\mathbf{u} + Q(\mathbf{u},\mathbf{v}) \\ L^*(\mathbf{g},\mathbf{h}) = \min_{\mathbf{u}\in D_{\mathbf{u}}} \max_{\mathbf{v}\in D_{\mathbf{v}}} \|\mathbf{v}\|\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|\mathbf{h}^T\mathbf{v} + Q(\mathbf{u},\mathbf{v}) \end{cases}$$

Then we have:

- *For all $t \in \mathbb{R}$*
$$\mathbb{P}\left(C^*(\mathbf{G}) \leqslant t\right) \leqslant 2\mathbb{P}\left(L^*(\mathbf{g},\mathbf{h}) \leqslant t\right)$$
- *If $D_{\mathbf{u}}, D_{\mathbf{v}}$ are convex sets and Q is convex-concave, then for all $t \in \mathbb{R}$*
$$\mathbb{P}\left(C^*(\mathbf{G}) \geqslant t\right) \geqslant 2\mathbb{P}\left(L^*(\mathbf{g},\mathbf{h}) \geqslant t\right)$$

In particular, for all $\mu \in \mathbb{R}$, $t > 0$,

$$\mathbb{P}\left(|C^*(\mathbf{G}) - \mu| \geqslant t\right) \leqslant 2\mathbb{P}\left(|L^*(\mathbf{g},\mathbf{h}) - \mu| \geqslant t\right) \tag{1.59}$$

This corollary, obtained by verifying the covariance conditions of proposition 3 for the Gaussian processes defining $C(\mathbf{G}), L(\mathbf{g},\mathbf{h})$ allows to study the concentration properties of convex-concave problems involving a dense random matrix by means of a simpler problem involving only two independent random vectors. In what follows, we will present a variant of the core derivation of the result from [281] which studies convex penalized generalized regression. Several technical steps are not reproduced, and pointers to the original paper will be given for their proofs and the full set of assumptions. We focus instead on the actual "algebra" that corollary 1 enables, and where ommited steps can be made rigorous for intuitive reasons.

An example : convex penalized regression

Consider the following regression problem

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{d}\left\{L(\mathbf{y} - \mathbf{A}\mathbf{w}) + r(\mathbf{w})\right\} \tag{1.60}$$

$$\text{where} \quad \mathbf{y} = \mathbf{A}\mathbf{w}_0 + \boldsymbol{\epsilon} \tag{1.61}$$

where L, r are convex functions, $\mathbf{A} \in \mathbb{R}^{n\times d}$ has i.i.d. $\mathcal{N}(0, 1/d)$ elements, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Delta\mathbf{I}_n)$ and $\mathbf{w}_0$ is sampled i.i.d. from a subgaussian distribution. Omitting the $\frac{1}{d}$ scaling for now, we may equivalently wirte the optimization problem as

$$\begin{aligned} &\min_{\mathbf{x}} L(\mathbf{y} - \mathbf{A}\mathbf{w}) + r(\mathbf{w}) \\ &= \min_{\mathbf{w}} L(\mathbf{A}(\mathbf{w}_0 - \mathbf{w}) + \boldsymbol{\epsilon}) + r(\mathbf{w}) \\ &= \min_{\mathbf{e}} L(\boldsymbol{\epsilon} - \sqrt{d}\mathbf{A}\mathbf{e}) + r(\mathbf{w}_0 + \sqrt{d}\mathbf{e}) \end{aligned}$$

where in the last lign we introduced the variable $\mathbf{e} = \frac{\mathbf{w}-\mathbf{w}_0}{\sqrt{d}}$. Reformulating the problem with an auxiliary variable $\mathbf{z} = \boldsymbol{\epsilon} - \sqrt{d}\mathbf{A}\mathbf{e}$, we can rewrite the objective cost with the corresponding Lagrange multiplier $\boldsymbol{\lambda}$

$$\begin{aligned} &\min_{\mathbf{e},\mathbf{z}} \quad L(\mathbf{z}) + r(\mathbf{w}_0 + \sqrt{d}\mathbf{e}) \quad \text{s.t. } \mathbf{z} = \boldsymbol{\epsilon} - \sqrt{d}\mathbf{A}\mathbf{e} \\ &\iff \min_{\mathbf{e},\mathbf{z}} \max_{\boldsymbol{\lambda}} L(\mathbf{z}) + r(\boldsymbol{\epsilon} + \sqrt{d}\mathbf{e}) + \boldsymbol{\lambda}^T \left(\mathbf{z} - \boldsymbol{\epsilon} + \sqrt{d}\mathbf{A}\mathbf{e}\right) \end{aligned}$$

Under appropriate growth conditions on the functions L, r (see assumption 1(b) from [281]), the compactness requirements to apply corollary 1 can be met (along with the convexity-concavity requirements which are straightforwardly verified), and we way now write the corresponding decoupled optimization problem:

$$\min_{\mathbf{e},\mathbf{z}} \max_{\boldsymbol{\lambda}} \|\boldsymbol{\lambda}\|_2 \mathbf{g}^T\mathbf{e} + \|\mathbf{e}\|_2 \mathbf{h}^T\boldsymbol{\lambda} + L(\mathbf{z}) + r(\mathbf{w}_0 + \sqrt{d}\mathbf{e}) + \boldsymbol{\lambda}^T(\mathbf{z} - \boldsymbol{\epsilon}) \tag{1.62}$$

where $\mathbf{g} \in \mathbb{R}^d$ and $\mathbf{h} \in \mathbb{R}^n$ are independent vectors with i.i.d. standard normal coordinates. Introducing the convex conjugate of r with dual variable $\boldsymbol{\mu}$, $r(\mathbf{w}_0 + \sqrt{d}\mathbf{e}) = \max_{\boldsymbol{\mu}}\{\boldsymbol{\mu}^T(\mathbf{w}_0 + \sqrt{d}\mathbf{e}) - r^*(\boldsymbol{\mu})\}$, which gives, reintroducing the scaling by $\frac{1}{n}$:

$$\min_{\mathbf{e},\mathbf{z}} \max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \frac{1}{d}\left(\|\boldsymbol{\lambda}\|_2 \mathbf{g}^T\mathbf{e} + \|\mathbf{e}\|_2 \mathbf{h}^T\boldsymbol{\lambda} + L(\mathbf{z}) + \boldsymbol{\lambda}^T(\mathbf{z} - \boldsymbol{\epsilon}) + \boldsymbol{\mu}^T(\mathbf{w}_0 + \sqrt{n}\mathbf{e}) - r^*(\boldsymbol{\mu})\right)$$

Here, due to the fact that $\mathbf{g}, \mathbf{h}$ may be negative, the problem is not convex-concave anymore. However, it is shown in [281] that, since this optimization problem is equivalent to a convex one, we may invert the order of minimization as if strong duality appplied. Then, letting $\alpha = \|\mathbf{e}\|_2 = \left\|\frac{\mathbf{w}-\mathbf{w}_0}{\sqrt{d}}\right\|_2$, and performing the optimization step on $\mathbf{e}$ which is now a linear optimization problem, we reach

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \min_{\alpha,\mathbf{z}} -\frac{\alpha}{d}\left\|\|\boldsymbol{\lambda}\|_2 \mathbf{g} + \sqrt{d}\boldsymbol{\mu}\right\|_2 + \frac{\alpha}{d}\mathbf{h}^T\boldsymbol{\lambda} + \frac{1}{d}L(\mathbf{z}) + \frac{1}{d}\boldsymbol{\lambda}^T(\mathbf{z} - \boldsymbol{\epsilon}) + \frac{1}{d}\boldsymbol{\mu}^T\mathbf{w}_0 - \frac{1}{d}r^*(\boldsymbol{\mu}).$$

letting $\beta = \frac{1}{\sqrt{d}}\|\boldsymbol{\lambda}\|_2$ (the problem is now convex so we may invert the order of minimization) and performing the linear optimization on $\boldsymbol{\lambda}$ gives the equivalent problem:

$$\max_{\beta,\boldsymbol{\mu}} \min_{\alpha,\mathbf{z}} -\frac{\alpha}{\sqrt{d}}\|\beta\mathbf{g} + \boldsymbol{\mu}\|_2 + \frac{\beta}{\sqrt{d}}\|\alpha\mathbf{h} + \mathbf{z} - \boldsymbol{\epsilon}\|_2 + \frac{1}{d}L(\mathbf{z}) + \frac{1}{d}\boldsymbol{\mu}^T\mathbf{w}_0 - \frac{1}{d}r^*(\boldsymbol{\mu})$$

We then introduce the following representation of the norm $\|\mathbf{t}\|_2 = \inf_{\tau>0} \frac{\tau}{2} + \frac{\|\mathbf{t}\|_2^2}{2}$, reaching

$$\max_{\beta,\boldsymbol{\mu},\tau_2>0} \min_{\alpha,\mathbf{z},\tau_1>0} -\frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} + \frac{1}{d}\left(-\frac{\alpha}{2\tau_2}\|\beta\mathbf{g} + \boldsymbol{\mu}\|_2^2 + \frac{\beta}{2\tau_1}\|\alpha\mathbf{h} + \mathbf{z} - \boldsymbol{\epsilon}\|_2^2 + g(\mathbf{z}) + \boldsymbol{\mu}^T\mathbf{w}_0 - f^*(\boldsymbol{\mu})\right)$$

Completing the squares in $\boldsymbol{\mu}$ and $\mathbf{w}_0$, inverting the sign in front of $\mathbf{g}$ (centered Gaussian) for convenience yields:

$$-\frac{\alpha}{2\tau_2}\|\boldsymbol{\mu} - \beta\mathbf{g}\|_2^2 = -\frac{\alpha}{2\tau_2}\left\|\boldsymbol{\mu} - \left(\frac{\tau_2}{\alpha}\mathbf{w}_0 + \beta\mathbf{g}\right)\right\|_2^2 - \mathbf{x}_0^T(\boldsymbol{\mu} - \beta\mathbf{g}) + \frac{\tau_2}{2\alpha}\|\mathbf{w}_0\|_2^2 \tag{1.63}$$

which gives

$$\max_{\beta,\tau_2>0}\min_{\alpha,\tau_1>0} -\frac{\alpha\tau_2}{2}+\frac{\beta\tau_1}{2}+\frac{\beta}{d}\mathbf{g}^T\mathbf{w}_0+\frac{\tau_2}{2d\alpha}\|\mathbf{w}_0\|_2^2 \\ +\frac{1}{d}\min_{\mathbf{z}}\left\{\frac{\beta}{2\tau_1}\|\mathbf{z}-\boldsymbol{\epsilon}-\alpha\mathbf{h}\|_2^2+g(\mathbf{z})\right\}-\frac{1}{d}\min_{\boldsymbol{\mu}}\left\{\frac{\alpha}{2\tau_2}\left\|\boldsymbol{\mu}-(\frac{\tau_2}{\alpha}\mathbf{w}_0+\beta\mathbf{g})\right\|_2^2+f^*(\boldsymbol{\mu})\right\}$$

Using the definition of Moreau envelopes and expression for Moreau envelopes of conjugate pairs (see [25] or the proofs for chapter 7):

$$\max_{\beta,\tau_2>0}\min_{\alpha,\tau_1>0} -\frac{\alpha\tau_2}{2}+\frac{\beta\tau_1}{2}+\frac{\alpha\beta^2}{2d\tau_2}\mathbf{g}^T\mathbf{g}+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\beta}g(.)}(\boldsymbol{\epsilon}+\alpha\mathbf{h})+\frac{1}{d}\mathcal{M}_{\frac{\alpha}{\tau_2}f(.)}\left(\mathbf{w}_0+\frac{\beta\alpha}{\tau_2}\mathbf{g}\right)$$

Assuming the loss and regularization functions f, g are separable and pseudo-Lipschitz of order 2, the following pointwise convergence occurs when $n, d \to \infty$ with $n/d = \gamma > 0$ for finite γ:

$$-\frac{\alpha\tau_2}{2}+\frac{\beta\tau_1}{2}+\frac{\alpha\beta^2}{2d\tau_2}\mathbf{g}^T\mathbf{g}+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\beta}g(.)}(\boldsymbol{\epsilon}+\alpha\mathbf{h})+\frac{1}{d}\mathcal{M}_{\frac{\alpha}{\tau_2}f(.)}\left(\mathbf{w}_0+\frac{\beta\alpha}{\tau_2}\mathbf{g}\right)\xrightarrow[n,d\to\infty]{a.s.} \\ -\frac{\alpha\tau_2}{2}+\frac{\beta\tau_1}{2}+\frac{\alpha\beta^2}{2\tau_2}+\gamma\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\beta}g(.)}(\epsilon+\alpha h)\right]+\mathbb{E}\left[\mathcal{M}_{\frac{\alpha}{\tau_2}f(.)}\left(w_0+\frac{\beta\alpha}{\tau_2}g\right)\right]$$

Uniform convergence can be proven using the convexity assumption, leading to the convergence of the extremum as well. In the high-dimensional proportional limit, the optimal cost function thus reduces to the scalar optimization problem.

$$\max_{\beta,\tau_2>0}\min_{\alpha,\tau_1>0} -\frac{\alpha\tau_2}{2}+\frac{\beta\tau_1}{2}+\frac{\alpha\beta^2}{2\tau_2}+\gamma\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\beta}g(.)}(\epsilon+\alpha h)\right]+\mathbb{E}\left[\mathcal{M}_{\frac{\alpha}{\tau_2}f(.)}\left(w_0+\frac{\beta\alpha}{\tau_2}g\right)\right] \tag{1.64}$$

The replica method, when applied to the same problem, gives the same result [12]. Using the differentiability results for Moreau envelopes presented in section 1.7.2, we can write the self-consistent system of non-linear equations, involving the proximal operators of f, g, solving the optimization problem (1.64). Once again using the properties of Moreau envelopes, the optimization problem (1.64) can be shown to be strictly convex-concave [281], proving the uniqueness of the optimal quarduplet $\alpha^*, \beta^*, \tau_1^*, \tau_2^*$. Corollary 3 then gives the following result

$$\lim_{d\to\infty}\frac{1}{d}\|\hat{\mathbf{w}}-\mathbf{w}_0\| \overset{a.s.}{=} \alpha^* \tag{1.65}$$

Using a stronger version of corollary 1 [204, 57, 176], one can actually prove a statement regarding the asymptotic distribution of $\hat{\mathbf{w}}$. Recall that the proximal operator is the unique solution to the minimization problem defining the Moreau envelope. In the case presented here, for any pseudo-Lipschitz function of order 2 $\psi : \mathbb{R} \to \mathbb{R}$

$$\lim_{n,d\to\infty}\frac{1}{d}\sum_{i=1}^{d}\psi(\hat{w}_i)=\mathbb{E}\left[\psi\left(\text{prox}_{\frac{\alpha^*}{\tau_2^*}f(.)}\left(w_0+\frac{\beta^*\alpha^*}{\tau_2^*}g\right)\right)\right] \tag{1.66}$$

where $w_0 \sim p_{w_0}$ and $g \sim \mathcal{N}(0,1)$. The result thus becomes very concrete : the estimator is asymptotically distributed as the ground truth w_0 with an added noise $\frac{\beta^*\alpha^*}{\tau_2^*}g$, to which the proximal operator of the regularisation is applied. In the case of an ℓ_1 penalty, we see that the soft-thresholding operator will put to zero coefficients that are smaller than a value uniquely prescribed by the solution

of the scalar optimization problem (1.64), on a noisy vector centred arount the ground-truth. The Gordon comparison theorem approach allows to straightforwardly turn the study of the asymptotic mean-squared error into a scalar optimization problem obtained by simplifying a decoupled problem with convenient convex optimization results. Moreover, several of the intermediate technical steps, such as verifying the compactness of the feasibility set or inverting the order of minimization when the problem presents negative weighted norms, remain similar for a variety of convex problems going beyond generalized regression. This makes this framework quite appealing, and we will use this approach in chapter 7 to study a more complex model.

Three main hurdles can be found to this approach : although we obtain the asymptotic value of the mean squared error, we do not obtain a full characterisation of asymptotic distribution of $\hat{\mathbf{w}}$. In [204], further inequalities are proven to obtain the full characterization of the asymptotic distribution of the LASSO with i.i.d. Gaussian matrices as discussed above, along with the finite size rates. The approach is quite tedious, and we will sketch in chapter 7 how to use it for a more complex problem. The other issue is that, for matrix valued estimators, the optimization problem involving the dense random matrix $\mathbf{G} \in \mathbb{R}^{n \times d}$ cannot be decoupled in the same form as in corollary 1. This prevents the Gordon approach to be used for ensembling or multiclass problems, which are formulated in terms of a (low-rank) matrix estimator. Finally, we mentioned in the introduction that the dynamics of several descent algorithms would be of interest, for which the Gordon approach is not well-suited. We thus turn to the method that we will use the most : iterative Gaussian conditioning.

1.7.4 Iterative Gaussian conditioning

The method that we will now present arose in the rigorous study of approximate message passing algorithms, notably [28, 42], and rests on a fundamental property of the Gaussian distribution : orthogonality and independence are equivalent for Gaussian random variables, and independence can be entirely characterized by their covariance matrices. Thus, in the Gaussian case, computing conditional expectations that are initially defined as orthogonal projections on an infinite dimensional space, becomes possible with finite dimensional projections. Intuitively, consider an $n \times d$ random matrix $\mathbf{A}$ with i.i.d. standard normal entries and a deterministic d-dimensional vector $\mathbf{w}$. We can then decompose $\mathbf{A}$ as

$$\mathbf{A} \overset{d}{=} \mathbf{A}\mathbf{P}_{\mathbf{x}} + \tilde{\mathbf{A}}\mathbf{P}_{\mathbf{x}}^{\perp} \tag{1.67}$$

where $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$, $\mathbf{P}_{\mathbf{x}} = \frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|_2^2}$ is the orthogonal projector on $\mathbf{x}$ and $\mathbf{P}_{\mathbf{x}}^{\perp} = \mathbf{I}_d - \mathbf{P}_{\mathbf{x}}$. A more generic statement can be found in [28] :

Lemma 2 (Gaussian matrices under linear constraints)**.** *Consider an $n \times d$ random matrix $\mathbf{A}$ with i.i.d. standard normal elements, and deterministic matrices $\mathbf{Q} \in \mathbb{R}^{d \times k}$, $\mathbf{M} \in \mathbb{R}^{n \times k}$, such that the projectors $\mathbf{P}_{\mathbf{M}} = \mathbf{M}\left(\mathbf{M}^\top\mathbf{M}\right)^{-1}\mathbf{M}^\top$ and $\mathbf{P}_{\mathbf{Q}} = \mathbf{Q}\left(\mathbf{Q}^\top\mathbf{Q}\right)^{-1}\mathbf{Q}^\top$ onto the subspaces spanned by the columns of $\mathbf{Q}$ and $\mathbf{M}$ exist. Then the conditional distribution of $\mathbf{A}$ given the random variables $\mathbf{A}\mathbf{Q}, \mathbf{A}^\top\mathbf{M}$ may be written*

$$\mathbf{A}|_{\mathbf{A}\mathbf{Q}, \mathbf{A}^\top\mathbf{M}} = \mathbf{P}_{\mathbf{M}}\mathbf{A} + \mathbf{A}\mathbf{P}_{\mathbf{Q}} - \mathbf{P}_{\mathbf{M}}\mathbf{A}\mathbf{P}_{\mathbf{Q}} + \mathbf{P}_{\mathbf{M}}^{\perp}\tilde{\mathbf{A}}\mathbf{P}_{\mathbf{Q}}^{\perp} \tag{1.68}$$

where $\mathbf{P}_{\mathbf{M}}^{\perp} = \mathbf{I}_n - \mathbf{P}_{\mathbf{M}}$, $\mathbf{P}_{\mathbf{Q}}^{\perp} = \mathbf{I}_d - \mathbf{P}_{\mathbf{Q}}$, and $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$.

As an example, let us study the dynamics of the gradient descent corresponding to the minimization problem

$$\hat{\mathbf{w}} \in \inf_{\mathbf{w}\in\mathbb{R}^d} \quad f(\mathbf{A}\mathbf{w}) \tag{1.69}$$

We will start with a sample splitting assumption, and then move to the generic case.

Gradient descent with sample splitting We start with sample splitting, i.e. a new batch of data is used at each iteration:

$$\forall t \in \mathbb{N}^* \quad \mathbf{w}^{t+1} = \mathbf{w}^t - \gamma^t (\mathbf{A}^t)^\top \nabla \mathbf{f}(\mathbf{A}^t \mathbf{w}^t) \tag{1.70}$$

where, for any $t \in \mathbb{N}$, $\mathbf{A}^t \in \mathbb{R}^{n\times d}$ is a matrix with i.i.d. Gaussian elements and variance $1/d$ independent on all other $\{\mathbf{A}^i\}_{i\neq t}$, $\gamma^t \in \mathbb{R}$ is a scalar step-size and $\mathbf{f}$ is a twice differentiable, deterministic function with pseudo-Lipschitz gradient $\nabla\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. We also assume that $\mathbf{f}$ is separable, with an elementwise operation f. The iteration is initialized with $\mathbf{w}^0 \in \mathbb{R}^d$, a random vector independent on $\mathbf{A}$ with i.i.d. subGaussian elements. Finally, assume that, when the dimensions of the problem are taken to infinity, we do so with finite ratio $\alpha = n/d$. Starting at $t = 0$, we condition equation (1.70) on (the sigma algebra generated by) $\mathbf{w}^0, \mathbf{A}^0\mathbf{w}^0$, and obtain

$$\mathbf{w}^1|_{\mathbf{w}^0,\mathbf{A}^0\mathbf{w}^0} = \mathbf{w}^0 - \gamma^0 \left(\mathbf{A}^0\mathbf{P}_{\mathbf{w}^0} + \tilde{\mathbf{A}}^0\mathbf{P}^\perp_{\mathbf{w}^0}\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \tag{1.71}$$

$$= \mathbf{w}^0 - \gamma^0 \mathbf{w}^0 \frac{1}{\|\mathbf{w}^0\|_2^2} \left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) - \gamma^0 \mathbf{P}^\perp_{\mathbf{w}^0}\tilde{\mathbf{A}}^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \tag{1.72}$$

Owing to the sample splitting assumption, the vector $\mathbf{A}^0\mathbf{w}^0$ has i.i.d. entries distributed according to $\mathcal{N}(0, \frac{1}{d}\|\mathbf{w}^0\|_2^2)$. We can then write

$$\frac{1}{\|\mathbf{w}^0\|_2^2} \left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) = \frac{1}{\frac{1}{d}\|\mathbf{w}^0\|_2^2} \frac{1}{d} \left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \tag{1.73}$$

The term $\frac{1}{d}\left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)$ is a scalar valued, pseudo-Lipschitz function of $\mathbf{A}^0\mathbf{w}^0$, and the subgaussian assumption on $\mathbf{w}^0$ ensures that the quantity $\frac{1}{d}\|\mathbf{w}^0\|_2^2$ converges almost surely to a finite, deterministic quantity. We can thus use lemma 1, the continuous mapping theorem (in the form of Slutsky's lemma), and Stein's lemma to show that

$$\frac{1}{\|\mathbf{w}^0\|_2^2} \left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \overset{\mathrm{P}}{\simeq} \alpha\mathbb{E}\left[f''(z^0)\right] \tag{1.74}$$

where $z^0 \sim \mathcal{N}(0, \rho^0)$ and we introduced $\rho^0 = \lim_{d\to\infty} \frac{1}{d}\|\mathbf{w}^0\|_2^2$. Turning to the part orthogonal to $\mathbf{w}^0$ and using the fact that the projector $\mathbf{P}_{\mathbf{w}^0}$ is of rank 1, the elements of $\tilde{\mathbf{A}}$ have variance $\frac{1}{d}$ and $\|\mathbf{w}^0\|_2^2$ is of order d, lemma 21 shows that

$$\frac{1}{\sqrt{d}} \left\|\mathbf{P}^\perp_{\mathbf{w}^0}\tilde{\mathbf{A}}^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) - (\tilde{\mathbf{A}}^0)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\right\|_2 \overset{\mathrm{P}}{\simeq} 0 \tag{1.75}$$

where $(\tilde{\mathbf{A}}^0)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)$ is a vector with i.i.d elements distributed as $\mathcal{N}(0, \frac{1}{d}\|\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\|_2^2)$. Once again, the function $\frac{1}{d}\|\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\|_2^2$ is scalar valued and pseudo-Lipschitz, thus lemma 1 and the continuous mapping theorem show that, for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ of order 2,

$$\frac{1}{d}\sum_{i=1}^{d} \psi(\left(\mathbf{P}^\perp_{\mathbf{w}^0}\tilde{\mathbf{A}}^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\right)_i) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\psi(u^0)\right] \tag{1.76}$$

where $u^0 \sim \mathcal{N}(0, \tau_0)$ and we have introduced $\tau_0 = \lim_{n,d\to\infty} \frac{1}{d}\|\nabla \mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\|_2^2 = \alpha\mathbb{E}\left[(f'(z^0))^2\right]$. Using these results, we may now lift the conditioning and use the definition of pseudo-Lipschitz function to recover the scalar equation describing the high-dimensional behaviour of $\mathbf{w}^1$. A straightforward induction shows that, for any $t \in \mathbb{N}$, the quantity $\frac{1}{d}\|\mathbf{w}^t\|_2^2$ is almost surely bounded, and the same conditioning argument can be applied along the sample splitting assumption to reach the following theorem

Theorem 2. *(High-dimensional dynamics of gradient descent with sample splitting) Consider the iteration Eq. (1.70) with its set of assumptions described above. Define the following discrete-time one-dimensional stochastic process, initialized with a subgaussian random variable ω^0 with variance ρ^0:*

$$\omega^{t+1} = \left(1 - \gamma^t \alpha \mathbb{E}\left[f''(z^t)\right]\right)\omega^t + \gamma^t u^t \tag{1.77}$$

where $z^t \sim \mathcal{N}(0, \rho^t)$ and $u^t \sim \mathcal{N}(0, \tau^t)$ are independent, and $\rho^t = \mathbb{E}\left[(\omega^t)^2\right]$, $\tau^t = \alpha\mathbb{E}\left[(f'(z^t))^2\right]$. Then, for any $t \in \mathbb{N}$ and any pseudo-Lipschitz function of order 2 $\psi : \mathbb{R} \to \mathbb{R}$, the following holds

$$\frac{1}{d}\sum_{i=1}^{d} \psi(w_i^t) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\psi(\omega^t)\right] \tag{1.78}$$

We have obtained a full description of the asymptotic distribution of $\mathbf{w}^t$ in terms of a scalar equation. The sample splitting assumption however, is unrealistic. Let us move to the generic case that corresponds to the usual gradient descent.

Gradient descent without sample splitting The proof becomes much more complicated without the sample splitting assumption, and the full result along with its proof, which recovers a result known as *dynamical mean-field theory* in physics, will be given in chapter 6, in which we also discuss the related litterature, both in theoretical physics and mathematics. Here we will only do the first few steps, to give a flavour of the problem, and to motivate the introduction of a stochastic correction at each time step, leading to *approximate message passing algorithms*. Let us rewrite the dynamics without the sample splitting assumption in the following way

$$\mathbf{v}^{t+1} = -\gamma^t \mathbf{A}^\top \mathbf{m}^t \tag{1.79}$$

$$\mathbf{m}^t = \nabla \mathbf{f}(\mathbf{r}^t) \tag{1.80}$$

$$\mathbf{r}^t = \mathbf{A}\sum_{k=0}^{t} \mathbf{v}^k \tag{1.81}$$

where $\mathbf{v}^t = \mathbf{w}^t - \mathbf{w}^{t-1}$ and $\mathbf{w}^{-1} = 0$. Then $\mathbf{v}^0 = \mathbf{w}^0$, assumed to be independent from $\mathbf{A}$ and sampled i.i.d. from a sub-gaussian distribution. Let's try to use Gaussian conditioning to decompose the different contributions at each time step, and see if concentration of measure allows to simplify independent terms. Starting at $t = 0$:

$$\mathbf{v}^0 = \mathbf{w}^0 \tag{1.82}$$

$$\mathbf{r}^0 = \mathbf{A}\mathbf{v}^0 \sim \mathcal{N}(0, \frac{1}{d}\left\|\mathbf{w}^0\right\|_2^2 \mathbf{I}_n) \tag{1.83}$$

$$\mathbf{v}^1 = -\gamma^0 \mathbf{A}^\top \nabla \mathbf{f}(\mathbf{r}^0) \tag{1.84}$$

Since $\mathbf{v}^0$ is assumed to be independent of the rest, we can consider the whole proof as done conditioned on the distribution of $\mathbf{v}^0$. Focusing on $\mathbf{v}^1$, conditioning on $\mathbf{r}^0$ and using the Gaussian conditioning lemma 2

$$\begin{aligned}
\mathbf{v}^1|_{\mathbf{r}^0} &= \mathbf{v}^1|_{\mathbf{A}\mathbf{v}^0} && (1.85)\\
&= -\gamma^0 \left(\mathbf{A}|_{\mathbf{A}\mathbf{v}^0}\right)^\top \nabla \mathbf{f}(\mathbf{r}^0) && (1.86)\\
&= -\gamma^0 \left(\mathbf{A}\mathbf{P}_{\mathbf{v}^0} + \tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{v}^0}\right)^\top \nabla \mathbf{f}(\mathbf{r}^0) && (1.87)\\
&= -\gamma^0 \frac{\mathbf{v}^0(\mathbf{v}^0)^\top}{\|\mathbf{v}^0\|_2^2}\mathbf{A}^\top \nabla \mathbf{f}(\mathbf{r}^0) - \gamma^0 \mathbf{P}^\perp_{\mathbf{v}^0}\tilde{\mathbf{A}}^\top \nabla \mathbf{f}(\mathbf{r}^0) && (1.88)\\
&= -\gamma^0 \frac{\mathbf{v}^0(\mathbf{r}^0)^\top}{\|\mathbf{v}^0\|_2^2}\nabla \mathbf{f}(\mathbf{r}^0) - \gamma^0 \mathbf{P}^\perp_{\mathbf{v}^0}\tilde{\mathbf{A}}^\top \nabla \mathbf{f}(\mathbf{r}^0) && (1.89)
\end{aligned}$$

using similar arguments as before (see e.g. the chapter 2), we will reach a similar statement as for the first step of the gradient descent with sample splitting. Moving to $\mathbf{r}^1$, we condition on (the sigma algebra generated by) $\mathbf{v}^1, \mathbf{r}^0$ to reach

$$\begin{aligned}
\mathbf{r}^1|_{\mathbf{r}^0,\mathbf{v}^1} &= \mathbf{r}^1|_{\mathbf{A}\mathbf{v}^0,\mathbf{A}^\top\mathbf{m}^0} && (1.90)\\
&= \mathbf{A}\left(\mathbf{v}^0 + \mathbf{v}^1\right)|_{\mathbf{A}\mathbf{v}^0,\mathbf{A}^\top\mathbf{m}^0} && (1.91)\\
&= \mathbf{A}\mathbf{v}^0 + \mathbf{A}|_{\mathbf{A}\mathbf{v}^0,\mathbf{A}^\top\mathbf{m}^0}\mathbf{v}^1 && (1.92)\\
&= \mathbf{A}\mathbf{v}^0 + \left(\mathbf{P}_{\mathbf{m}^0}\mathbf{A} + \mathbf{A}\mathbf{P}_{\mathbf{v}^0} - \mathbf{P}_{\mathbf{m}^0}\mathbf{A}\mathbf{P}_{\mathbf{v}^0} + \mathbf{P}^\perp_{\mathbf{m}^0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{v}^0}\right)\mathbf{v}^1 && (1.93)\\
&= \mathbf{A}\mathbf{v}^0 + \left(\mathbf{P}_{\mathbf{m}^0}\mathbf{A}\mathbf{P}^\perp_{\mathbf{v}^0} + \mathbf{A}\mathbf{P}_{\mathbf{v}^0} + \mathbf{P}^\perp_{\mathbf{m}^0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{v}^0}\right)\mathbf{v}^1 && (1.94)\\
&= \underbrace{\mathbf{A}\mathbf{v}^0 + \mathbf{A}\mathbf{P}_{\mathbf{v}^0}\mathbf{v}^1 + \mathbf{P}^\perp_{\mathbf{m}^0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{v}^0}\mathbf{v}^1}_{\mathbf{I}_1} + \underbrace{\mathbf{P}_{\mathbf{m}^0}\mathbf{A}\mathbf{P}^\perp_{\mathbf{v}^0}\mathbf{v}^1}_{\mathbf{I}_2} && (1.95)
\end{aligned}$$

We will show that the term $\mathbf{I}_1$ constitutes an additive Gaussian process with correlation across all time steps, while the term $\mathbf{I}_2$ will build up a memory kernel. Starting with $\mathbf{I}_1$:

$$\mathbf{I}_1 \xrightarrow[n,d\to\infty]{Pl2} \mathbf{A}\mathbf{v}^0 + \mathbf{A}\mathbf{v}^0\frac{(\mathbf{v}^0)^\top\mathbf{v}^1}{\|\mathbf{v}^0\|_2^2} + \tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{v}^0}\mathbf{v}^1 \tag{1.96}$$

where $\tilde{\mathbf{A}}$ is independent on $\mathbf{A}, \mathbf{v}^0, \mathbf{r}^0, \mathbf{v}^1$, and we remind that the notation $\xrightarrow[n,d\to\infty]{Pl2}$ informally denotes that two random variables asymptotically give the same value for any pseudo-Lipschitz function of order 2. It is straightforward to check that this term converges to a Gaussian process with cross correlations equal to the inner product of successive iterates $\mathbf{w}$ (recall that the $\mathbf{v}$ are the increments in $\mathbf{w}$). The term $\mathbf{I}_2$ can be rewritten

$$\begin{aligned}
\mathbf{I}_2 &= \mathbf{m}^0\left((\mathbf{m}^0)^\top\mathbf{m}^0\right)^{-1}(\mathbf{m}^0)^\top\mathbf{A}\mathbf{P}^\perp_{\mathbf{v}^0}\mathbf{v}^1 && (1.97)\\
&= -\mathbf{m}^0\left((\mathbf{m}^0)^\top\mathbf{m}^0\right)^{-1}\left(\frac{1}{\gamma^0}\mathbf{v}^1\right)^\top\mathbf{P}^\perp_{\mathbf{v}^0}\mathbf{v}^1 && (1.98)\\
& && (1.99)
\end{aligned}$$

where, using the result for $\mathbf{v}^1$ at Eq.(1.89), we have that

$$(\mathbf{v}^1)^\top\mathbf{P}^\perp_{\mathbf{v}^0}\mathbf{v}^1 \xrightarrow[n,d\to\infty]{P} (\gamma^0)^2\frac{1}{d}\mathbb{E}\left[\left\|\nabla \mathbf{f}(\mathbf{r}^0)\right\|_2^2\right] \tag{1.100}$$

which leads to, using the definition of $\mathbf{m}^0$ and Eq.(1.83),

$$\mathbf{I}_2 \xrightarrow[n,d\to\infty]{Plk} \mathbf{m}^0\gamma^0 = -\gamma_0\nabla\mathbf{f}(\mathbf{r}^0) \tag{1.101}$$

which is the first term of a memory kernel. We will show in chapter 6 how to continue this proof using an induction. The curious reader may have a look at Theorem 9 from chapter 6, and see that the full result is somewhat impractical. In the introduction, we mentioned approximate message passing algorithms : we will now show how a stochastic correction at each time step may define an iteration with much simpler dynamics, while retaining all the relevant information for a wide family of problems.

1.7.5 Approximate message-passing

As discussed in subsection 1.4, AMP iterations first originate in statistical physics as Gaussian relaxation of belief-propagation on dense graphs, see e.g. [154], and their derivation is usually presented by formulating an inference problem as a factor graph and simplifying the messages in the high-dimensional limit using heuristic arguments. The set of non-linear equations describing the dynamics of messages in this limit is called *state evolution* equations, and have been the subject of mathematical proofs, notably by Erwin Bolthausen [41], and subsequently in [28, 135, 37]. Our proof in chapters 2 and 3 is based on similar ideas. The main benefit of these equations is that they track the exact asymptotic distribution of the iterates of the algorithm with a simple Markovian recursion at each time step, and this without any sample splitting assumption. From the mathematical point of view, if one is willing to forget about the physical intuition, AMP iterations can thus be seen as a family of sequence with an "appropriate" correction that considerably simplifies the dynamics without losing relevant information. We stress that this is not the standard way of presenting AMP iterations, but it is more in tune with the results presented in this thesis.

Recall the equation (1.89) we had for $\mathbf{v}^1$ on the first step of the natural gradient descent

$$\mathbf{v}^1|_{\mathbf{r}^0} = -\gamma^0\frac{\mathbf{v}^0(\mathbf{r}^0)^\top}{\|\mathbf{v}^0\|_2^2}\nabla\mathbf{f}(\mathbf{r}^0) - \gamma^0\mathbf{P}^\perp_{\mathbf{v}^0}\tilde{\mathbf{A}}^\top\nabla\mathbf{f}(\mathbf{r}^0) \tag{1.102}$$

where we have seen that the term $-\gamma^0\frac{\mathbf{v}^0(\mathbf{r}^0)^\top}{\|\mathbf{v}^0\|_2^2}\nabla\mathbf{f}(\mathbf{r}^0)$ converges in the pseudo-Lipschitz sense to a previous iterate $\mathbf{v}^0$ with asymptotically deterministic prefactor $-\gamma^0\mathbb{E}\left[f''(z^0)\right]$ where the distribution of z^0 may be evaluated from the previous iteration. The second term is simply an additive independent Gaussian. The idea is thus to remove, at each time step, a term of the form $b^t\mathbf{v}^t$ that cancels the first part, the so-called *Onsager* correction. This way, we only keep the information that is "new". Note that the conditional expectation elegantly captures this intuition : at each iteration, we only keep the part that is not measurable according to the σ-algebra generated by previous iterates. We now move to the proof of state evolution equations for the simplest instance of an AMP iteration, in the form it originally took to generate solutions of the Sherrington-Kirkpatrick problem at high-temperature in [41, 42].

Let $\mathbf{G} \in \mathrm{GOE}(n)$, $\{f^t\}_{t\in\mathbb{N}}$ a sequence of separable, pseudo-Lipschitz functions of order 2. The iterates $\mathbf{x}^t$ then take the form

$$\mathbf{x}^{t+1} = \mathbf{A}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{1.103}$$

$$\mathbf{m}^t = \mathbf{f}_t(\mathbf{x}^t) \tag{1.104}$$

with initialization at $\mathbf{x}^0 \in \mathbb{R}^n$, for instance with i.i.d. subGaussian coordinates.

$$b_t = \frac{1}{n}\text{div}\left(\mathbf{f}^t(\mathbf{x}^t)\right) \tag{1.105}$$

Definition 3 (state evolution iterates)**.** *The state evolution iterates are composed of one infinite-dimensional array $(\boldsymbol{\kappa}^{s,r})_{r,s>0}$ of scalars. This array is generated as follows. Define the first state evolution iterate*

$$\kappa^{1,1} = \mathbb{E}\left[(f^0(x^0))^2\right] \tag{1.106}$$

Recursively, once $\kappa^{s,r}, 0 \leqslant s,r \leqslant t$ are defined for some $t \geqslant 1$, take $z^0 = x^0$ and $(z^1,\dots,z^t) \in \mathbb{R}^t$ a centered Gaussian vector of covariance $(\boldsymbol{\kappa}^{s,r})_{s,r\leqslant t}$. We then define new state evolution iterates

$$\kappa^{t+1,s+1} = \kappa^{s+1,t+1} = \mathbb{E}\left[f^s(z^s)f^t(z^t)\right], \qquad s \in \{0,\dots,t\}.$$

The following property then holds for the AMP iteration (1.103)-(1.104).

Theorem 3. *Define, as above, $z^0 = x^0$ and $(z^1,\dots,z^t) \in \mathbb{R}^t$ a centered Gaussian vector of covariance $(\boldsymbol{\kappa}^{s,r})_{s,r\leqslant t}$. Then for any pseudo-Lipschitz function $\Phi : \mathbb{R}^{t+1} \to \mathbb{R}$ of order 2,*

$$\frac{1}{n}\sum_{i=1}^n \Phi\left((\mathbf{x}^0,\mathbf{x}^1,\dots,\mathbf{x}^t)_i\right) \overset{\text{P}}{\simeq} \mathbb{E}\left[\Phi\left(x^0,z^1,\dots,z^t\right)\right].$$

Define the σ-algebra $\mathfrak{S}_t = \sigma(\mathbf{x}^1,\mathbf{x}^2,...,\mathbf{x}^t)$. We then have :

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \mathbf{A}|_{\mathfrak{S}_t}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{1.107}$$

because $\mathbf{m}^t, \mathbf{m}^{t-1}, b^t$ are $\mathfrak{S}_t$-measurable.
A straightforward induction shows that conditioning on $\mathfrak{S}_t$ is equivalent to conditioning on the gaussian space generated by $\mathbf{Am}^0, \mathbf{Am}^1, ..., \mathbf{Am}^{t-1}$. We may then apply lemma 2 for a symmetric matrix (GOE(n)), to obtain :

$$\begin{aligned}
\mathbf{A}|_{\mathfrak{S}_t} &= \mathbb{E}\left[\mathbf{A}|\mathfrak{S}_t\right] + \mathcal{P}_t(\mathbf{A}) && (1.108)\\
&= \mathbf{A} - \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}} + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}} && (1.109)
\end{aligned}$$

where $\mathbf{M}_{t-1} = [\mathbf{m}^0|...|\mathbf{m}^{t-1}]$ and $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$. Using this on symmetric AMP iteration, we get :

$$\begin{aligned}
\mathbf{x}^{t+1}|_{\mathfrak{S}_t} &= \left(\mathbf{A} - \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}} + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\right)\mathbf{m}^t - b_t\mathbf{m}^{t-1} && (1.110)\\
&= \left(\mathbf{A} - \left(Id - \mathbf{P}_{\mathbf{M}_{t-1}}\right)\mathbf{A}\left(Id - \mathbf{P}_{\mathbf{M}_{t-1}}\right)\right)\mathbf{m}^t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t - b_t\mathbf{m}^{t-1} && (1.111)\\
&= \left(\mathbf{A}\mathbf{P}_{\mathbf{M}_{t-1}} + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{P}^T_{\mathbf{M}_{t-1}}\right)\mathbf{m}^t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t - b_t\mathbf{m}^{t-1} && (1.112)\\
&= \mathbf{A}\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1}. && (1.113)
\end{aligned}$$

The proof of the state evolutione equations is then done by induction, so we assume (after proving the initialization), that Theorem 3 is true up to time t. Assuming $\mathbf{M}_{t-1}$ has full rank (we will handle rigorously the existence of projectors in the proofs of chapter 3), we may define $\boldsymbol{\alpha}_t$ as the coefficients of the projection of $\mathbf{m}^t$ onto the columns of $\mathbf{M}_{t-1}$, $\boldsymbol{\alpha}^t = \left(\mathbf{M}^{\top}_{t-1}\mathbf{M}_{t-1}\right)^{-1}\mathbf{M}^{\top}_{t-1}\mathbf{m}^t$, which gives:

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \mathbf{A}\mathbf{M}_{t-1}\boldsymbol{\alpha}_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1}. \tag{1.114}$$

Using the definition of the symmetric AMP iteration, we have $\mathbf{A}\mathbf{M}_{t-1} = \mathbf{X}_{t-1} + [0|\mathbf{M}_{t-2}]\,\mathbf{B}_t$ where $\mathbf{X}_{t-1} = [\mathbf{x}^1|...|\mathbf{x}^t]$ and $\mathbf{B}_t$ is a diagonal matrix containing the Onsager terms up to time t. Then:

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = (\mathbf{X}_{t-1} + [0|\mathbf{M}_{t-2}]\,\mathbf{B}_t)\,\boldsymbol{\alpha}_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1} \tag{1.115}$$

$$= \underbrace{\mathbf{X}_{t-1}\boldsymbol{\alpha}_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t}_{\mathbf{I}_1} + \underbrace{[0|\mathbf{M}_{t-2}]\,\mathbf{B}_t\boldsymbol{\alpha}_t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1}}_{\mathbf{I}_2} \tag{1.116}$$

The term $\mathbf{I}_1$ in the above expression is a combination of previous terms with an additional new Gaussian one, coming from the independent copy $\tilde{\mathbf{A}}$. Checking the covariance of this term matches the state evolution equation for $t+1$. The term $\mathbf{I}_2$ cancels out in the high-dimensional limit, which is the main benefit of the Onsager correction. Let us sketch out how to cancel $\mathbf{I}_2$. We shall focus on the term

$$\mathcal{A} = \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} \tag{1.117}$$

$$= \mathbf{M}_{t-1}(\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}\mathbf{M}^T_{t-1}\mathbf{A}\mathbf{m}^t_{\perp} \tag{1.118}$$

$$= \mathbf{M}_{t-1}(\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}(\mathbf{A}\mathbf{M}_{t-1})^T\mathbf{m}^t_{\perp} \tag{1.119}$$

Then $(\mathbf{A}\mathbf{M}_{t-1})^T = (\mathbf{X}_{t-1} - [0|\mathbf{M}_{t-2}]\,\mathbf{B}_t)^T$ so that

$$(\mathbf{A}\mathbf{M}_{t-1})^T\mathbf{m}^t_{\perp} = (\mathbf{X}^T_{t-1} - \mathbf{B}^T_t\,[0|\mathbf{M}_{t-2}]^T)\mathbf{m}^t_{\perp} \xrightarrow[n\to\infty]{Plk} \mathbf{X}^T_{t-1}\mathbf{m}^t_{\perp} \tag{1.120}$$

Note that here, we have used an orthogonal decomposition of random vectors as if they were deterministic. Here however, using the induction hypothesis we can precisely write down what projections converge to, and deterministic limits are obtained for projection coefficients due to concentration of measure. In the case of AMP iterations, inner products of iterates essentially converge to their covariances due to the state evolution equations. For the proof of the DMFT equations however, we will see in chapter 6 that one must pay extra attention to the deterministic limits of projection coefficients. Back to the AMP sketch of proof, we obtain

$$\mathcal{A} = \mathbf{M}_{t-1}(\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}\mathbf{X}^T_{t-1}\mathbf{m}^t_{\perp} \tag{1.121}$$

$$= \mathbf{M}_{t-1}(\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}\mathbf{X}^T_{t-1}(\mathbf{m}^t - \mathbf{m}^t_{\parallel}) \tag{1.122}$$

$$= \mathbf{M}_{t-1}(\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}\mathbf{X}^T_{t-1}(f^t(\mathbf{x}^t) - \mathbf{M}^{t-1}\alpha^t) \tag{1.123}$$

$$= \mathbf{M}_{t-1}(\frac{1}{n}\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}\frac{1}{n}\mathbf{X}^T_{t-1}(f^t(\mathbf{x}^t) - \mathbf{M}^{t-1}\alpha^t) \tag{1.124}$$

where we have made the $\frac{1}{n}$ appear to highlight the two averaged inner-products of pseudo-Lipschitze functions. We will now use the induction hypothesis to simplify these terms,

$$\frac{1}{n}\mathbf{X}^T_{t-1}f^t(\mathbf{x}^t) = \begin{bmatrix} \frac{1}{n}\sum_i x_i^{(1)} f(x_i^{(t)}) \\ \frac{1}{n}\sum_i x_i^{(2)} f(x_i^{(t)}) \\ \dots \\ \frac{1}{n}\sum_i x_i^{(t)} f(x_i^{(t)}) \end{bmatrix} \tag{1.125}$$

$$\overset{\mathrm{P}}{\simeq} \begin{bmatrix} \mathbb{E}[z^1 f(z^{(t)})] \\ \mathbb{E}[z^2 f(z^{(t)})] \\ \dots \\ \mathbb{E}[z^t f(z^{(t)})] \end{bmatrix} \tag{1.126}$$

which, using Stein's lemma yields

$$\frac{1}{n}\mathbf{X}_{t-1}^T f^t(\mathbf{x}^t) \overset{\text{P}}{\simeq} \begin{bmatrix} \kappa_{1,t}\mathbb{E}[f'(z^{(t)})] \\ \kappa_{2,t}\mathbb{E}[f'(z^{(t)})] \\ \dots \\ \kappa_{t,t}\mathbb{E}[f'(z^{(t)})] \end{bmatrix} = b_t \begin{bmatrix} \kappa_{1,t} \\ \kappa_{2,t} \\ \dots \\ \kappa_{t,t} \end{bmatrix} \tag{1.127}$$

Then, the state evolution equations also give that

$$\frac{1}{n}\mathbf{m}^{s-1}\mathbf{m}^{t-1} \overset{\text{P}}{\simeq} \kappa_{s,t} \tag{1.128}$$

and therefore

$$\frac{1}{n}\mathbf{X}_{t-1}^T f^t(\mathbf{x}^t) \overset{\text{P}}{\simeq} \frac{1}{n}b_t \begin{bmatrix} (\mathbf{m}^0)^T\mathbf{m}^{t-1} \\ (\mathbf{m}^1)^T\mathbf{m}^{t-1} \\ \dots \\ (\mathbf{m}^{t-1})^T\mathbf{m}^{t-1} \end{bmatrix} = \frac{1}{N}b_t\mathbf{M}_{t-1}^T\mathbf{m}^{t-1} \tag{1.129}$$

We can deal in a similar way with the term $\frac{1}{n}\mathbf{X}_{t-1}^\top\mathbf{M}^{t-1}\boldsymbol{\alpha}^t$, such that

$$\frac{1}{n}\mathbf{X}_{t-1}^\top\mathbf{M}^{t-1}\boldsymbol{\alpha}^t \overset{\text{P}}{\simeq} \frac{1}{n}\mathbf{M}_{t-1}^T\left[0|\mathbf{M}_{t-2}\right]\mathbf{B}_t\boldsymbol{\alpha}_t \tag{1.130}$$

and finally

$$\mathcal{A} = \left(\mathbf{M}_{t-1}(\mathbf{M}_{t-1}^T\mathbf{M}_{t-1})^{-1}N\right)\left(\frac{1}{N}\mathbf{X}_{t-1}^T(f^t(\mathbf{x}^t) - \mathbf{M}^{t-1}\alpha^t)\right) \tag{1.131}$$

$$\xrightarrow[n\to\infty]{Plk} \mathbf{M}_{t-1}(\mathbf{M}_{t-1}^T\mathbf{M}_{t-1})^{-1}\mathbf{M}_{t-1}^T\underbrace{\left(b_t\mathbf{m}^{t-1} - \left[0|\mathbf{M}_{t-2}\right]\mathbf{B}_t\boldsymbol{\alpha}_t\right)}_{\in \text{span}(\mathbf{M}_{t-1})} \tag{1.132}$$

$$\xrightarrow[n\to\infty]{Plk} b_t\mathbf{m}^{t-1} - \left[0|\mathbf{M}_{t-2}\right]\mathbf{B}_t\boldsymbol{\alpha}_t \tag{1.133}$$

where the last line is obtained by explicitly writing the projection coefficients again. This is precisely the part needed to cancel $\mathbf{I}_2$, concluding the sketch of proof. The benefit of this proof method is that it directly gives the asymptotic equivalent of the distribution of each iterate, and the iterative projection argument can be extended to matrix-valued variables [135] and non-separaable non-linearities [37]. There are many more AMP iterations in the litterature, which we will discuss in the next chapter, where we will present an extension of a similar proof to problems that may involve several random matrices, in particular multilayer models [194, 188], and beyond.

Converging trajectories Now that we have presented the main dynamical tools to analyze the high-dimensional asymptotics, how do we use them to obtain resuls on a given estimator ? The idea is to design an AMP iteration whose fixed point matches the optimality condition of the optimization problem defining the estimator of interest, which, for strictly convex feasible problems, is enough to characterize the unique solution of the problem. This proof idea was pioneered in [29, 82] for the LASSO and unregularized logistic regression with i.i.d. Gaussian data, and in parts II and III we will build upon this method to study problems with generic convex loss and regularization, structured data and ensembles of estimators.

Adding a planted model In all the derivations presented above, we have neglected the presence of a teacher model that depends on a product $\mathbf{Aw}^*$ for a given ground-truth $\mathbf{w}^*$. As we will see in the next chapters, any low-rank perturbation such as a spike in the matrix $\mathbf{A}$ or a dependency of a non-linearity on the teacher output $\mathbf{Aw}^*$ can be accounted for by introducing additional order parameters and further conditioning arguments. We make this quantitative for a wide range of cases We can also deal with such dependencies with further orthogonal decompositions and Lagrange multipliers on the cost function of interest, which we will do in part II when studying convex problems, where strong duality allows to freeze cumbersome order parameters and optimize on the remaining variables.

Part I

High-dimensional dynamics : graph-based AMP iterations and first order methods

Chapter 2

Graph-based AMP iterations

This chapter presents the results of a joint work with R. Berthier, published in [110]. In the introduction, we sketched the main steps of proof for SE equations of AMP iterations based on iterative conditioning, which becomes quite tedious when all the steps are made rigorous. The papers [28, 135, 37] use this proof method to rigorously obtain the SE equations for the symmetric AMP iteration Eq.(1.103-1.103), and the asymmetric AMP iteration originally obtained in [83, 240, 154] for the probabilistic formulation of generalized linear models, respectively for separable functions and vector-valued iterates; block-separable functions and matrix valued iterates; and non-separable functions with matrix-valued iterates. However, many new AMP iterations along with their SE equations were heuristically derived for problems going well-beyond generalized linear models, notably in [194, 188, 13] where composite iterations involving a finite number of different random matrices are proposed to evaluate marginals from Hopfield models, multilayer neural networks with random weights and low-rank matrix estimation with deep generative priors. Here we propose to index AMP iterations on an oriented graph which may be composed arbitrarily, provided a certain structure is respected. We then prove SE equations for any AMP iteration indexed on such a graph, using an embedding argument based on a symmetric iteration with matrix-valued iterates and non-separable update functions, for which we prove the SE equations using the iterative conditioning scheme of Erwin Bolthausen. Extensions of the main theorem, such as spatial coupling or low-rank perturbations of the Gaussian random matrices, are finally proposed along with examples of applications.

AMP algorithms are iterative equations solving inference problems involving high-dimensional random variables with random interactions [83, 300]. For the typical case in which AMP iterations were initially studied, the interactions involve an i.i.d. Gaussian matrix. These algorithms are inspired from Bolthausen's iterative solution of the celebrated Thouless-Anderson-Palmer (TAP) equations of spin glass theory [196, 42, 43]. However, they are usually derived as heuristic relaxations of the belief propagation equations [227] on dense factor graphs in a manner often encountered in the context of statistical physics of disordered systems. A central property of AMP iterations is that the distribution of their outputs can be tracked rigorously in the high-dimensional limit by low-dimensional equations called *state evolution* (SE). This property can be seen as similar to the concept of density evolution from coding theory [243], but in the case of dense factor graphs.

In recent years, the growing interest in high-dimensional inference and learning problems has motivated the introduction of approximate-message passing algorithms as solutions to many inference

problems, and as analytical tools—thanks to the SE equations—to study the statistical properties of learned estimators, notably starting with the LASSO [28, 153, 83]. A number of extensions were then proposed for inference problems of growing complexity: generalized linear modelling and robust m-estimators [240, 82, 300], low-rank matrix reconstruction [241, 165], principal component analysis (PCA) [75, 164], inference in deep multilayer networks with random weights [188], matrix-valued inference problems [15] or matrix recovery under generative priors [14], among others. Interestingly, AMP algorithms can be composed with one another to solve inference problems obtained by combining factor graphs, as demonstrated in [14], where each part of the factor graph represents an elaborate prior and inference process. This demonstrates the adaptability of such iterations, even more so as the state evolution equations are shown to hold, often heuristically, for these composite structures.

Contributions. As the diversity of inference problems and AMP iterations increases, it is important to identify a common structure underlying the known AMP algorithms. Such a partial unification was done in [135, 37]: symmetric and asymmetric AMP iterations are treated in a common framework. However, these results do not apply to the more recent AMP iterations designed for more complex problems presenting multilayered structures or ones obtained by combining factor graphs.

Our first contribution is to show how AMP algorithms are naturally indexed by a graph that determines its form. Seeing AMP algorithms as supported by this graph helps understanding the iterations, especially the multi-layer ones, in a unified way. In this regard, we hope that our framework will be used as a tool to generate new AMP iterations. Roughly speaking, the graph underlying the AMP iteration represents the interaction of the high-dimensional variables of the associated inference problem. However, this graph is not the factor graph representing the inference problem that sometimes appears in the derivation of AMP equations, see [153] for example. The factor graph is microscopic, in the sense that it disappears when taking the dense limit leading to the AMP equations. On the contrary, the graph that we consider here is macroscopic: it structures the AMP iteration itself. It is insensitive to the underlying inference problem that has generated the AMP equation; for instance, it can be used in both Bayes optimal or non-Bayes optimal scenarios.

The second contribution of this chapter is to use the graph framework to show that all graph-based AMP iterations admit a rigorous SE description. This generalizes the previous works of [28, 135, 37] on SE to more complex iterations. Using our result, writing and proving the state evolution equations is reduced to the identification of a specific structure in the AMP iteration, instead of heuristically deriving or reproducing the rigorous proof entirely for problems of increasing complexity. In particular, it gives a theoretical grounding for the analysis of AMP on recent multi-layer structures [188, 15, 14]. Related to [188], this chapter proves that AMP algorithms are a rigorously grounded approach to understanding multi-layer neural networks, albeit only when the weights are random and when we perform inference with an AMP algorithm. Still, in a context where theory struggles to explain the behavior of multi-layered neural networks, it is interesting to see that this particular case can be rigorously studied, even for deep architectures.

We illustrate the flexibility of our framework by applying it to diverse inference problems mentioned above, notably multilayer generalized linear estimation problems and low-rank matrix recovery with deep generative priors. We also show how our results can be extended to handle matrix-valued variables, combined with the spatial coupling framework introduced in [153, 135], and how low-rank perturbations such as spikes in the random matrices or additional dependencies of the non-linearities on linear observations change the state evolution equations.

Related work. There is a rich literature of proofs of state evolution equations, notably starting with Bolthausen's iterative scheme [42, 43] based on Gaussian conditioning. The technique was then adapted and extended to the case of a more generic AMP iteration related to the LASSO problem in [28], where it is mentioned that Gaussian conditioning methods also appear in [79] to tackle fundamental random convex geometry problems. The analysis was then extended to matrix-valued variables with block-separable non-linearities in [135] and for vector-valued variables with non-separable non-linearities in [37], which also show that symmetric AMP and asymmetric AMP can be treated in the same framework. Our proof is partly based on the same iterative Gaussian conditioning method but is additionally combined with an embedding specific to the graph framework. To the best of our knowledge, the latter part of the proof is novel.

Another line of work—called VAMP (vector approximate message passing) algorithms—handles rotationnally invariant matrices [242] with generic spectrum. This family of VAMP iterations is obtained using a Gaussian parametrization of *expectation propagation* [202, 219], a variational inference algorithm based on iterative moment-matching between a chosen form of probability distribution (e.g., Gaussian nodes on a factor graph) and a target distribution observed through empirical data. These iterations also verify SE equations proven with a similar conditioning method [278, 242], handling a different kind of randomness than i.i.d. Gaussian matrices. The SE proof for VAMP iterations was then extended to multilayer inference problems and their matrix-valued counterparts in [97, 222]. In these works, the conditioning method is applied in a sequential manner to each layer of the problem, making it specific to multilayer inference problems. On the contrary, our proof method is not restricted to sequential multilayer estimation as mentioned in the contributions, and does not rely on iterating through the graph. However, our proof does not apply to all rotationnally invariant matrices. We handle mostly Gaussian or GOE matrices, with extensions to correlated Gaussian matrices, products of Gaussian matrices and spatially coupled Gaussian matrices. This is discussed in greater detail in Sections 2.2 and 2.3.

Outline of the chapter. The chapter is organised as follows: we start by presenting the indexation of AMP iterations by an oriented graph in Section 2.1. Several conceptual examples are provided. We present the state evolution equations on any graph-supported AMP iteration in Section 2.2, along with its proof, which constitutes the main technical contribution of this chapter. We then move to applications to inference problems in Section 2.3 and conclude on related open problems in Section 2.4. All proofs of auxiliary results are deferred to the Appendix.

Notations. We adopt similar notations to those of [37]. Differences are mainly due to the matrix variables framework.

We denote scalars with lowercase letters, vectors with bold lowercase letters and matrices with bold uppercase ones. Inner products are denoted by brackets $\langle ., . \rangle$, and the canonical inner products are chosen for vectors and matrices, i.e., $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$, $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}\left(\mathbf{X}^\top \mathbf{Y}\right)$. The associated norms are respectively denoted $\|.\|_2$ and $\|.\|_F$ for the Frobenius norm.

For two random variables X and Y, and a σ-algebra $\mathfrak{S}$, we use $X|_{\mathfrak{S}} \overset{d}{=} Y$ to mean that for any integrable function ϕ and any $\mathfrak{S}$-measurable bounded random variable Z, $\mathbb{E}\left[\phi(X)Z\right] = \mathbb{E}\left[\phi(Y)Z\right]$. For two sequences of random variables X_n, Y_n, we write $X_n \overset{\mathrm{P}}{\simeq} Y_n$ when their difference converges in probability to 0, i.e., $X_n - Y_n \xrightarrow{P} 0$.

We use $\mathbf{I}_N$ to denote the $N \times N$ identity matrix, and $0_{N\times N}$ the $N \times N$ matrix with zero entries. We use $\sigma_{min}(\mathbf{Q})$ and $\sigma_{max}(\mathbf{Q}) = \|\mathbf{Q}\|_{op}$ to denote the minimum and maximum singular values of

a given matrix $\mathbf{Q}$. For two matrices $\mathbf{Q}$ and $\mathbf{P}$ with the same number of rows, we denote their horizontal concatenation with $[\mathbf{P}|\mathbf{Q}]$. The orthogonal projector onto the range of a given matrix $\mathbf{M}$ is denoted $\mathbf{P}_{\mathbf{M}}$, and let $\mathbf{P}_{\mathbf{M}}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{M}}$.

Let $\mathcal{S}_q^+$ denote the space of positive semi-definite matrices of size $q \times q$. For any matrix $\boldsymbol{\kappa} \in \mathcal{S}_q^+$ and a random matrix $\mathbf{Z} \in \mathbb{R}^{N \times q}$ we write $\mathbf{Z} \sim \mathbf{N}(0, \boldsymbol{\kappa} \otimes \mathbf{I}_N)$ if $\mathbf{Z}$ is a matrix with jointly Gaussian entries such that for any $1 \leqslant i, j \leqslant q$, $\mathbb{E}[\mathbf{Z}^i(\mathbf{Z}^j)^\top] = \boldsymbol{\kappa}_{i,j}\mathbf{I}_N$, where $\mathbf{Z}^i, \mathbf{Z}^j$ denote the i-th and j-th columns of $\mathbf{Z}$. The i-th line of the matrix $\mathbf{Z}$ is denoted $\mathbf{Z}_i$.

If $f : \mathbb{R}^{N \times q} \to \mathbb{R}^{N \times q}$ is an function and $i \in \{1, \dots N\}$, we write $f_i : \mathbb{R}^{N \times q} \to \mathbb{R}^q$ the component of f generating the i-th line of its image, i.e., if $\mathbf{X} \in \mathbb{R}^{N \times q}$,

$$f(\mathbf{X}) = \begin{bmatrix} f_1(\mathbf{X}) \\ \vdots \\ f_N(\mathbf{X}) \end{bmatrix} \in \mathbb{R}^{N \times q}.$$

We write $\frac{\partial f_i}{\partial \mathbf{X}_i}$ the $q \times q$ Jacobian containing the derivatives of f_i with respect to (w.r.t.) the i-th line $\mathbf{X}_i \in \mathbb{R}^q$:

$$\frac{\partial f_i}{\partial \mathbf{X}_i} = \begin{bmatrix} \frac{\partial (f_i(\mathbf{X}))_1}{\partial \mathbf{X}_{i1}} & \cdots & \frac{\partial (f_i(\mathbf{X}))_1}{\partial \mathbf{X}_{iq}} \\ \vdots & & \vdots \\ \frac{\partial (f_i(\mathbf{X}))_q}{\partial \mathbf{X}_{i1}} & \cdots & \frac{\partial (f_i(\mathbf{X}))_q}{\partial \mathbf{X}_{iq}} \end{bmatrix} \in \mathbb{R}^{q \times q}. \tag{2.1}$$

2.1 Graph-based AMP iterations

We start by defining the class of graphs indexing AMP iterations.

Definition 4 (graph notions)**.** *A* finite directed graph*—also simply called* graph *in the following—is a pair $G = (V, \overrightarrow{E})$ where V is a finite set, called the* vertex set*, and $\overrightarrow{E}$ is a subset of $V \times V$, called the* edge set*. This definition of graphs uses directed edges and allows loops.*

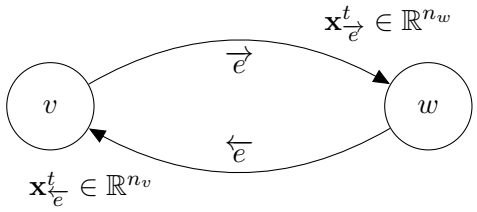
A graph $G = (V, \overrightarrow{E})$ is said to be symmetric *if for all $v, w \in \overrightarrow{E}$, $(v, w) \in \overrightarrow{E}$ if and only if $(w, v) \in \overrightarrow{E}$.*

The degree $\deg v$ *of a node $v \in V$ is the number of edges of which it is the end-node. In symmetric graphs, it is also the number of edges of which v is the starting-node.*

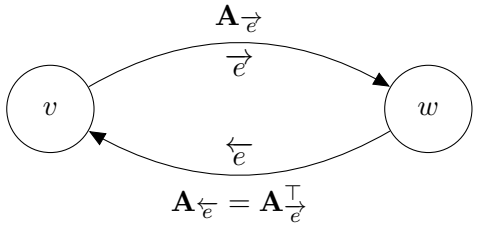
Graph notations. Given a symmetric graph $G = (V, \overrightarrow{E})$, the following notations are useful. We sometimes write $v \to w$ to mean that $\overrightarrow{e} = (v, w)$ is an edge of the graph. We say that v is the starting-node of $\overrightarrow{e}$ and w the end-node of $\overrightarrow{e}$. We denote $\overleftarrow{e} = (w, v) \in \overrightarrow{E}$ the symmetric edge of $\overrightarrow{e}$. If $\overrightarrow{e}$ is a loop, then $\overleftarrow{e} = \overrightarrow{e}$. We write $\overrightarrow{e} \to \overrightarrow{e}'$ as a shorthand to say that the end-node of $\overrightarrow{e} \in \overrightarrow{E}$ is the starting-node of $\overrightarrow{e}' \in \overrightarrow{E}$. Note that for any $\overrightarrow{e} \in \overrightarrow{E}$, $\overleftarrow{e} \to \overrightarrow{e}$.

Iteration. We now fix a symmetric finite directed graph $G = (V, \overrightarrow{E})$. We associate an AMP iteration supported by the graph G as follows.

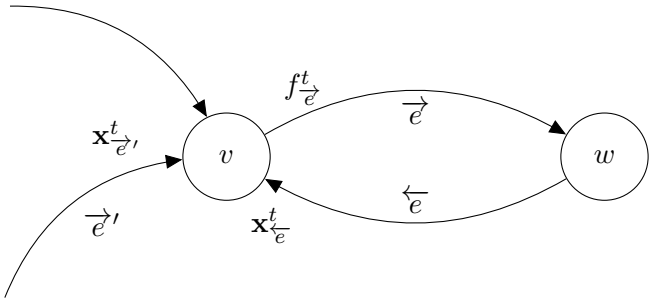
- The variables $\mathbf{x}_{\overrightarrow{e}}^t$ of the AMP iteration are indexed by the iteration number $t \in \mathbb{N}$ and the oriented edges of the graph $\overrightarrow{e} \in \overrightarrow{E}$.



- All variables associated to edges $\overrightarrow{e} = (v, w)$ with end-node $w \in V$ have a same dimension $n_w \in \mathbb{N}_{>0}$, i.e., $\mathbf{x}^t_{\overrightarrow{e}} \in \mathbb{R}^{n_w}$. We define $N = \sum_{(v,w)\in\overrightarrow{E}} n_w$ the sum of the dimensions of all variables.
- Matrices of the AMP iteration are also indexed by the edges of the graph. If $\overrightarrow{e} = (v, w) \in \overrightarrow{E}$, $\mathbf{A}_{\overrightarrow{e}} \in \mathbb{R}^{n_w \times n_v}$. These matrices must satisfy the symmetry condition $\mathbf{A}_{(v,w)} = \mathbf{A}^\top_{(w,v)}$. In particular, this implies that matrices $\mathbf{A}_{(v,v)} \in \mathbb{R}^{n_v \times n_v}$ associated to loops $(v, v) \in \overrightarrow{E}$ must be symmetric.



- Non-linearities of the AMP iteration are also indexed by the edges of the graph (and possibly by the iteration number t). If $t \geqslant 0$ and $\overrightarrow{e} = (v, w) \in \overrightarrow{E}$, $f^t_{(v,w)}\left(\left(\mathbf{x}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)$ is a function of all the variables of the edges whose end-node is the starting-node v of $\overrightarrow{e}$, as denoted by the condition $\overrightarrow{e}' \to \overrightarrow{e}$. It is a function from $(\mathbb{R}^{n_v})^{\deg v}$ to $\mathbb{R}^{n_v}$.



Once these parameters $(\mathbf{A}_{\overrightarrow{e}})_{\overrightarrow{e}\in\overrightarrow{E}}$ and $\left(f^t_{\overrightarrow{e}}\right)_{t\geqslant 0,\overrightarrow{e}\in\overrightarrow{E}}$ are given, we can choose an arbitrary initial condition $\mathbf{x}^0_{\overrightarrow{e}} \in \mathbb{R}^{n_w}$ for all oriented edges $\overrightarrow{e} \in \overrightarrow{E}$ of the graph. We define recursively the AMP iterates $\left(\mathbf{x}^t_{\overrightarrow{e}}\right)_{t\geqslant 0,\overrightarrow{e}\in\overrightarrow{E}}$, by the iteration: for all $t \geqslant 0$, $\overrightarrow{e} \in \overrightarrow{E}$,

$$\mathbf{x}^{t+1}_{\overrightarrow{e}} = \mathbf{A}_{\overrightarrow{e}}\mathbf{m}^t_{\overrightarrow{e}} - b^t_{\overrightarrow{e}}\mathbf{m}^{t-1}_{\overleftarrow{e}}\,, \tag{2.2}$$

$$\mathbf{m}^t_{\overrightarrow{e}} = f^t_{\overrightarrow{e}}\left(\left(\mathbf{x}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)\,, \tag{2.3}$$

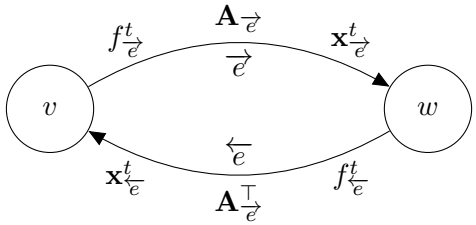
where $b^t_{\overrightarrow{e}}$ is the so-called *Onsager term*

$$b^t_{\overrightarrow{e}} = \frac{1}{N} \operatorname{Tr} \frac{\partial f^t_{\overrightarrow{e}}}{\partial \mathbf{x}_{\overleftarrow{e}}} \left(\left(\mathbf{x}^t_{\overrightarrow{e}'} \right)_{\overrightarrow{e}' : \overrightarrow{e}' \to \overrightarrow{e}} \right) \qquad \in \mathbb{R} \,. \tag{2.4}$$

The above partial derivative makes sense as $\overleftarrow{e} \to \overrightarrow{e}$, thus $\mathbf{x}_{\overleftarrow{e}}$ is a variable of $f^t_{\overrightarrow{e}}$. Note that in (2.2), the Onsager term multiplies the vector $\mathbf{m}^{t-1}_{\overleftarrow{e}}$ indexed by the symmetric edge $\overleftarrow{e}$ of $\overrightarrow{e}$.

Let us derive some simple particular cases of this framework, first to recover the classical asymmetric and symmetric AMP iterations, and second to cover multi-layer AMP iterations.

Asymmetric AMP. The asymmetric AMP iteration appeared first in the literature to solve the compressed sensing problem [83] and then more generally to tackle generalized linear estimation, see, e.g., [240, 255, 82]. It corresponds to a simple underlying graph composed of two nodes and two symmetric directed edges between them.

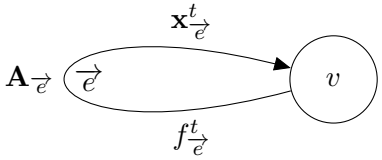


In this case, the graph AMP equations (2.2)-(2.3) give

$$\begin{aligned} \mathbf{x}^{t+1}_{\overrightarrow{e}} &= \mathbf{A}_{\overrightarrow{e}} \mathbf{m}^t_{\overrightarrow{e}} - b^t_{\overrightarrow{e}} \mathbf{m}^{t-1}_{\overleftarrow{e}} \,, \\ \mathbf{m}^t_{\overrightarrow{e}} &= f^t_{\overrightarrow{e}} \left(\mathbf{x}^t_{\overleftarrow{e}} \right) \,, \\ \mathbf{x}^{t+1}_{\overleftarrow{e}} &= \mathbf{A}^\top_{\overrightarrow{e}} \mathbf{m}^t_{\overleftarrow{e}} - b^t_{\overleftarrow{e}} \mathbf{m}^{t-1}_{\overrightarrow{e}} \,, \\ \mathbf{m}^t_{\overleftarrow{e}} &= f^t_{\overleftarrow{e}} \left(\mathbf{x}^t_{\overrightarrow{e}} \right) \,. \end{aligned} \tag{2.5}$$

The corresponding state evolution (SE) property was proved in [28] for the separable case and in [37] in the non-separable case. Note that the time indices proposed here are different from the ones appearing in these works. The time index convention adopted here generalizes better to more elaborate graphs. We show how to recover the usual time indices in Appendix 3.1.

Symmetric AMP. The symmetric AMP iteration is central to our discussion as we show that all graph AMP iterations can be reduced to this case (with matrix-valued iterates, as detailed below). It is already known that the asymmetric case can be reduced to this case [135]. The symmetric AMP iteration appears, e.g., when solving the low-rank matrix recovery problem [241, 75], or community detection in graphs [74]. It corresponds to the degenerate graph with only one node and one loop.

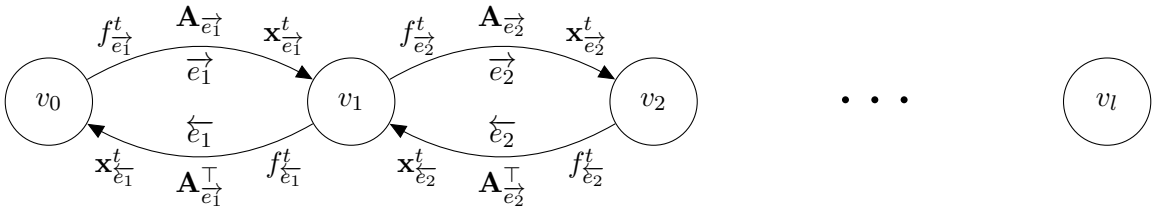


Recall that $\overleftarrow{e} = \overrightarrow{e}$ as $\overrightarrow{e}$ is a loop. In this case, the graph AMP equations (2.2)-(2.3) give

$$\begin{aligned} \mathbf{x}_{\overrightarrow{e}}^{t+1} &= \mathbf{A}_{\overrightarrow{e}} \mathbf{m}_{\overrightarrow{e}}^{t} - b_{\overrightarrow{e}}^{t} \mathbf{m}_{\overrightarrow{e}}^{t-1} , \\ \mathbf{m}_{\overrightarrow{e}}^{t} &= f_{\overrightarrow{e}}^{t} \left(\mathbf{x}_{\overrightarrow{e}}^{t} \right) , \end{aligned} \tag{2.6}$$

Here, as there is a single edge $\overrightarrow{e}$, the indexes are superfluous and could be dropped. For these equations, the SE property was proved in [135] for the separable case and in [37] in the non-separable case. Note that the results of [135] allow matrix-valued variables.

Multi-layer AMP. The multi-layer AMP iteration appears when considering inference problems through a multi-layer random neural network, see [188]. They correspond to a line graph whose length l is the number of layers.



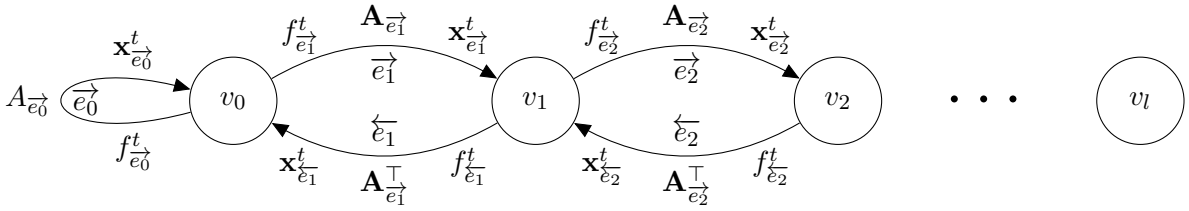
In this case, the graph AMP equations (2.2)-(2.3) give

$$\begin{aligned} \mathbf{x}_{\overrightarrow{e_1}}^{t+1} &= \mathbf{A}_{\overrightarrow{e_1}} \mathbf{m}_{\overrightarrow{e_1}}^{t} - b_{\overrightarrow{e_1}}^{t} \mathbf{m}_{\overleftarrow{e_1}}^{t-1} , \\ \mathbf{m}_{\overrightarrow{e_1}}^{t} &= f_{\overrightarrow{e}_1}^{t} \left(\mathbf{x}_{\overleftarrow{e_1}}^{t} \right) , \\ \mathbf{x}_{\overleftarrow{e_1}}^{t+1} &= \mathbf{A}_{\overrightarrow{e_1}}^{\top} \mathbf{m}_{\overleftarrow{e_1}}^{t} - b_{\overleftarrow{e_1}}^{t} \mathbf{m}_{\overrightarrow{e_1}}^{t-1} , \\ \mathbf{m}_{\overleftarrow{e_1}}^{t} &= f_{\overleftarrow{e_1}}^{t} \left(\mathbf{x}_{\overrightarrow{e_1}}^{t}, \mathbf{x}_{\overleftarrow{e_2}}^{t} \right) , \\ \\ \mathbf{x}_{\overrightarrow{e_2}}^{t+1} &= \mathbf{A}_{\overrightarrow{e_2}} \mathbf{m}_{\overrightarrow{e_2}}^{t} - b_{\overrightarrow{e_2}}^{t} \mathbf{m}_{\overleftarrow{e_2}}^{t-1} , \\ \mathbf{m}_{\overrightarrow{e_2}}^{t} &= f_{\overrightarrow{e}_2}^{t} \left(\mathbf{x}_{\overleftarrow{e_1}}^{t}, \mathbf{x}_{\overleftarrow{e_2}}^{t} \right) , \\ \mathbf{x}_{\overleftarrow{e_2}}^{t+1} &= \mathbf{A}_{\overrightarrow{e_2}}^{\top} \mathbf{m}_{\overleftarrow{e_2}}^{t} - b_{\overleftarrow{e_2}}^{t} \mathbf{m}_{\overrightarrow{e_2}}^{t-1} , \\ \mathbf{m}_{\overleftarrow{e_2}}^{t} &= f_{\overleftarrow{e_2}}^{t} \left(\mathbf{x}_{\overrightarrow{e_2}}^{t}, \mathbf{x}_{\overleftarrow{e_3}}^{t} \right) , \end{aligned} \tag{2.7}$$

$$\vdots$$

Note that the non-linearities now take several variables as inputs when there are several incoming edges at a node.

Spiked matrix model under generative multi-layer priors. Of course, the structures described above can be combined to tackle new AMP iterations. For instance, the paper [15] studies the recovery of noisy symmetric rank-1 matrix when the spike comes from a known multi-layer generative prior. The associated AMP iteration corresponds to the following graph, where the loop corresponds to the spike recovery and the other edges correspond to multi-layer prior on the spike.



In this case, the graph AMP equations (2.2)-(2.3) give

$$\begin{aligned}
\mathbf{x}^{t+1}_{\overrightarrow{e_0}} &= \mathbf{A}_{\overrightarrow{e_0}}\mathbf{m}^t_{\overrightarrow{e_0}} - b^t_{\overrightarrow{e_0}}\mathbf{m}^{t-1}_{\overrightarrow{e_0}}\,,\\
\mathbf{m}^t_{\overrightarrow{e_0}} &= f^t_{\overrightarrow{e_0}}\left(\mathbf{x}^t_{\overrightarrow{e_0}}, \mathbf{x}^t_{\overleftarrow{e_1}}\right)\,,
\end{aligned}$$

$$\begin{aligned}
\mathbf{x}^{t+1}_{\overrightarrow{e_1}} &= \mathbf{A}_{\overrightarrow{e_1}}\mathbf{m}^t_{\overrightarrow{e_1}} - b^t_{\overrightarrow{e_1}}\mathbf{m}^{t-1}_{\overleftarrow{e_1}}\,,\\
\mathbf{m}^t_{\overrightarrow{e_1}} &= f^t_{\overrightarrow{e}_1}\left(\mathbf{x}^t_{\overrightarrow{e_0}}, \mathbf{x}^t_{\overleftarrow{e_1}}\right)\,,\\
\mathbf{x}^{t+1}_{\overleftarrow{e_1}} &= \mathbf{A}^\top_{\overrightarrow{e_1}}\mathbf{m}^t_{\overleftarrow{e_1}} - b^t_{\overleftarrow{e_1}}\mathbf{m}^{t-1}_{\overrightarrow{e_1}}\,,\\
\mathbf{m}^t_{\overleftarrow{e_1}} &= f^t_{\overleftarrow{e_1}}\left(\mathbf{x}^t_{\overrightarrow{e_1}}, \mathbf{x}^t_{\overleftarrow{e_2}}\right)\,,
\end{aligned} \tag{2.8}$$

$$\vdots$$

2.2 State evolution for graph-based AMP iterations

In this section, we start by presenting the most straightforward form of our result, and show afterwards how several refinements can be added.

2.2.1 Main theorem

AMP algorithms admit a state evolution description under two major assumptions: that the interactions matrices $\mathbf{A}_{\overrightarrow{e}}$ are sufficiently random—in our case Gaussian or GOE—and that the dimensions $n = (n_v)_{v\in V}$ of all the variables converge to infinity with fixed ratios.

Assumptions. We make the following assumptions:

(A1) The matrices $(\mathbf{A}_{\overrightarrow{e}})_{\overrightarrow{e}\in\overrightarrow{E}}$ are random and independent, up to the symmetry condition $\mathbf{A}_{\overleftarrow{e}} = \mathbf{A}^\top_{\overrightarrow{e}}$. Moreover, if $(v,w) \in \overrightarrow{E}$ is not a loop in G, i.e., $v \neq w$, then $\mathbf{A}_{(v,w)}$ has independent centered Gaussian entries with variance $1/N$. If $(v,v) \in \overrightarrow{E}$ is a loop in G, then $\mathbf{A}_{(v,v)}$ has independent entries (up to the symmetry $\mathbf{A}_{(v,v)} = \mathbf{A}^\top_{(v,v)}$), centered Gaussian with variance $2/N$ on the diagonal and variance $1/N$ off the diagonal.

(A2) For all $v \in V$, $n_v \to \infty$ and n_v/N converges to a well-defined limit $\delta_v \in [0,1]$. We denote by $n \to \infty$ the limit under this scaling.

(A3) For all $t \in \mathbb{N}$ and $\overrightarrow{e} \in \overrightarrow{E}$, the non-linearity $f^t_{\overrightarrow{e}}$ is pseudo-Lipschitz of finite order, uniformly with respect to the problem dimensions $n = (n_v)_{v\in V}$ (see Definition 1 in Appendix 3.5).

(A4) For all $\overrightarrow{e} \in E$, $\left\|\mathbf{x}^0_{\overrightarrow{e}}\right\|_2/\sqrt{N}$ converges to a finite constant as $n \to \infty$.

(A5) For all $\overrightarrow{e} \in E$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{N} \left\langle f^0_{\overrightarrow{e}} \left(\left(\mathbf{x}^0_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right), f^0_{\overrightarrow{e}} \left(\left(\mathbf{x}^0_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\rangle$$

(A6) Let $(\kappa_{\overrightarrow{e}})_{\overrightarrow{e}\in E}$ be an array of bounded non-negative reals and $\mathbf{Z}_{\overrightarrow{e}} \sim \mathbf{N}(0, \kappa_{\overrightarrow{e}} \mathbf{I}_{n_w})$ independent random variables for all $\overrightarrow{e}$. For all $\overrightarrow{e} \in E$, for any $t \in \mathbb{N}_{>0}$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{N} \mathbb{E}\left[\left\langle f^0_{\overrightarrow{e}} \left(\left(\mathbf{x}^0_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right), f^t_{\overrightarrow{e}} \left(\left(\mathbf{Z}^t_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\rangle \right].$$

(A7) Consider any array of 2×2 positive definite matrices $(\boldsymbol{S}_{\overrightarrow{e}})_{\overrightarrow{e}\in E}$ and the collection of random variables $(\mathbf{Z}_{\overrightarrow{e}}, \mathbf{Z}'_{\overrightarrow{e}}) \sim \mathbf{N}(0, \boldsymbol{S}_{\overrightarrow{e}} \otimes \mathbf{I}_{n_w}))$ defined independently for each edge $\overrightarrow{e}$. Then for any $\overrightarrow{e} \in E$ and $s, t > 0$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{N} \mathbb{E}\left[\left\langle f^s_{\overrightarrow{e}} \left((\mathbf{Z}^s_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right), f^t_{\overrightarrow{e}} \left(\left(\tilde{\mathbf{Z}}^t_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\rangle \right].$$

Remark on the assumptions. In the literature, the random matrices $\mathbf{A}_{(v,w)}$ of AMP iterations are often scaled with variances $1/n_w$. To recover the desired scaling, it is sufficient to rescale the non-linearity on which a given matrix acts with the corresponding aspect ratio δ_w.

Definition 5 (State evolution iterates)**.** *The state evolution iterates are composed of one infinite-dimensional array $(\boldsymbol{\kappa}^{s,r}_{\overrightarrow{e}})_{r,s>0}$ of real values for each edge $\overrightarrow{e} \in \overrightarrow{E}$. These arrays are generated as follows. Define the first state evolution iterates*

$$\boldsymbol{\kappa}^{1,1}_{\overrightarrow{e}} = \lim_{n\to\infty} \frac{1}{N} \left\| f^0_{\overrightarrow{e}} \left(\left(\mathbf{x}^0_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\|_2^2, \qquad \overrightarrow{e} \in \overrightarrow{E}.$$

Recursively, once $(\boldsymbol{\kappa}^{s,r}_{\overrightarrow{e}})_{s,r\leqslant t, \overrightarrow{e}\in\overrightarrow{E}}$ are defined for some $t \geqslant 1$, define independently for each $\overrightarrow{e} \in \overrightarrow{E}$, $\mathbf{Z}^0_{\overrightarrow{e}} = \mathbf{x}^0_{\overrightarrow{e}}$ and $(\mathbf{Z}^1_{\overrightarrow{e}}, \dots, \mathbf{Z}^t_{\overrightarrow{e}})$ a centered Gaussian random vector of covariance $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}})_{r,s\leqslant t} \otimes I_{n_w}$. We then define new state evolution iterates

$$\boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}} = \boldsymbol{\kappa}^{s+1,t+1}_{\overrightarrow{e}} = \lim_{n\to\infty} \frac{1}{N} \mathbb{E}\left[\left\langle f^s_{\overrightarrow{e}} \left((\mathbf{Z}^s_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right), f^t_{\overrightarrow{e}} \left(\left(\mathbf{Z}^t_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\rangle \right]$$
for all $s \in \{1, \dots, t\}, \overrightarrow{e} \in \overrightarrow{E}$.

Theorem 4. *Assume (A1)-(A7). Define, as above, independently for each $\overrightarrow{e} = (v, w) \in \overrightarrow{E}$, $\mathbf{Z}^0_{\overrightarrow{e}} = \mathbf{x}^0_{\overrightarrow{e}}$ and $(\mathbf{Z}^1_{\overrightarrow{e}}, \dots, \mathbf{Z}^t_{\overrightarrow{e}})$ a centered Gaussian random vector of covariance $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}})_{r,s\leqslant t} \otimes \mathbf{I}_{n_w}$. Then for any sequence of uniformly (in n) pseudo-Lipschitz function $\Phi : \mathbb{R}^{(t+1)N} \to \mathbb{R}$,*

$$\Phi\left((\mathbf{x}^s_{\overrightarrow{e}})_{0\leqslant s\leqslant t, \overrightarrow{e}\in\overrightarrow{E}} \right) \overset{\mathrm{P}}{\sim} \mathbb{E}\left[\Phi\left((\mathbf{Z}^s_{\overrightarrow{e}})_{0\leqslant s\leqslant t, \overrightarrow{e}\in\overrightarrow{E}} \right) \right]$$

2.2.2 Reduction of graph-based AMP iterations to the matrix-valued, non-separable symmetric case

The core strategy in the proof of Theorem 4 is to reduce the graph AMP iteration (2.2)-(2.4) into a symmetric AMP iteration with matrix-valued iteration, i.e., an iteration of the form

$$\mathbf{X}^{t+1} = \mathbf{A}\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top \qquad \in \mathbb{R}^{N\times q}, \tag{2.9}$$

$$\mathbf{M}^t = f^t(\mathbf{X}^t) \qquad \in \mathbb{R}^{N\times q}, \tag{2.10}$$

$$\mathbf{b}_t = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial f_i^t}{\partial \mathbf{X}_i}(\mathbf{X}^t) \qquad \in \mathbb{R}^{q\times q}. \tag{2.11}$$

Here, $\mathbf{A}$ is a $N\times N$ GOE matrix, the iterates $\mathbf{X}^t, \mathbf{M}^t$ are $N\times q$ matrices, and $f^t : \mathbb{R}^{N\times q} \to \mathbb{R}^{N\times q}$ are non-separable non-linearities. A rigorous SE description for this iteration is established in Appendix 3.2; it is an extension of the results of [135, 37].

In this section, we show that the graph AMP iteration (2.2)-(2.4) can be formulated as a symmetric AMP iteration (2.9)-(2.11) with matrix iterates. In Appendix 3.2.2, this reduction is used to show that Theorem 4 follows from its equivalent on symmetric iterations.

Let $q = |\overrightarrow{E}|$, $\overrightarrow{e}_1, \dots, \overrightarrow{e}_l$ be the loops of G and $\overrightarrow{e}_{l+1}, \overleftarrow{e}_{l+1}, \dots, \overrightarrow{e}_m, \overleftarrow{e}_m$ be the other edges of the graph. Define

$$\mathbf{X}^0 = \begin{pmatrix} \mathbf{x}^0_{\overrightarrow{e}_1} & & & & & & & \\ & \ddots & & & & & * & \\ & & \mathbf{x}^0_{\overrightarrow{e}_l} & & & & & \\ & & & \mathbf{x}^0_{\overrightarrow{e}_{l+1}} & & & & \\ & & & & \mathbf{x}^0_{\overleftarrow{e}_{l+1}} & & & \\ & & & & & \ddots & & \\ & * & & & & & \mathbf{x}^0_{\overrightarrow{e}_m} & \\ & & & & & & & \mathbf{x}^0_{\overleftarrow{e}_m} \end{pmatrix} \in \mathbb{R}^{N\times q}.$$

where $*$ denotes entries whose values do not matter for what follows. Let $\mathbf{A}$ be a $N\times N$ GOE matrix such that

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{\overrightarrow{e}_1} & & & & & & & \\ & \ddots & & & & & * & \\ & & \mathbf{A}_{\overrightarrow{e}_l} & & & & & \\ & & & * & \mathbf{A}_{\overrightarrow{e}_{l+1}} & & & \\ & & & \mathbf{A}_{\overleftarrow{e}_{l+1}} & * & & & \\ & & & & & \ddots & & \\ & * & & & & & * & \mathbf{A}_{\overrightarrow{e}_m} \\ & & & & & & \mathbf{A}_{\overleftarrow{e}_m} & * \end{pmatrix}.$$

Finally, define the non-linearities $f_t : \mathbb{R}^{N\times q} \to \mathbb{R}^{N\times q}$ as

$$f^t\begin{pmatrix} \mathbf{x}_{\overrightarrow{e}_1} & & & & & & & & \\ & \ddots & & & & & * & & \\ & & \mathbf{x}_{\overrightarrow{e}_l} & & & & & & \\ & & & \mathbf{x}_{\overrightarrow{e}_{l+1}} & & & & & \\ & & & & \mathbf{x}_{\overleftarrow{e}_{l+1}} & & & & \\ & & & & & \ddots & & & \\ & * & & & & & \mathbf{x}_{\overrightarrow{e}_m} & & \\ & & & & & & & \mathbf{x}_{\overleftarrow{e}_m} \end{pmatrix} \tag{2.12}$$

$$= \begin{pmatrix} f^t_{\overrightarrow{e}_1}\left((\mathbf{x}_{\overrightarrow{e}})_{\overrightarrow{e}:\overrightarrow{e}\to\overrightarrow{e}_1}\right) & & & & & & & & \\ & \ddots & & & & & & 0 & \\ & & f^t_{\overrightarrow{e}_l}(\dots) & & & & & & \\ & & & 0 & f^t_{\overleftarrow{e}_{l+1}}(\dots) & & & & \\ & & & f^t_{\overrightarrow{e}_{l+1}}(\dots) & 0 & & & & \\ & & & & & \ddots & & & \\ & 0 & & & & & 0 & f^t_{\overleftarrow{e}_m}(\dots) & \\ & & & & & & f^t_{\overrightarrow{e}_m}(\dots) & 0 & \end{pmatrix}$$

Lemma 3. *Define $\mathbf{X}^0$, $\mathbf{A}$ and f^t as above. Then the iterates $\mathbf{X}^t$ of the symmetric AMP iteration (2.9)-(2.11) are of the form*

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}^t_{\overrightarrow{e}_1} & & & & & & & \\ & \ddots & & & & & * & \\ & & \mathbf{x}^t_{\overrightarrow{e}_l} & & & & & \\ & & & \mathbf{x}^t_{\overrightarrow{e}_{l+1}} & & & & \\ & & & & \mathbf{x}^t_{\overleftarrow{e}_{l+1}} & & & \\ & & & & & \ddots & & \\ & * & & & & & \mathbf{x}^t_{\overrightarrow{e}_m} & \\ & & & & & & & \mathbf{x}^t_{\overleftarrow{e}_m} \end{pmatrix} \in \mathbb{R}^{N\times q},$$

where $\mathbf{x}^t_{\overrightarrow{e}}$ denote the iterates of the graph-AMP iteration (2.2)-(2.4).

Proof. We proceed by induction. Assume that $\mathbf{X}^t$ and $\mathbf{X}^{t-1}$ are indeed of this form and we show the claim for $\mathbf{X}^{t+1}$. We use equations (2.9)-(2.11) to compute $\mathbf{X}^{t+1}$; we start by computing the

Onsager term $\mathbf{b}_t = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial f_i^t}{\partial \mathbf{X}_i}(\mathbf{X}^t) \in \mathbb{R}^{q\times q}$. From the formula for f^t, we compute

$$\mathbf{b}_t = \frac{1}{N}\begin{pmatrix} \mathrm{Tr}\,\frac{\partial f^t_{\overrightarrow{e}_1}}{\partial \mathbf{x}_{\overrightarrow{e}_1}}(\dots) & & & & & & \\ & \ddots & & & & 0 & \\ & & \mathrm{Tr}\,\frac{\partial f^t_{\overrightarrow{e}_l}}{\partial \mathbf{x}_{\overrightarrow{e}_l}}(\dots) & & & & \\ & & & 0 & \mathrm{Tr}\,\frac{\partial f^t_{\overrightarrow{e}_{l+1}}}{\partial \mathbf{x}_{\overleftarrow{e}_{l+1}}}(\dots) & & \\ & 0 & & \mathrm{Tr}\,\frac{\partial f^t_{\overleftarrow{e}_{l+1}}}{\partial \mathbf{x}_{\overrightarrow{e}_{l+1}}}(\dots) & 0 & & \\ & & & & & \ddots & \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{b}^t_{\overrightarrow{e}_1} & & & & & \\ & \ddots & & & 0 & \\ & & \mathbf{b}^t_{\overrightarrow{e}_l} & & & \\ & & & 0 & \mathbf{b}^t_{\overrightarrow{e}_{l+1}} & \\ & 0 & & \mathbf{b}^t_{\overleftarrow{e}_{l+1}} & 0 & \\ & & & & & \ddots \end{pmatrix}.$$

Then we can now compute

$$\mathbf{X}^{t+1} = \mathbf{A}\mathbf{M}^t - \mathbf{M}^{t-1}\mathbf{b}_t^\top .$$

First,

$$\mathbf{A}\mathbf{M} = \begin{pmatrix} \mathbf{A}_{\overrightarrow{e}_1} & & & & & \\ & \ddots & & & & \\ & & \mathbf{A}_{\overrightarrow{e}_l} & & & \\ & & & * & \mathbf{A}_{\overrightarrow{e}_{l+1}} & \\ & & & \mathbf{A}_{\overleftarrow{e}_{l+1}} & * & \\ & & & & & \ddots \end{pmatrix} \begin{pmatrix} f^t_{\overrightarrow{e}_1}(.) & & & & & \\ & \ddots & & & & \\ & & f^t_{\overrightarrow{e}_l}(.) & & & \\ & & & 0 & f^t_{\overleftarrow{e}_{l+1}}(.) & \\ & & & f^t_{\overrightarrow{e}_{l+1}}(.) & 0 & \\ & & & & & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{A}_{\overrightarrow{e}_1} f^t_{\overrightarrow{e}_1}\left(\left(\mathbf{x}^t_{\overrightarrow{e}}\right)_{\overrightarrow{e}:\overrightarrow{e}\to\overrightarrow{e_1}}\right) & & & & & \\ & \ddots & & & * & \\ & & \mathbf{A}_{\overrightarrow{e}_l} f^t_{\overrightarrow{e}_l}(.) & & & \\ & & & \mathbf{A}_{\overrightarrow{e}_{l+1}} f^t_{\overrightarrow{e}_{l+1}}(.) & & \\ & * & & & \mathbf{A}_{\overleftarrow{e}_{l+1}} f^t_{\overleftarrow{e}_{l+1}}(.) & \\ & & & & & \ddots \end{pmatrix}.$$

Second,

$$\mathbf{M}^{t-1}\mathbf{b}_t = \begin{pmatrix} f^{t-1}_{\overrightarrow{e}_1}(.) & & & & & & \\ & \ddots & & & & & \\ & & f^{t-1}_{\overrightarrow{e}_l}(.) & & & & \\ & & & 0 & f^{t-1}_{\overleftarrow{e}_{l+1}}(.) & & \\ & & & f^{t-1}_{\overrightarrow{e}_{l+1}}(.) & 0 & & \\ & & & & & \ddots & \end{pmatrix} \begin{pmatrix} \mathbf{b}^t_{\overrightarrow{e}_1} & & & & & & \\ & \ddots & & & & & \\ & & \mathbf{b}^t_{\overrightarrow{e}_l} & & & & \\ & & & 0 & \mathbf{b}^t_{\overleftarrow{e}_{l+1}} & & \\ & & & \mathbf{b}^t_{\overrightarrow{e}_{l+1}} & 0 & & \\ & & & & & \ddots & \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{b}^t_{\overrightarrow{e}_1} f^{t-1}_{\overrightarrow{e}_1}\left(\left(\mathbf{x}^t_{\overrightarrow{e}}\right)_{\overrightarrow{e}:\overrightarrow{e}\to\overrightarrow{e_1}}\right) & & & & & & \\ & \ddots & & & & & \\ & & \mathbf{b}^t_{\overrightarrow{e}_l} f^{t-1}_{\overrightarrow{e}_l}(.) & & & & \\ & & & \mathbf{b}^t_{\overrightarrow{e}_{l+1}} f^{t-1}_{\overleftarrow{e}_{l+1}}(.) & & & \\ & & & & \mathbf{b}^t_{\overleftarrow{e}_{l+1}} f^{t-1}_{\overrightarrow{e}_{l+1}}(.) & & \\ & & & & & \ddots & \end{pmatrix}.$$

Thus, combining the above equations, we obtain

$$\mathbf{X}^{t+1} = \mathbf{AM} - \mathbf{M}^{t-1}\mathbf{b}_t^\top$$

$$= \begin{pmatrix} \mathbf{A}_{\overrightarrow{e}_1} f^t_{\overrightarrow{e}_1}(.) - b^t_{\overrightarrow{e}_1} f^{t-1}_{\overrightarrow{e}_1}(.) & & & & & & \\ & \ddots & & & & & * \\ & & \mathbf{A}_{\overrightarrow{e}_l} f^t_{\overrightarrow{e}_l}(.) - b^t_{\overrightarrow{e}_l} f^{t-1}_{\overrightarrow{e}_l}(.) & & & & \\ & & & \mathbf{A}_{\overrightarrow{e}_{l+1}} f^t_{\overrightarrow{e}_{l+1}}(.) - b^t_{\overrightarrow{e}_{l+1}} f^{t-1}_{\overleftarrow{e}_{l+1}}(.) & & \\ & & & & & \ddots & \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{x}^{t+1}_{\overrightarrow{e}_1} & & & & \\ & \ddots & & * & \\ & & \mathbf{x}^{t+1}_{\overrightarrow{e}_l} & & \\ & * & & \mathbf{x}^{t+1}_{\overrightarrow{e}_{l+1}} & \\ & & & & \ddots \end{pmatrix}.$$

This proves the induction. □

2.2.3 Useful extensions

Here we present several refinements of Theorem 4 that can be obtained in a straightforward fashion and appear often in statistical inference problems.

Matrix-valued variables. The variables $\mathbf{x}_{\overrightarrow{e}}, \mathbf{m}_{\overrightarrow{e}}$ initially defined as vectors can be extended to matrices with a finite number of columns, and the non-linearities $f^t_{\overrightarrow{e}}$ are then matrix-valued functions of matrix-valued variables.

- $n_v \in \mathbb{N}_{>0}$ is now the number of lines of the variables coming in node $v \in V$. The definition $N = \sum_{(v,w)\in\overrightarrow{E}} n_w$ remains the same.
- Let $q_{\overrightarrow{e}} \in \mathbb{N}_{>0}$ be the number of columns of $\mathbf{x}^t_{\overrightarrow{e}}$. We assume that, for all $\overrightarrow{e} \in E$, $q_{\overrightarrow{e}} = q_{\overleftarrow{e}}$, and the $q_{\overrightarrow{e}}$ remain constant, independently of $n \to \infty$.

- The initial condition becomes $\mathbf{x}^0_{(v,w)} \in \mathbb{R}^{n_w \times q_{(v,w)}}$, for all edges $\overrightarrow{e} = (v,w)$.
- Non-linearities f_t indexed by the edge $\overrightarrow{e} = (v,w) \in \overrightarrow{E}$, $f^t_{(v,w)}((x^t_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}})$ are now functions from $\times_{\overrightarrow{e}'\to\overrightarrow{e}}\mathbb{R}^{n_v \times q_{\overrightarrow{e}'}}$ to $\mathbb{R}^{n_v \times q_{(v,w)}}$.

The AMP iterates are then recursively defined with:

$$\mathbf{x}^{t+1}_{\overrightarrow{e}} = \mathbf{A}_{\overrightarrow{e}}\mathbf{m}^t_{\overrightarrow{e}} - \mathbf{m}^{t-1}_{\overleftarrow{e}}(\mathbf{b}^t_{\overrightarrow{e}})^\top \quad \in \mathbb{R}^{n_w \times q_{\overrightarrow{e}}}\,, \tag{2.13}$$

$$\mathbf{m}^t_{\overrightarrow{e}} = f^t_{\overrightarrow{e}}\left(\left(\mathbf{x}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)\,, \tag{2.14}$$

where each Onsager term is now a matrix given by:

$$\mathbf{b}^t_{\overrightarrow{e}} = \frac{1}{N}\sum_{i=1}^{n_v}\frac{\partial f^t_{\overrightarrow{e},i}}{\partial \mathbf{x}_{\overleftarrow{e},i}}\left(\left(\mathbf{x}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right) \qquad \in \mathbb{R}^{q_{\overrightarrow{e}}\times q_{\overrightarrow{e}}}\,.$$

where we used the notation from Eq.(2.1). The state evolution equations then read

$$\boldsymbol{\kappa}^{1,1}_{\overrightarrow{e}} = \lim_{n\to\infty}\frac{1}{N} f^0_{\overrightarrow{e}}(\mathbf{x}^0_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}})^\top f^0_{\overrightarrow{e}}(\mathbf{x}^0_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}) \quad \in \mathbb{R}^{q_{\overrightarrow{e}}\times q_{\overrightarrow{e}}}\,, \quad \overrightarrow{e} \in \overrightarrow{E}\,.$$

$$\boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}} = \boldsymbol{\kappa}^{s+1,t+1}_{\overrightarrow{e}} = \lim_{n\to\infty}\frac{1}{N}\mathbb{E}\left[f^s_{\overrightarrow{e}}((\mathbf{Z}^s_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}})^\top f^t_{\overrightarrow{e}}((\mathbf{Z}^t_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}})\right] \quad \in \mathbb{R}^{q_{\overrightarrow{e}}\times q_{\overrightarrow{e}}}$$
$$\text{for all} \quad 1 \leqslant s \leqslant t\,, \overrightarrow{e} \in \overrightarrow{E}\,.$$

where the Gaussian fields generalize straightforwardly to $\mathbf{Z}^t_{\overrightarrow{e}} \sim \mathbf{N}(0, \boldsymbol{\kappa}^{t,t}_{\overrightarrow{e}} \otimes \mathbf{I}_{n_w}) \in \mathbb{R}^{n_w \times q_{\overrightarrow{e}}}$ for each edge. Using these generalized definitions, the above statement of Theorem 4 and its proof can be adapted easily. We give examples throughout Section 2.3.

Additional random variables in the non-linearities. Many inference problems are formulated with a "planted" signal, i.e., a ground truth signal parametrizing the function the statistician tries to reconstruct, sometimes called *teacher* in statistical physics. This often leads to the dependence of certain non-linearities on additional random variables. As long as they appropriately concentrate and are independent on the rest of the problem, they can be treated in straightforward fashion with an additional average in the SE equations as done in [135], where the summability is reduced to second-order moments conditions due to the separability of the update functions. However it is not always straightforward to isolate the independent contribution in the teacher which is often generated using the matrices found in the AMP algorithm, effectively introducing a correlation between the matrices and non-linearities. In appendix 3.4, we propose a generic way to deal with such dependencies with two additional results in the form of Lemmas 15 and Lemma 16. These two lemmas may be combined at will to deal with a wide range of perturbations relevant to inference problems. We now give an example of graph to which we apply those results, recovering the full SE equations of [188, 14]: consider any instance of the family of AMP iterations presented in Section 2.1, indexed on a given oriented graph $G = (V, E)$, i.e.

$$\mathbf{x}^{t+1}_{\overrightarrow{e}} = \hat{\mathbf{A}}_{\overrightarrow{e}}\mathbf{m}^t_{\overrightarrow{e}} - b^t_{\overrightarrow{e}}\mathbf{m}^{t-1}_{\overleftarrow{e}}\,, \tag{2.15}$$

$$\mathbf{m}^t_{\overrightarrow{e}} = \tilde{f}^t_{\overrightarrow{e}}\left(\left(\mathbf{x}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)\,, \tag{2.16}$$

where, in the notation of Lemma 3, for any symmetric edge $\overrightarrow{e}$ from the set $\{\overrightarrow{e}_1, ..., \overrightarrow{e}_l\}$, $\hat{\mathbf{A}}_{\overrightarrow{e}} = \mathbf{A}_{\overrightarrow{e}} + \frac{1}{N}\mathbf{v}_{\overrightarrow{e}}\mathbf{v}_{\overleftarrow{e}}^\top$, and $\tilde{f}^t_{\overrightarrow{e}}(.) = f^t_{\overrightarrow{e}}(.)$. Furthermore, for any asymmetric edge $\overrightarrow{e}$ from the set $\{\overrightarrow{e}_{l+1}, ..., \overrightarrow{e}_m\}$, $\hat{\mathbf{A}}_{\overrightarrow{e}} = \mathbf{A}_{\overrightarrow{e}}$ and $\tilde{f}^t_{\overrightarrow{e}}(.) = f^t(\varphi_{\overrightarrow{e}}(\mathbf{A}_{\overrightarrow{e}}\mathbf{w}_{\overrightarrow{e}}), .)$. The following lemma then gives the SE equations for this iteration:

Lemma 4. *Assume that (A1)-(A7) are verified. Further assume that, for any $\overrightarrow{e} \in \overrightarrow{E}$, $\frac{1}{\sqrt{N}}\|\mathbf{v}_{\overrightarrow{e}}\|_F$ and $\frac{1}{\sqrt{N}}\|\mathbf{w}_{\overrightarrow{e}}\|_F$ converge to finite constants as $N \to \infty$. For any symmetric edge $\overrightarrow{e}$ from the set $\{\overrightarrow{e}_1, ..., \overrightarrow{e}_l\}$, define the following SE recursion:*

$$\boldsymbol{\mu}^0_{\overrightarrow{e}},\ \boldsymbol{\kappa}^{1,1}_{\overrightarrow{e}} = \lim_{N\to\infty} \frac{1}{N} f^0_{\overrightarrow{e}}\left(\left(\mu^0_{\overrightarrow{e}'}\mathbf{v}_{\overrightarrow{e}'} + \mathbf{x}^0_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)^\top f^0_{\overrightarrow{e}}\left(\left(\mu^0_{\overrightarrow{e}'}\mathbf{v}_{\overrightarrow{e}'} + \mathbf{x}^0_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right) \tag{2.17}$$

$$\boldsymbol{\mu}^{s+1}_{\overrightarrow{e}} = \lim_{N\to+\infty} \frac{1}{N}\mathbb{E}\left[(\mathbf{v}_{\overrightarrow{e}})^\top f^s_{\overrightarrow{e}}((\mu^s_{\overrightarrow{e}'}\mathbf{v}_{\overrightarrow{e}'} + \mathbf{Z}^s_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}})\right] \tag{2.18}$$

$$\boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}} = \boldsymbol{\kappa}^{s+1,t+1}_{\overrightarrow{e}} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[f^s_{\overrightarrow{e}}((\mu^s_{\overrightarrow{e}'}\mathbf{v}_{\overrightarrow{e}'} + \mathbf{Z}^s_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}})^\top f^t_{\overrightarrow{e}}\left(\left(\mu^t_{\overrightarrow{e}'}\mathbf{v}_{\overrightarrow{e}'} + \mathbf{Z}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)\right], \quad s \in \{0, \dots, t\}. \tag{2.19}$$

where $(\mathbf{Z}^1_{\overrightarrow{e}}, \dots, \mathbf{Z}^t_{\overrightarrow{e}})$ is a centered Gaussian random vector of covariance $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}})_{r,s\leqslant t} \otimes \mathbf{I}_{n_w}$. Then, for any sequence of uniformly (in n) pseudo-Lipschitz function $\Phi : \mathbb{R}^{(t+1)n_w} \to \mathbb{R}$:

$$\Phi\left((\mathbf{x}^s_{\overrightarrow{e}})_{0\leqslant s\leqslant t, \overrightarrow{e}\in\overrightarrow{E}_{sym}}\right) \overset{\mathrm{P}}{\sim} \mathbb{E}\left[\Phi\left((\mu^s_{\overrightarrow{e}}\mathbf{v}_{\overrightarrow{e}} + \mathbf{Z}^s_{\overrightarrow{e}})_{0\leqslant s\leqslant t, \overrightarrow{e}\in\overrightarrow{E}_{sym}}\right)\right] \tag{2.20}$$

For any asymmetric edge $\overrightarrow{e}$ from the set $\{\overrightarrow{e}_{l+1}, ..., \overrightarrow{e}_m\}$, define the following SE recursion :

$$\boldsymbol{\nu}^0_{\overrightarrow{e}}, \hat{\boldsymbol{\nu}}^0_{\overrightarrow{e}}, \boldsymbol{\kappa}^{1,1}_{\overrightarrow{e}} = \frac{1}{N} f^0_{\overrightarrow{e}}((\mathbf{x}^0_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}})^\top f^0_{\overrightarrow{e}}((\mathbf{x}^0_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}) \tag{2.21}$$

$$\boldsymbol{\nu}^{t+1}_{\overrightarrow{e}} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\mathbf{w}^\top_{\overrightarrow{e}} f^t_{\overrightarrow{e}}\left(\varphi_{\overrightarrow{e}}(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}}}), \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}'}}, \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}'}}\boldsymbol{\nu}^t_{\overrightarrow{e}'} + \mathbf{w}_{\overleftarrow{e}'}\hat{\boldsymbol{\nu}}^t_{\overrightarrow{e}'} + \mathbf{Z}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)\right] \tag{2.22}$$

$$\hat{\boldsymbol{\nu}}^{t+1}_{\overrightarrow{e}} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f^t_{\overrightarrow{e},i}}{\partial \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}}, i}, \varphi_{\overrightarrow{e}}}\left(\varphi_{\overrightarrow{e}}(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}}}), \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}'}}, \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}'}}\boldsymbol{\nu}^t_{\overrightarrow{e}'} + \mathbf{w}_{\overleftarrow{e}'}\hat{\boldsymbol{\nu}}^t_{\overrightarrow{e}'} + \mathbf{Z}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right)\right] \tag{2.23}$$

$$\begin{aligned}
&\boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}} = \boldsymbol{\kappa}^{s+1,t+1}_{\overrightarrow{e}} = \\
&\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Bigg[\left(f^s_{\overrightarrow{e}}\left(\varphi_{\overrightarrow{e}}(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}}}), \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}'}}, \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}'}}\boldsymbol{\nu}^s_{\overrightarrow{e}'} + \mathbf{w}_{\overleftarrow{e}'}\hat{\boldsymbol{\nu}}^s_{\overrightarrow{e}'} + \mathbf{Z}^s_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right) - \mathbf{w}_{\overrightarrow{e}}\rho^{-1}_{\mathbf{w}_{\overrightarrow{e}}}\boldsymbol{\nu}^{s+1}_{\overrightarrow{e}}\right)^\top \\
&\left(f^t_{\overrightarrow{e}}\left(\varphi_{\overrightarrow{e}}(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}}}), \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}'}}, \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}'}}\boldsymbol{\nu}^t_{\overrightarrow{e}'} + \mathbf{w}_{\overleftarrow{e}'}\hat{\boldsymbol{\nu}}^t_{\overrightarrow{e}'} + \mathbf{Z}^t_{\overrightarrow{e}'}\right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}}\right) - \mathbf{w}_{\overrightarrow{e}}\rho^{-1}_{\mathbf{w}_{\overrightarrow{e}}}\boldsymbol{\nu}^{t+1}_{\overrightarrow{e}}\right)\Bigg]
\end{aligned} \tag{2.24}$$

where $(\mathbf{Z}^1_{\overrightarrow{e}}, \dots, \mathbf{Z}^t_{\overrightarrow{e}})$ is a centered Gaussian random vector of covariance $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}})_{r,s\leqslant t} \otimes \mathbf{I}_{n_w}$. Then, for any sequence of uniformly (in n) pseudo-Lipschitz function $\Phi : \mathbb{R}^{(t+1)n_w} \to \mathbb{R}$:

$$\Phi\left((\mathbf{x}^s_{\overrightarrow{e}})_{0\leqslant s\leqslant t, \overrightarrow{e}\in\overrightarrow{E}_{asym}}\right) \overset{\mathrm{P}}{\sim} \mathbb{E}\left[\Phi\left((\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}}}\rho^{-1}_{\mathbf{w}_{\overrightarrow{e}}}\boldsymbol{\nu}^s_{\overrightarrow{e}} + \mathbf{w}_{\overleftarrow{e}}\hat{\boldsymbol{\nu}}^s_{\overrightarrow{e}} + \mathbf{Z}^s_{\overrightarrow{e}})_{0\leqslant s\leqslant t, \overrightarrow{e}\in\overrightarrow{E}_{asym}}\right)\right] \tag{2.25}$$

Note the dependence on $\mathbf{w}_{\overleftarrow{e}}$ of the SE quantities indexed by $\overrightarrow{e}$, which comes from evaluating the matrix products defining the terms in $\mathbf{m}^t, \hat{\mathbf{m}}^t$. In the AMP litterature, non-linearities often take the form $\tilde{f}^t_{\overrightarrow{e}}(.) = f^t(\varphi_{\overrightarrow{e}}(\mathbf{A}_{\overleftarrow{e}}\mathbf{w}_{\overrightarrow{e}}), .)$, i.e. with a dependence on the random matrix of the opposite

edge. This only changes the arrows in $\mathbf{W}_0$ i.e.

$$\mathbf{W}_0 = \begin{pmatrix} 0 & & & & & & & \\ & \ddots & & & & & 0 & \\ & & 0 & & & & & \\ & & & 0 & \mathbf{w}_{\overrightarrow{e}_{l+1}} & & & \\ & & & \mathbf{w}_{\overleftarrow{e}_{l+1}} & 0 & & & \\ & & & & & \ddots & & \\ & 0 & & & & & 0 & \mathbf{w}_{\overrightarrow{e}_m} \\ & & & & & & \mathbf{w}_{\overleftarrow{e}_m} & 0 \end{pmatrix}, \tag{2.26}$$

and the corresponding arrows in the SE equations above. It is indeed what is observed in, e.g. [240, 188, 14]. Examples are given throughout Section 2.3.

Structured and correlated matrices. Products of Gaussian matrices can be considered by choosing identities as non-linearities on given edges of the graph. This was done heuristically in [188] to study structured inference problems. Gaussian matrices with generic covariances can also be considered, i.e., $\mathbf{A} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$ where $\mathbf{Z}$ is an i.i.d. $\mathbf{N}(0, \frac{1}{d})$ matrix and $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive definite matrix. Indeed, the covariance matrix can be absorbed in the non-linearity as a non-separable component. Depending on the non-linearity, expressions may simplify as functions of the spectral distribution of $\mathbf{\Sigma}$. Examples are given in Section 2.3.5.

Spatial coupling. *Spatial coupling* was introduced and studied in [153, 153, 135, 80] as a mean to reach information theoretic limits in compressed sensing. The idea is to write the state evolution equations when the random matrices have a block structure of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1l} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \dots & \mathbf{A}_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{k1} & \mathbf{A}_{k2} & \dots & \mathbf{A}_{kl} \end{bmatrix} \in \mathbb{R}^{N\times d},$$

each $\mathbf{A}_{ij} \in \mathbb{R}^{N_i\times d_j}$ has i.i.d. $\mathbf{N}(0, \frac{\sigma_{ij}}{d})$ entries and $N_i/N, d_j/d$ are constant aspect ratios, where $\sum_i N_i = N$ and $\sum_j d_j = d$. The proof of SE equations with this kind of matrices was proposed in [135] and relies on a matrix-valued symmetric AMP iteration similar to the one used in our proof, with a family of non-linearities acting on blocks of variables, with a separable effect on each block. Since our proof extends the matrix-valued, symmetric AMP iteration to the fully non-separable case, the same ideas can be applied to our framework to include spatially coupled matrices on each edge of the oriented graph presented in the previous section (with the added possibility of non-separable effects on each block). We now give an example in Section 2.3.6.

2.3 Applications to inference problems

In this section we illustrate our main theorem by showing how several AMP iterations established heuristically in the literature are included in our framework, in particular [188, 15, 14, 178], and how straightforward generalizations can be considered. We adopt an optimization viewpoint for each problem, omitting the probabilistic inference formulation at the origin of these iterations for simplicity.

2.3.1 A building block: AMP for generalized linear models

We start with a known AMP iteration for which the state evolution equations were already proven, and build upon the intuition it gives to present more elaborate iterations. Consider the task of optimizing a penalized cost functions of the form

$$\hat{\mathbf{x}} \in \min_{\mathbf{x}\in\mathbb{R}^d} g(\mathbf{A}\mathbf{x}, \mathbf{y}) + f(\mathbf{x}) \tag{2.27}$$

where the vector of labels $\mathbf{y}$ is typically assumed to be generated from another process as

$$\mathbf{y} = \phi(\mathbf{A}\mathbf{x}_0),$$

with $\mathbf{x}_0 \in \mathbb{R}^d$ generated from a given distribution $p_{\mathbf{x}_0}$ independent from the matrix $\mathbf{A}$, $\mathbf{A} \in \mathbb{R}^{N\times d}$ is a matrix with i.i.d. $\mathbf{N}(0, \frac{1}{d})$ elements, and ϕ a given function. The goal is then to reconstruct the vector $\mathbf{x}_0$. This formulation is at the basis of many of the fundamental estimation methods in machine learning: least-squares, LASSO, logistic regression, etc. Approximate-message passing algorithms were proposed for this task, notably in [83, 28, 240, 153, 135], and take the generic form of the asymmetric AMP iteration (2.5) where $\mathbf{A}_{\overrightarrow{e}} = \mathbf{A}$. Intuitively, the functions $f^t_{\overrightarrow{e}}, f^t_{\overleftarrow{e}}$ each correspond to one of the functions g, f from (2.27) and respectively output an estimate of the quantities $\mathbf{A}\hat{\mathbf{x}}, \hat{\mathbf{x}}$. As prescribed by the form of the generative model, we expect the update function associated to the loss $g(., \mathbf{y})$ to be correlated with the matrix $\mathbf{A}$, thus preventing a direct application of the SE equations of Theorem 4, and requiring the results of Lemma4.

2.3.2 Multilayer generalized linear estimation

Consider now the problem of recovering a vector $\mathbf{x}_0$ from a more complex generative model involving a multilayer neural network with random weights:

$$\mathbf{y} = \phi_L(\mathbf{A}_L\phi_{L-1}\left(\mathbf{A}_{L-1}(...\phi_1(\mathbf{A}_1\mathbf{x}_0))\right)$$

where one has access to the final output $\mathbf{y}$ and would like to reconstruct the intermediate ones and input $\mathbf{x}_0$. For each layer $1 \leqslant l \leqslant L$ the matrix $\mathbf{A}_l \in \mathbb{R}^{N_{l+1}\times N_l}$ has i.i.d. $\mathbf{N}(0, \frac{1}{N_l})$ with $N_{l+1}/N_l = \delta_l$. The idea is to solve this sequentially using asymmetric AMP iterations similar to the one presented in the previous section. This approach was originally proposed in [188] under the name multilayer AMP (MLAMP). For any $1 \leqslant l \leqslant L+1$, define

$$\begin{aligned} \mathbf{x}_l &= \phi_{l-1}(\mathbf{A}_{l-1}\phi_{l-2}(..\phi_1(\mathbf{A}_1\mathbf{x}_0))), \\ \text{such that} \quad \mathbf{x}_{l+1} &= \phi_l(\mathbf{A}_l\mathbf{x}_l) \quad \text{and} \quad \mathbf{x}_{L+1} = \mathbf{y} \end{aligned}$$

The intuition is the following : each $\mathbf{x}_l$ is then estimated using the asymmetric AMP corresponding to the problem

$$\hat{\mathbf{x}}_l = \arg\min_{\mathbf{x}\in\mathbb{R}^{N_l}} \; g_l(\mathbf{A}_l\mathbf{x}, \mathbf{y}_l) + f_l(\mathbf{x})$$

the output of which is used to estimate the next, i.e., $\mathbf{y}_l = \hat{\mathbf{x}}_{l+1}$, whose statistical properties are given by the SE equations. The complete derivation of the iteration involves writing the belief-propagation (BP) equations on the factor graph corresponding to the multilayer inference problem, capturing all the interactions between the different iterates. These SE equations were derived heuristically in [188] for Bayes-optimal inference, and this paper proves them in the generic case.

2.3.3 Spiked matrix with generative prior

In the same spirit as the composition of generalized linear models defining MLAMP, different tasks can be composed to obtain richer instances of inference problems. For instance in [15], the reconstruction of a low-rank matrix under a generative prior is considered using an AMP iteration. A rank-one matrix is observed, blurred by Gaussian noise:

$$\mathbf{Y} = \sqrt{\frac{\lambda}{d}}\mathbf{v}_0\mathbf{v}_0^\top + \mathbf{W}$$

where $\mathbf{W} \in GOE(N)$, and the vector $\mathbf{v}_0 \in \mathbb{R}^N$ is assumed to be generated from a multilayer neural network with random weights

$$\mathbf{v}_0 = \phi_L(\mathbf{A}_L\phi_{L-1}\left(\mathbf{A}_{L-1}(...\phi_1(\mathbf{A}_1\mathbf{x}_0))\right)$$

for a given ground truth vector $\mathbf{x}_0 \in \mathbb{R}^{N_1}$, matrices $\{\mathbf{A}_l \in \mathbb{R}^{N_{l+1}\times N_l}\}_{1\leqslant l\leqslant L}$ and non-linearities $\{\phi_l\}_{1\leqslant l\leqslant L}$. The AMP iteration to estimate $\mathbf{v}_0$ from $\mathbf{Y}$ was first proposed in [241, 75], and takes the form of a symmetric AMP (2.6). Similarly to MLAMP, the output of this iteration can then be used as input, leading to the AMP iteration proposed in [14], which corresponds to the AMP iteration (2.8). This paper proves the state evolution equations for this iteration.

2.3.4 An example with matrix-valued variables

Matrix valued variables are encountered in scenarios such as committee machines [15] or multiclass learning problems [178], or more generically when a finite ensemble of predictors is learned. Consider the matrix-valued extension of the generalized linear estimation problem Eq.(2.27).

$$\begin{aligned}\hat{\mathbf{X}} &\in \underset{\mathbf{X}\in\mathbb{R}^{N\times q}}{\arg\min}\, g(\mathbf{A}\mathbf{X}, \mathbf{Y}) + f(\mathbf{X})\\ \text{where} \quad \mathbf{Y} &= \phi(\mathbf{A}\mathbf{X}_0))\end{aligned}$$

where $\mathbf{X}_0 \in \mathbb{R}^{N\times q}$ and $q \in \mathbb{N}$ is kept finite. The SE equations for the asymmetric AMP with matrix valued-variables are included in the result of [135]. This can be directly generalized to a multilayer matrix inference problem by considering a generative model of the form

$$\mathbf{Y} = \phi_L(\mathbf{A}_L\phi_{L-1}\left(\mathbf{A}_{L-1}(...\phi_1(\mathbf{A}_1\mathbf{X}_0))\right)$$

and successive application of the matrix-valued asymmetric AMP as proposed for MLAMP in Section 2.3.2. The state evolution equations for this problem is included in our framework using the results from Section 2.2.3.

2.3.5 An example with structured random matrices

Consider a generalized linear inference task where the data is now represented by a Gaussian matrix with a covariance $\mathbf{\Sigma} \neq \mathbf{I}_d$. This can be dealt with using the non-separable framework. Assuming the covariance matrix is full-rank, we can equivalently work with the variable $\tilde{\mathbf{x}} = \mathbf{\Sigma}^{1/2}\mathbf{x}$, and solve

$$\underset{\tilde{\mathbf{x}}}{\arg\min}\, g(\tilde{\mathbf{A}}\tilde{\mathbf{x}}, \mathbf{y}) + f(\mathbf{\Sigma}^{-1/2}\tilde{\mathbf{x}}).$$

where $\tilde{A}$ is now ana i.i.d. Gaussian matrix. This will modify the update function associated to f, becoming $f(\mathbf{\Sigma}^{-1/2}.)$, which is non-separable, even if the function f is initially assumed to be

separable. The validity of the SE equations for this case follows from the results of [37]. This manipulation can also be done on any layer of MLAMP, for a given set of covariance matrices $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_{\boldsymbol{L}}$ associated to each random matrix $\mathbf{A}_1, ..., \mathbf{A}_L$, with vector or matrix-valued variables. The validity of the SE equations in this case follows from the results of this paper. In the convex GLM case (2-layer), the fixed point of the state evolution equations with a generic covariance gives the same result as (a particular case of) the exact asymptotics recently proposed in [176] to study different feature maps in generalized linear models.

2.3.6 An example of spatial coupling with non-separable non-linearities

Here we briefly describe an inference problem recently studied in [178] that can be solved using spatial coupling on a non-separable AMP iteration. Consider the problem of classifying a high-dimensional Gaussian mixture with a finite number K of clusters, described by the joint density

$$P(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} y_k \pi_k \mathbf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\mathbf{x} \in \mathbb{R}^d$ is a sample, $\mathbf{y} \in \mathbb{R}^K$ is a binary label vector, $\{\pi_k\}_k$ are the cluster probabilities such that $\sum_{k=1}^{K} \pi_k = 1$, $\{\boldsymbol{\mu}_k\}_{1 \leqslant k \leqslant K}$ are the means and $\{\boldsymbol{\Sigma}_k\}_{1 \leqslant k \leqslant K}$ are positive definite covariances, using a convex generalized linear model, i.e.,

$$\mathbf{X} \in \underset{\mathbf{X} \in \mathbb{R}^{d \times K}}{\arg\min}\, g(\mathbf{AX}, \mathbf{Y}) + f(\mathbf{X})$$

where $\mathbf{Y} \in \mathbb{R}^{N \times K}$ is the concatenated matrix of one-hot encoded labels. The matrix $\mathbf{A}$ representing N samples of the Gaussian mixture can be written as a block diagonal matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{Z}_1 \boldsymbol{\Sigma}_1^{1/2} & & & \\ & \mathbf{Z}_2 \boldsymbol{\Sigma_2}^{1/2} & & \\ & & \dots & \\ & & & \mathbf{Z}_K \boldsymbol{\Sigma_K}^{1/2} \end{bmatrix} \in \mathbb{R}^{N \times Kd}$$

where the $\mathbf{Z}_k \in \mathbb{R}^{N_k \times d}$ are i.i.d. $\mathbf{N}(0, \frac{1}{d})$ independent matrices, with N_k the number of samples coming from each cluster. This type of matrix can be embedded into an AMP iteration using the spatial coupling technique to handle the block structure and the non-separable framework to deal with the covariances on each block. The validity of the SE equations for the combination of spatial coupling and non-separable effects is proven by this paper. This is also an example where the teacher distribution is independent of the Gaussian matrices that will appear in the AMP iteration, as the multinomial distribution prescribing cluster membership is independent of the Gaussian cloud of each cluster.

2.4 Perspectives

We have shown that AMP algorithms can be unified in an intuitive way by means of an oriented graph, and that this representation leads to a modular, effective and extended proof of state evolution equations. Several problems follow from the results presented here.

Connecting back to the factor graph. We do not relate our proposed graphical representation of the AMP iterations with the factor graphs of the probabilistic inference problems that generated them. Understanding this relation would clarify the statistical inference problems that can be solved using AMP iterations. The applications that motivated this paper use our framework with only very simple graphs—line graphs, sometimes with a loop. However, the framework accepts much more complicated graphs, potentially with more loops. In future work, we hope to explore the new statistical problems and AMP iterations that can be analyzed using these graphs.

Rotationally invariant matrices. As shown in [242, 97, 222, 95], the Gaussian conditioning method at the core of AMP proofs can be reproduced with right rotationally invariant matrices with generic spectrum. Extending the results of the present paper to this family of matrices requires finding the appropriate form of the graph iteration and is an open problem.

Universality and finite size corrections. State evolution proofs are amenable to both finite size analysis [251, 180] and universality proofs [27, 62]. Although both problems were tackled in simpler settings in these papers, their techniques could be combined with the embedding proposed in the proof of Theorem 4 to prove finite size rates and universality properties for any graph supported AMP.

Chapter 3

Proofs for the Graph-based AMP iterations

3.1 Changing time indices

Here we show how the time index convention usually encountered in earlier instances of the asymmetric AMP iteration can be recovered from the one used in this proof. Consider two successive iterations of the asymmetric AMP (2.5):

$$\begin{aligned}
\mathbf{x}_{\vec{e}}^{t+1} &= \mathbf{A}_{\vec{e}}\mathbf{m}_{\vec{e}}^{t} - b_{\vec{e}}^{t}\mathbf{m}_{\overleftarrow{e}}^{t-1}, & \mathbf{x}_{\vec{e}}^{t} &= \mathbf{A}_{\vec{e}}\mathbf{m}_{\vec{e}}^{t-1} - b_{\vec{e}}^{t-1}\mathbf{m}_{\overleftarrow{e}}^{t-2}, \\
\mathbf{m}_{\vec{e}}^{t} &= f_{\vec{e}}^{t}\left(\mathbf{x}_{\overleftarrow{e}}^{t}\right), & \mathbf{m}_{\vec{e}}^{t-1} &= f_{\vec{e}}^{t-1}\left(\mathbf{x}_{\overleftarrow{e}}^{t-1}\right), \\
\mathbf{x}_{\overleftarrow{e}}^{t+1} &= \mathbf{A}_{\vec{e}}^{\top}\mathbf{m}_{\overleftarrow{e}}^{t} - b_{\overleftarrow{e}}^{t}\mathbf{m}_{\vec{e}}^{t-1}, & \mathbf{x}_{\overleftarrow{e}}^{t} &= \mathbf{A}_{\vec{e}}^{\top}\mathbf{m}_{\overleftarrow{e}}^{t-1} - b_{\overleftarrow{e}}^{t-1}\mathbf{m}_{\vec{e}}^{t-2}, \\
\mathbf{m}_{\overleftarrow{e}}^{t} &= f_{\overleftarrow{e}}^{t}\left(\mathbf{x}_{\vec{e}}^{t}\right) & \mathbf{m}_{\overleftarrow{e}}^{t-1} &= f_{\overleftarrow{e}}^{t-1}\left(\mathbf{x}_{\vec{e}}^{t-1}\right)
\end{aligned} \tag{3.1}$$

which requires initializing both $\mathbf{x}_{\vec{e}}$ and $\mathbf{x}_{\overleftarrow{e}}$, and updates them simultaneously at each iteration. We see that to evaluate $\mathbf{x}_{\vec{e}}^{t+1}$ (resp. $\mathbf{x}_{\overleftarrow{e}}^{t+1}$), we only need the previous value of $\mathbf{x}_{\overleftarrow{e}}^{t}$ (resp. $\mathbf{x}_{\vec{e}}^{t}$) and $\mathbf{x}_{\vec{e}^{t-1}}$ (resp. $\mathbf{x}_{\overleftarrow{e}}^{t-1}$). Thus only half of the iterates can be computed, independently of the other half, using the following formulae (setting the other update functions to zero):

$$\begin{aligned}
\mathbf{x}_{\overleftarrow{e}}^{2t+1} &= \mathbf{A}_{\vec{e}}^{\top}\mathbf{m}_{\overleftarrow{e}}^{2t} - b_{\overleftarrow{e}}^{2t}\mathbf{m}_{\vec{e}}^{2t-1}, \\
\mathbf{m}_{\overleftarrow{e}}^{2t} &= f_{\overleftarrow{e}}^{2t}\left(\mathbf{x}_{\vec{e}}^{2t}\right), \\
\mathbf{x}_{\vec{e}}^{2t} &= \mathbf{A}_{\vec{e}}\mathbf{m}_{\vec{e}}^{2t-1} - b_{\vec{e}}^{2t-1}\mathbf{m}_{\overleftarrow{e}}^{2t-2}, \\
\mathbf{m}_{\vec{e}}^{2t-1} &= f_{\vec{e}}^{2t-1}\left(\mathbf{x}_{\overleftarrow{e}}^{2t-1}\right)
\end{aligned} \tag{3.2}$$

which only requires one value at initialization and at each iteration. The usual time indices found in , e.g., [37] are then recovered with the following mapping:

$$\begin{aligned}
\mathbf{x}_{\overleftarrow{e}}^{2t+1} &= \mathbf{u}^{t+1} \\
\mathbf{x}_{\vec{e}}^{2t} &= \mathbf{v}^{t} \\
f_{\overleftarrow{e}}^{2t}(.) &= g_t(.) \\
f_{\vec{e}}^{2t-1}(.) &= e_t(.)
\end{aligned}$$

Note that this simplification is specific to the graph structure underlying the asymmetric AMP iteration.

3.2 Matrix-valued symmetric AMP iterations with non-separable non-linearities

3.2.1 State evolution description

In this section, we present the state evolution equations for a symmetric AMP iteration with non-separable non-linearities and matrix-valued variables. This is an extension of the results of [135, 37]. This result underlies the proof of state evolution equations for graph-based AMP iterations.

Consider an initial (deterministic) matrix $\mathbf{X}^0 \in \mathbb{R}^{N\times q}$ and a sequence of deterministic functions $\{f^t : \mathbb{R}^{N\times q} \to \mathbb{R}^{N\times q}\}_{t\in\mathbb{N}}$. For the reader's convenience, we recall here the symmetric AMP iteration (2.9)-(2.11).

Symmetric AMP iteration. Let $\mathbf{X}^0 \in \mathbb{R}^{N\times q}$ and define recursively,

$$\mathbf{X}^{t+1} = \mathbf{A}\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top \qquad \in \mathbb{R}^{N\times q}, \tag{3.3}$$

$$\mathbf{M}^t = f^t(\mathbf{X}^t) \qquad \in \mathbb{R}^{N\times q}, \tag{3.4}$$

$$\mathbf{b}^t = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial f_i^t}{\partial \mathbf{X}_i}(\mathbf{X}^t) \qquad \in \mathbb{R}^{q\times q}. \tag{3.5}$$

where $\mathbf{b}^t$ is the Onsager correction term. We now list the necessary assumptions.

Assumptions.

(B1) $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a GOE(N) matrix, i.e., $\mathbf{A} = \mathbf{G} + \mathbf{G}^\top$ for $\mathbf{G} \in \mathbb{R}^{N\times N}$ with i.i.d. entries $G_{ij} \sim \mathbf{N}(0, 1/(2N))$.

(B2) For each $t \in \mathbb{N}$, $f^t : \mathbb{R}^{N\times q} \to \mathbb{R}^{N\times q}$ is pseudo-Lipschitz of order k, uniformly in N.

(B3) $\|\mathbf{X}^0\|_F/\sqrt{N}$ converges to a finite constant as $N \to \infty$.

(B4) The following limit exists and is finite:

$$\lim_{N\to\infty} \frac{1}{N} f^0(\mathbf{X}^0)^\top f^0(\mathbf{X}^0) \in \mathbb{R}^{q\times q} \tag{3.6}$$

(B5) For any $t \in \mathbb{N}_{>0}$ and any $\boldsymbol{\kappa} \in \mathcal{S}_q^+$, the following limit exists and is finite:

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[f^0(\mathbf{X}^0)^\top f^t(\mathbf{Z})\right] \in \mathbb{R}^{q\times q} \tag{3.7}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times q}$, $\mathbf{Z} \sim \mathbf{N}(0, \boldsymbol{\kappa} \otimes \mathbf{I}_N)$.

(B6) For any $s, t \in \mathbb{N}_{>0}$ and any $\boldsymbol{\kappa} \in \mathcal{S}_{2q}^+$, the following limit exists and is finite:

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[f^s(\mathbf{Z}^s)^\top f^t(\mathbf{Z}^t)\right] \in \mathbb{R}^{q\times q} \tag{3.8}$$

where $(\mathbf{Z}^s, \mathbf{Z}^t) \in (\mathbb{R}^{N\times q})^2$,$(\mathbf{Z}^s, \mathbf{Z}^t) \sim \mathbf{N}(0, \boldsymbol{\kappa} \otimes \mathbf{I}_N)$.

Under these assumptions, we define the *state evolution* iteration related to the AMP iteration (3.3)-(3.5).

Definition 6 (state evolution iterates)**.** *The state evolution iterates are composed of one infinite-dimensional array $(\boldsymbol{\kappa}^{s,r})_{r,s>0}$ of real matrices. This array is generated as follows. Define the first state evolution iterate*

$$\boldsymbol{\kappa}^{1,1} = \lim_{N\to\infty} \frac{1}{N} f^0(\mathbf{X}^0)^\top f^0(\mathbf{X}^0) \tag{3.9}$$

Recursively, once $\boldsymbol{\kappa}^{s,r}, 0 \leqslant s,r \leqslant t$ are defined for some $t \geqslant 1$, take $\mathbf{Z}^0 = \mathbf{X}^0$ and $(\mathbf{Z}^1,\dots,\mathbf{Z}^t) \in (\mathbb{R}^{n\times q})^t$ a centered Gaussian vector of covariance $(\boldsymbol{\kappa}^{s,r})_{s,r\leqslant t}\otimes \mathbf{I}_N$. We then define new state evolution iterates

$$\boldsymbol{\kappa}^{t+1,s+1} = \boldsymbol{\kappa}^{s+1,t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[f^s(\mathbf{Z}^s)^\top f^t(\mathbf{Z}^t)\right], \qquad s\in\{0,\dots,t\}.$$

The following property then holds for the AMP iteration (3.3)-(3.5).

Theorem 5. *Assume (B1)-(B6). Define, as above, $\mathbf{Z}^0 = \mathbf{X}^0$ and $(\mathbf{Z}^1,\dots,\mathbf{Z}^t) \in (\mathbb{R}^{N\times q})^t$ a centered Gaussian vector of covariance $(\boldsymbol{\kappa}^{s,r})_{s,r\leqslant t}\otimes \mathbf{I}_N$. Then for any sequence $\Phi_N : (\mathbb{R}^{N\times q})^{t+1} \to \mathbb{R}$ of pseudo-Lipschitz functions,*

$$\Phi_N\left(\mathbf{X}^0,\mathbf{X}^1,\dots,\mathbf{X}^t\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\Phi_N\left(\mathbf{Z}^0,\mathbf{Z}^1,\dots,\mathbf{Z}^t\right)\right].$$

Given the above result, we can expect the Onsager correction $\mathbf{b}^t$ to verify

$$\mathbf{b}^t \overset{\mathrm{P}}{\simeq} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{Z}_i}(\mathbf{Z}^t)\right] \in \mathbb{R}^{q\times q}. \tag{3.10}$$

where $\mathbf{Z}^t \sim \mathbf{N}(0, \boldsymbol{\kappa}_{t,t}\otimes \mathbf{I}_n)$. In fact, similarly to [37], Theorem 5 can be shown to hold for the AMP iteration ((3.3)-(3.5)) with any estimator $\hat{\mathbf{b}}^t$ satisfying

$$\hat{\mathbf{b}}^t(\mathbf{X}^0,\hat{\mathbf{M}}^0,...,\hat{\mathbf{M}}^{t-1},\hat{\mathbf{X}}^t) \overset{\mathrm{P}}{\simeq} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{Z}_i}(\mathbf{Z}^t)\right] \in \mathbb{R}^{q\times q}. \tag{3.11}$$

. The state evolution actually hold for the AMP Eq.(2.9-2.11) with any estimator $\hat{\mathbf{b}}_t$ converging in probability to the expectation on the r.h.s. of Eq.((3.10)). This is formalized in the following corollary:

Theorem 6. *Consider the AMP iteration*

$$\hat{\mathbf{X}}^{t+1} = \mathbf{A}\hat{\mathbf{M}}^t - \hat{\mathbf{M}}^{t-1}\hat{\mathbf{b}}_t^\top \qquad \in \mathbb{R}^{N\times q} \tag{3.12}$$

$$\hat{\mathbf{M}}^t = f^t(\hat{\mathbf{X}}^t) \qquad \in \mathbb{R}^{N\times q} \tag{3.13}$$

initialized with $\mathbf{X}^0$ as Eq. (2.9-2.11), and where $\hat{\mathbf{b}}_t(\mathbf{X}^0,\hat{\mathbf{M}}^0,...,\hat{\mathbf{M}}^{t-1},\hat{\mathbf{X}}^t)$ is an estimator of $\mathbf{b}^t$. Under the set of assumptions (A1-A6), and provided the estimator $\hat{b}_t$ verifies

$$\hat{\mathbf{b}}_t(\mathbf{X}^0,\hat{\mathbf{M}}^0,...,\hat{\mathbf{M}}^{t-1},\hat{\mathbf{X}}^t) \overset{\mathrm{P}}{\simeq} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f_t^i}{\partial \mathbf{Z}_i}(\mathbf{Z}^t)\right] \in \mathbb{R}^{q\times q}. \tag{3.14}$$

then for any $t\in\mathbb{N}$

$$\lim_{N\to\infty}\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{X}}^{t+1}-\mathbf{X}^{t+1}\right\|_F \overset{\mathrm{P}}{\simeq} 0, \quad \lim_{N\to\infty}\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^{t}-\mathbf{M}^{t}\right\|_F \overset{\mathrm{P}}{\simeq} 0 \tag{3.15}$$

and the iterates $\hat{\mathbf{M}}^t, \hat{\mathbf{X}}^t$ verify the state evolution equations.

The proof of this corollary is also provided in Appendix 3.3.

3.2.2 Application: proof of Theorem 4

In Section 2.2.2, we have seen that the graph AMP iteration (2.2)-(2.4) can be rewritten as a symmetric AMP iteration of the form (2.9)-(2.11). Here, we check that applying Theorem 5 on the symmetric iteration after performing the reduction indeed gives Theorem 4.

Define the state evolution iterates as in Definition 6. Here, due to the expression (2.12) of the non-linearities, the state evolution iterates are diagonal:

$$\boldsymbol{\kappa}^{1,1} = \lim_{N\to\infty} \frac{1}{N} \begin{pmatrix} \left\| f^0_{\overrightarrow{e}_1}((\mathbf{x}^0_{\overrightarrow{e}})_{\overrightarrow{e}:\overrightarrow{e}\to\overrightarrow{e}_1}) \right\|^2 & & 0 \\ & \ddots & \\ 0 & & \left\| f^0_{\overleftarrow{e}_m}((\mathbf{x}^0_{\overrightarrow{e}})_{\overrightarrow{e}:\overrightarrow{e}\to\overleftarrow{e}_m}) \right\|^2 \end{pmatrix} \tag{3.16}$$

and

$$\boldsymbol{\kappa}^{t+1,s+1} = \boldsymbol{\kappa}^{s+1,t+1} = \lim_{N\to\infty} \frac{1}{N} \begin{pmatrix} \mathbb{E} f^s_{\overrightarrow{e}_1}(\dots)^\top f^t_{\overrightarrow{e}_1}(\dots) & & 0 \\ & \ddots & \\ 0 & & \mathbb{E} f^s_{\overleftarrow{e}_m}(\dots)^\top f^t_{\overleftarrow{e}_m}(\dots) \end{pmatrix}.$$

Let $\mathbf{Z}^t \in \mathbb{R}^{N\times q}$ be the variable from Definition 6. Decompose

$$\mathbf{Z}^t = \begin{pmatrix} \mathbf{Z}^t_{\overrightarrow{e}_1} & & * \\ & \ddots & \\ * & & \mathbf{Z}^t_{\overleftarrow{e}_m} \end{pmatrix}.$$

where $\mathbf{Z}^t_{(v,w)} \in \mathbb{R}^{n_w}$. The diagonal structure of the state evolution iterates means that $\mathbf{Z}^t_{\overrightarrow{e}}$ and $\mathbf{Z}^t_{\overrightarrow{e}'}$ are independent when $\overrightarrow{e} \neq \overrightarrow{e}'$. We thus find that

$$\boldsymbol{\kappa}^{s,t} = \begin{pmatrix} \boldsymbol{\kappa}^{s,t}_{\overrightarrow{e}_1} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{\kappa}^{s,t}_{\overleftarrow{e}_m} \end{pmatrix},$$

where the $\boldsymbol{\kappa}^{s,t}_{\overrightarrow{e}}$ are those defined in Section 2.2 and the variables $\mathbf{Z}^t_{\overrightarrow{e}}$ are the same as those defined in Section 2.2.

These elements show that Theorem 4 follows from the application of Theorem 5.

3.3 Proof of Theorem 5

Once the concentration lemmas of Appendix 3.5 are established for matrix valued-variables, the proof follows closely that of [37]. We include the main steps (with minor changes) for completeness nonetheless.

As an intermediate step, we introduce the following AMP iteration initialized with $X^0 \in \mathbb{R}^{N\times q}$:

$$\mathbf{X}^{t+1} = \mathbf{A}\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top \qquad \in \mathbb{R}^{N\times q} \tag{3.17}$$

$$\mathbf{M}^t = f^t(\mathbf{X}^t) \qquad \in \mathbb{R}^{N\times q}, \tag{3.18}$$

$$\mathbf{b}_t = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f^t_i}{\partial \mathbf{Z}_i}(\mathbf{Z}^t)\right] \qquad \in \mathbb{R}^{q\times q}. \tag{3.19}$$

where the Onsager term has been replaced by the expectation in Eq.(3.10) using the state evolution recursion, i.e., $\mathbf{Z}^t \in \mathbb{R}^{N\times q} \sim \mathbf{N}(0, \boldsymbol{\kappa}_{t,t} \otimes \mathbf{I}_N)$.
We denote this recursion with the shorthand $\{\mathbf{X}^t, \mathbf{M}^t | f^t, \mathbf{X}^0\}$. The following lemma is an analog of Theorem 5 for the iteration (3.17)-(3.19).

Lemma 5. *Define, as above,* $\mathbf{Z}^0 = \mathbf{X}^0$ *and* $(\mathbf{Z}^1, \dots, \mathbf{Z}^t) \in (\mathbb{R}^{N\times q})^t$ *a centered Gaussian vector of covariance* $\begin{pmatrix} \boldsymbol{\kappa}^{1,1} & \dots & \boldsymbol{\kappa}^{1,t} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}^{t,1} & \dots & \boldsymbol{\kappa}^{t,t} \end{pmatrix} \otimes \mathbf{I}_N$. *Then for any sequence* $\Phi_N : (\mathbb{R}^{N\times q})^{t+1} \to \mathbb{R}$ *of pseudo-Lipschitz functions, the iterates of* (3.17)-(3.19) *satisfy*

$$\Phi_N\left(\mathbf{X}^0, \mathbf{X}^1, \dots, \mathbf{X}^t\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\Phi_N\left(\mathbf{Z}^0, \mathbf{Z}^1, \dots, \mathbf{Z}^t\right)\right].$$

3.3.1 Proof outline and intermediate lemmas

The main idea is to analyze an iteration that behaves well under Gaussian conditioning and that asymptotically approximates (3.17)-(3.19).

Matrix LoAMP. We consider the following iteration, a matrix-valued version of the LoAMP iteration introduced in [37]. The sequence of functions f^t and initialization $\mathbf{X}^0$ are the same as for the AMP orbit $\{\mathbf{X}^t, \mathbf{M}^t | f^t, \mathbf{X}^0\}$. Initialize $\mathbf{Q}^0 = f^0(\mathbf{X}^0)$, and recursively define

$$\mathbf{H}^{t+1} = \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}} \mathbf{A} \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}} \mathbf{Q}^t + \boldsymbol{\mathcal{H}}_{t-1} \boldsymbol{\alpha}^t \quad \in \mathbb{R}^{N\times q}, \tag{3.20}$$

$$\mathbf{Q}^t = f^t(\mathbf{H}^t) \quad \in \mathbb{R}^{N\times q}, \tag{3.21}$$

where at each step, the matrices $\boldsymbol{\mathcal{Q}}_{t-1}, \boldsymbol{\alpha}^t, \boldsymbol{\mathcal{H}}_{t-1}$ are defined as

$$\boldsymbol{\mathcal{Q}}_{t-1} = \left[\mathbf{Q}^0 | \mathbf{Q}^1 | \dots | \mathbf{Q}^{t-1}\right] \quad \in \mathbb{R}^{N\times tq}, \tag{3.22}$$

$$\boldsymbol{\alpha}^t = (\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1} \boldsymbol{\mathcal{Q}}_{t-1}^\top \mathbf{Q}^t \quad \in \mathbb{R}^{tq\times q}, \tag{3.23}$$

$$\boldsymbol{\mathcal{H}}_{t-1} = \left[\mathbf{H}^1 | \mathbf{H}^2 | \dots | \mathbf{H}^t\right] \quad \in \mathbb{R}^{N\times tq}, \tag{3.24}$$

$\mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}} = \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1} \boldsymbol{\mathcal{Q}}_{t-1}^\top$ is the orthogonal projector on the subspace spanned by the columns of $\boldsymbol{\mathcal{Q}}_{t-1}$, and $\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}} = \mathbf{I}_N - \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}$.
We denote this recursion with the shorthand $\{\mathbf{H}^t, \mathbf{Q}^t | f^t, \mathbf{X}^0\}$. The inverse $(\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}$ in the projector may not always be properly defined if $\boldsymbol{\mathcal{Q}}_{t-1}$ is either rank-deficient or has vanishing singular values. We thus introduce the following assumption as in [37], which ensures the proper definition of the projector.

Assumption 1 (Non-degeneracy)**.** *We say that the LoAMP iterates satisfy the non-degeneracy assumption if :*

- *almost surely, for all* t *and all* $N \geqslant t$, $\boldsymbol{\mathcal{Q}}_{t-1}$ *has full column rank.*
- *for all* t, *there exists some constant* $c_t > 0$ *—independent of* N*—such that almost surely, there exists* N_0 *(random) such that, for* $N \geqslant N_0$, $\sigma_{\min}(\boldsymbol{\mathcal{Q}}_{t-1})/\sqrt{N} \geqslant c_t > 0$.

We now study the LoAMP iteration, starting with the non-degenerate case.

The non-degenerate case. The following lemma gives the distribution of the Long-AMP iterates when conditioned on the previous ones.

Lemma 6. *Consider the LoAMP iteration $\{\mathbf{H}^t, \mathbf{Q}^t | f_t, \mathbf{X}^0\}$ and assume it satisfies the non-degeneracy assumption. For any $t \in \mathbb{N}$, let $\mathfrak{S}_t$ be the σ-algebra generated by the collection of random variables $\mathbf{H}^1, \mathbf{H}^2, ..., \mathbf{H}^t$. Then*

$$\mathbf{H}^{t+1}|_{\mathfrak{S}_t} \overset{d}{=} \mathbf{P}^{\perp}_{\mathcal{Q}_{t-1}} \tilde{\mathbf{A}} \mathbf{P}^{\perp}_{\mathcal{Q}_{t-1}} \mathbf{Q}^t + \boldsymbol{\mathcal{H}}_{t-1} \boldsymbol{\alpha}^t \tag{3.25}$$

where $\tilde{\mathbf{A}}$ is a copy of $\mathbf{A}$ independent of $\mathfrak{S}_t$.

The next lemma characterizes the high-dimensional geometry and distribution of the LoAMP iterates, notably that they verify the state evolution equations.

Lemma 7. *Consider the LoAMP recursion $\{\mathbf{H}^t, \mathbf{Q}^t | f_t, \mathbf{X}^0\}$ and suppose it satisfies the non-degeneracy assumption. Then*

a) for all $0 \leqslant s, r \leqslant t$,

$$\frac{1}{N}(\mathbf{H}^{s+1})^\top \mathbf{H}^{r+1} \overset{\mathrm{P}}{\backsimeq} \frac{1}{N}(\mathbf{Q}^s)^\top \mathbf{Q}^r \quad \in \mathbb{R}^{q \times q}\,, \tag{3.26}$$

b) for any $t \in \mathbb{N}$, for any sequence of uniformly order-k pseudo-Lipschitz functions $\{\phi_N : (\mathbb{R}^{N \times q})^{t+2} \to \mathbb{R}\}$,

$$\Phi_N(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^{t+1}) \overset{\mathrm{P}}{\backsimeq} \mathbb{E}[\Phi_N(\mathbf{X}^0, \mathbf{Z}^1, ..., \mathbf{Z}^{t+1})] \tag{3.27}$$

where

$$(\mathbf{Z}^1, ..., \mathbf{Z}^{t+1}) \sim \mathbf{N}(0, (\boldsymbol{\kappa}^{s,r})_{s,r \leqslant t} \otimes \mathbf{I}_N) \tag{3.28}$$

The next two lemmas show that the iterates of the Long-AMP recursion are arbitrary close to those of the original symmetric AMP in the high-dimensional limit.

Lemma 8. *For each iteration t of the LoAMP iteration $\{\mathbf{H}^t, \mathbf{Q}^t | f^t, \mathbf{X}^0\}$, consider the recursion*

$$\hat{\mathbf{H}}^{t+1} = \mathbf{A}\mathbf{Q}^t - \mathbf{Q}^{t-1}(\mathbf{b}^t)^\top \quad \textit{where} \quad \mathbf{b}^t = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} \frac{\partial f_i^t}{\partial \mathbf{Z}_i}(\mathbf{Z}^t)\right] \quad \in \mathbb{R}^{q \times q} \tag{3.29}$$

$$\mathbf{Q}^t = f^t(\mathbf{H}^t) \tag{3.30}$$

where we take $\hat{\mathbf{H}}^1 = \mathbf{A}\mathbf{Q}^0$ and $\mathbf{Z}^t \sim \mathbf{N}(0, \mathbf{K}_{t,t} \otimes \mathbf{I}_N)$ with $\mathbf{K}_{t,t}$ defined by the state evolution. Then for any $t \in \mathbb{N}$, $\frac{1}{\sqrt{N}}\left\|\mathbf{H}^{t+1} - \hat{\mathbf{H}}^{t+1}\right\|_F \xrightarrow[N\to\infty]{P} 0$.

Lemma 9. *Consider the symmetric AMP iteration $\{\mathbf{X}^t, \mathbf{M}^t | f_t, \mathbf{X}^0\}$ and the LongAMP iteration $\{\mathbf{H}^t, \mathbf{Q}^t | f_t, \mathbf{X}^0\}$. Suppose that LongAMP satisfies the non-degeneracy assumption. Then for any $t \in \mathbb{N}$,*

$$\frac{1}{\sqrt{N}}\left\|\mathbf{H}^{t+1} - \mathbf{X}^{t+1}\right\|_F \xrightarrow[N\to\infty]{P} 0 \quad \textit{and} \quad \frac{1}{\sqrt{N}}\left\|\mathbf{Q}^t - \mathbf{M}^t\right\|_F \xrightarrow[N\to\infty]{P} 0 \tag{3.31}$$

Combining the previous results, and assuming the non-degeneracy is verified, Lemma 5 holds true.

Relaxing the non-degeneracy hypothesis This paragraph shows how the non-degeneracy assumption is relaxed using a perturbative argument as done in [37]. Define the randomly perturbed functions

$$f^t_{\epsilon\mathbf{Y}^t} = f^t(.) + \epsilon\mathbf{Y}^t \tag{3.32}$$

where $\mathbf{Y}^t \in \mathbb{R}^{N\times q}$ is a matrix with i.i.d. $\mathbf{N}(0,1)$ entries independent of the original matrix $\mathbf{A}$. We denote $\mathbf{Y}$ the set of random matrices $(\mathbf{Y}^0, \mathbf{Y}^1, ..., \mathbf{Y}^t) \in (\mathbb{R}^{N\times q})^{t+1}$.

Lemma 10. *The AMP iteration defined with the functions $f^t_{\epsilon\mathbf{Y}}$ and initialized with $\mathbf{X}^0$ verifies Assumptions (B4) – (B6). Furthermore, define the associated state evolution iteration $\{\boldsymbol{\kappa}^{s,t}_\epsilon | f^t_{\epsilon\mathbf{Y}}, \mathbf{X}^0\}$, initialized with*

$$\boldsymbol{\kappa}^{1,1}_\epsilon = \lim_{N\to\infty} \frac{1}{N}(f^0_{e\mathbf{Y}}(\mathbf{X}^0))^\top (f^0_{e\mathbf{Y}}(\mathbf{X}^0)) \tag{3.33}$$

and

$$\boldsymbol{\kappa}^{s+1,t+1}_\epsilon = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[(f^s_{\epsilon\mathbf{Y}}(\mathbf{Z}^{\epsilon,s})^\top f^t_{\epsilon\mathbf{Y}}(\mathbf{Z}^{\epsilon,t})\right] \tag{3.34}$$

where $(\mathbf{Z}^{\epsilon,1}, ..., \mathbf{Z}^{\epsilon,t}) \sim \mathbf{N}(0, (\boldsymbol{\kappa}^{s,r})^\epsilon_{s,r\leqslant t} \otimes \mathbf{I}_N)$ and the expectations are taken w.r.t. $\mathbf{Z}^{\epsilon,1}, ..., \mathbf{Z}^{\epsilon,t}$ but not on $\mathbf{Y}$. Then the state evolution $\{\boldsymbol{\kappa}^{s,t}_\epsilon | f^t_{\epsilon\mathbf{Y}}, \mathbf{X}^0\}$ is almost surely non-random.

Lemma 11. *Denote $\boldsymbol{\mathcal{Q}}^{\epsilon\mathbf{Y}}_{t-1}$ the $N \times tq$ matrix associated with the LoAMP iterates $\{\mathbf{H}^{\epsilon\mathbf{Y},t}, \mathbf{Q}^{\epsilon\mathbf{Y},t} | f^t_{\epsilon\mathbf{Y}}, \mathbf{X}^0\}$. Assume $\epsilon > 0$. Then for $N \geqslant t$, the matrix $\boldsymbol{\mathcal{Q}}^{\epsilon\mathbf{Y}}_{t-1}$ almost surely has full column-rank. Furthermore, there exists a constant $c_{t,\epsilon}$, independent of n, such that, almost surely, there exists N_0 (random) such that, for $N \geqslant N_0$, $\sigma_{min}(\boldsymbol{\mathcal{Q}}^{\epsilon\mathbf{Y}}_{t-1})/\sqrt{N} \geqslant c_{t,\epsilon} > 0$.*

The next two lemmas show uniform convergence of the perturbed state evolution averages to the original one when the perturbation vanishes.

Lemma 12. *Let $\{\Phi_N : \mathbb{R}^{N\times tq} \to \mathbb{R}^{q\times q}\}_{N>0}$ be a sequence of uniformly pseudo-Lipschitz functions of order k. Let $\boldsymbol{\kappa}, \tilde{\boldsymbol{\kappa}}$ be two $tq \times tq$ covariance matrices and $\mathbf{Z} \sim \mathbf{N}(0, \boldsymbol{\kappa} \otimes \mathbf{I}_N)$, $\tilde{\mathbf{Z}} \sim \mathbf{N}(0, \tilde{\boldsymbol{\kappa}} \otimes \mathbf{I}_N)$. Then*

$$\lim_{\tilde{\boldsymbol{\kappa}}\to\boldsymbol{\kappa}} \sup_{N\geqslant 1} \mathbb{E}[\Phi_N(\mathbf{Z})] - \mathbb{E}[\Phi_N(\tilde{\mathbf{Z}})] = 0\,. \tag{3.35}$$

Lemma 13. *For any $s, t \geqslant 1$, $\boldsymbol{\kappa}^{s,t}_\epsilon \xrightarrow[\epsilon\to 0]{} \boldsymbol{\kappa}^{s,t}$.*

This last lemma shows that the iterates of the AMP orbit defined with the randomly perturbed functions (3.32), denoted $\{\mathbf{X}^{\epsilon\mathbf{Y},t}, \mathbf{M}^{\epsilon\mathbf{Y},t} | f^t_{\epsilon\mathbf{Y}}, \mathbf{X}^0\}$, is arbitrarily close to the original AMP orbit $\{\mathbf{X}^t, \mathbf{M}^t | f^t, \mathbf{X}^0\}$ when the perturbation is taken to zero.

Lemma 14. *Consider the symmetric AMP orbit defined by $\{\mathbf{X}^t, \mathbf{M}^t | f^t, \mathbf{X}^0\}$ and the corresponding perturbed orbit defined by $\{\mathbf{X}^{\epsilon\mathbf{Y},t}, \mathbf{M}^{\epsilon\mathbf{Y},t} | f^t_{\epsilon\mathbf{Y}}, \mathbf{X}^0\}$. Assume that, for some $t \in \mathbb{N}$. Then there exist functions $h_t(\epsilon)$, $h'_t(\epsilon)$, independent of N, such that*

$$\lim_{\epsilon\to 0} h_t(\epsilon) = \lim_{\epsilon\to 0} h'_t(\epsilon) = 0 \tag{3.36}$$

and for all $\epsilon \leqslant 1$, with high probability,

$$\frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t} - \mathbf{M}^t\right\|_F \leqslant h'_t(\epsilon)\,, \tag{3.37}$$

$$\frac{1}{\sqrt{N}}\left\|\mathbf{X}^{\epsilon\mathbf{Y},t+1} - \mathbf{X}^{t+1}\right\|_F \leqslant h_t(\epsilon)\,. \tag{3.38}$$

Combining these lemmas, we now prove Lemma 5.

3.3.2 Proof of Lemma 5 and Theorem 5

Theorem 5 follows from Lemma 5 similarly to the proof of Corollary 2 from [37].

Proof of Lemma 5. The lemmas presented in the previous section ensure the following:

- Lemma 11 and 5 ensure the AMP iteration defined with randomly perturbed functions verifies the non-degeneracy assumptions and the perturbed state evolution equations, i.e.,

$$\Phi_N\left(\mathbf{X}^0, \mathbf{X}^{\epsilon,1}, \dots, \mathbf{X}^{\epsilon_{\mathbf{Y}},t}\right) \overset{\mathrm{P}}{\backsimeq} \mathbb{E}\left[\Phi_N\left(\mathbf{Z}^{\epsilon,0}, \mathbf{Z}^{\epsilon,1}, \dots, \mathbf{Z}^{\epsilon,t}\right)\right].$$

  for any sequence of pseudo-Lispchitz functions Φ_N, where $(\mathbf{Z}^{\epsilon,0}, \mathbf{Z}^{\epsilon,1}, \dots, \mathbf{Z}^{\epsilon,t})$ are defined as in Eq.(3.33).

- We have shown that the perturbed state evolution converges to the original one for vanishing perturbations, i.e.,

$$\sup_{N\geqslant 1}\left|\mathbb{E}\left[\Phi_N\left(\mathbf{Z}^0, \mathbf{Z}^1, \dots, \mathbf{Z}^t\right)\right] - \mathbb{E}\left[\Phi_N\left(\mathbf{Z}^{\epsilon,0}, \mathbf{Z}^{\epsilon,1}, \dots, \mathbf{Z}^{\epsilon,t}\right)\right]\right| \xrightarrow[\epsilon\to 0]{} 0$$

  using Lemma 12 and 13.

- Lemma 14 ensures the AMP orbit $\{\mathbf{X}^{\epsilon_{\mathbf{Y}},t}, \mathbf{M}^{\epsilon_{\mathbf{Y}},t} | f^t_{\epsilon_{\mathbf{Y}}}, \mathbf{X}^0\}$ uniformly approximates the $\{\mathbf{X}^t, \mathbf{M}^t | f^t, \mathbf{X}^0\}$ one.

In light of these results, consider the following decomposition: for any $\eta \geqslant 0$:

$$\begin{aligned}
&\mathbb{P}\left(\left|\Phi_N\left(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t\right) - \mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^1, ..., \mathbf{Z}^t\right)\right]\right| \geqslant \eta\right) \\
&\leqslant \mathbb{P}\left(\left|\Phi_N\left(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t\right) - \Phi_N\left(\mathbf{X}^0, \mathbf{X}^{\epsilon_{\mathbf{Y}},1}, ..., \mathbf{X}^{\epsilon_{\mathbf{Y}},t}\right)\right| \geqslant \frac{\eta}{3}\right) \\
&\quad + \mathbb{P}\left(\left|\Phi_N\left(\mathbf{X}^0, \mathbf{X}^{\epsilon_{\mathbf{Y}},1}, ..., \mathbf{X}^{\epsilon_{\mathbf{Y}},t}\right) - \mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^{\epsilon,1}, ..., \mathbf{Z}^{\epsilon,t}\right)\right]\right| \geqslant \frac{\eta}{3}\right) \\
&\quad + \mathbb{P}\left(\left|\mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^{\epsilon,1}, ..., \mathbf{Z}^{\epsilon,t}\right)\right] - \mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^1, ..., \mathbf{Z}^t\right)\right]\right| \geqslant \frac{\eta}{3}\right)
\end{aligned}$$

Starting with the first term of the r.h.s., the pseudo-Lipschitz property and the triangle inequality give

$$\begin{aligned}
&\left|\Phi_N(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t) - \Phi_N(\mathbf{X}^0, \mathbf{X}^{\epsilon_{\mathbf{Y}},1}, ..., \mathbf{X}^{\epsilon_{\mathbf{Y}},t})\right| \leqslant \\
&L\left(1 + 2\frac{\|\mathbf{X}^0\|_F^{k-1}}{n^{k-1}} + \sum_{i=1}^t \frac{\|\mathbf{X}^i\|_F^{k-1}}{n^{(k-1)/2}} + \sum_{i=1}^t \frac{\|\mathbf{X}^{\epsilon,i}\|_F^{k-1}}{N^{(k-1)/2}}\right) \sum_{i=1}^t \frac{\|\mathbf{X}^{\epsilon,i} - \mathbf{X}^i\|_F}{\sqrt{N}} \\
&\leqslant L\left(1 + 2\frac{\|\mathbf{X}^0\|_F^{k-1}}{n^{(k-1)/2}} + \sum_{i=1}^t \frac{\|\mathbf{X}^i - \mathbf{X}^{\epsilon,i} + \mathbf{X}^{\epsilon,i}\|_F^{k-1}}{n^{(k-1)/2}} + \sum_{i=1}^t \frac{\|\mathbf{X}^{\epsilon,i}\|_F^{k-1}}{n^{(k-1)/2}}\right) \sum_{i=1}^t \frac{\|\mathbf{X}^{\epsilon,i} - \mathbf{X}^i\|_F}{\sqrt{N}} \\
&\leqslant L\left(1 + 2\frac{\|\mathbf{X}^0\|_F^{k-1}}{n^{(k-1)/2}} + \sum_{i=1}^t \frac{\|\mathbf{X}^i - \mathbf{X}^{\epsilon,i}\|_F^{k-1}}{n^{(k-1)/2}} + 2\sum_{i=1}^t \frac{\|\mathbf{X}^{\epsilon,i}\|_F^{k-1}}{n^{(k-1)/2}}\right) \sum_{i=1}^t \frac{\|\mathbf{X}^{\epsilon,i} - \mathbf{X}^i\|_F}{\sqrt{N}} \\
&\leqslant L\left(1 + 2C_0^{k-1} + \sum_{i=1}^t h_i(\epsilon)^{k-1} + 2\sum_{i=1}^t C^{k-1}_{\epsilon_{\mathbf{Y}},t}\right) \sum_{i=1}^t h_i(\epsilon) \quad \text{w.h.p.}
\end{aligned}$$

where we used assumption (B3) for the convergence of $\|\mathbf{X}_0\|_F/\sqrt{N}$ to a finite constant, the well-defined state evolution of the perturbed orbit $\{\mathbf{X}^{\epsilon\mathbf{Y},t}, \mathbf{M}^{\epsilon\mathbf{Y},t} | f^t_{\epsilon\mathbf{Y}}, \mathbf{X}^0\}$ for convergence of $\|\mathbf{X}^{\epsilon,i}\|/\sqrt{N}$ to finite constants $C_{\epsilon\mathbf{Y},t}$ and Lemma 14 to replace the differences $\|\mathbf{X}^{\epsilon,i} - \mathbf{X}^i\|_F$ by the functions $h_i(\epsilon)$ with high probability. This gives, for any $\eta > 0$:

$$\lim_{\epsilon\to 0} \limsup_{N\to\infty} \mathbb{P}\left(\left|\Phi_N\left(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t\right) - \Phi_N\left(\mathbf{X}^0, \mathbf{X}^{\epsilon\mathbf{Y},1}, ..., \mathbf{X}^{\epsilon\mathbf{Y},t}\right)\right| \geqslant \frac{\eta}{3}\right) = 0 \tag{3.39}$$

The state evolution for the perturbed AMP then gives

$$\lim_{\epsilon\to 0} \limsup_{N\to\infty} \mathbb{P}\left(\left|\Phi_N\left(\mathbf{X}^0, \mathbf{X}^{\epsilon\mathbf{Y},1}, ..., \mathbf{X}^{\epsilon\mathbf{Y},t}\right) - \mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^{\epsilon,1}, ..., \mathbf{Z}^{\epsilon,t}\right)\right]\right| \geqslant \frac{\eta}{3}\right) = 0 \tag{3.40}$$

and Lemma 12 guarantees:

$$\lim_{\epsilon\to 0} \mathbb{P}\left(\left|\mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^{\epsilon\mathbf{Y},1}, ..., \mathbf{Z}^{\epsilon\mathbf{Y},t}\right)\right] - \mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^1, ..., \mathbf{Z}^t\right)\right]\right| \geqslant \frac{\eta}{3}\right) = 0 \tag{3.41}$$

for all N. From this we deduce

$$\mathbb{P}\left(\left|\Phi_N\left(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t\right) - \mathbb{E}\left[\Phi_N\left(\mathbf{X}^0, \mathbf{Z}^1, ..., \mathbf{Z}^t\right)\right]\right| \geqslant \eta\right) \xrightarrow[N\to\infty]{} 0 \tag{3.42}$$

which is the desired result. □

We now turn to the proof of Theorem 6.

Proof of Theorem 6. The property is verified at $t = 0$ straightforwardly from the initial conditions : $\hat{\mathbf{X}}^0 = \mathbf{X}^0$ and $\hat{\mathbf{M}}^0 = \mathbf{M}^0 = f_0(\mathbf{X}^0)$.
Consider now that Corollary 6 is verified up to time $t-1$. Then, using the pseudo-Lipschitz property:

$$\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^t - \mathbf{M}^t\right\|_F = \frac{1}{\sqrt{N}}\left\|f_t(\hat{\mathbf{X}}^t) - f_t(\mathbf{X}^t)\right\|_F \leqslant \left(1 + \frac{\left\|\hat{\mathbf{X}}^t\right\|_F^{k-1}}{n^{(k-1)/2}} + \frac{\|\mathbf{X}^t\|_F^{k-1}}{n^{(k-1)/2}}\right)\frac{\left\|\hat{\mathbf{X}}^t - \mathbf{X}^t\right\|_F}{\sqrt{N}} \tag{3.43}$$

$\mathbf{X}^t$ verifies a well-defined state evolution using Theorem 5, thus $\lim_{N\to\infty} \frac{\|\mathbf{X}^t\|_F^{k-1}}{n^{(k-1)/2}} \leqslant C_t$ for a given bounded constants C_t. To bound $\left\|\hat{\mathbf{X}}^t\right\|_F/\sqrt{N}$, we can write:

$$\frac{\left\|\hat{\mathbf{X}}^t\right\|_F}{\sqrt{N}} = \frac{\left\|\hat{\mathbf{X}}^t + \mathbf{X}^t - \mathbf{X}^t\right\|_F}{\sqrt{N}} \leqslant \frac{\|\mathbf{X}^t\|_F}{\sqrt{N}} + \frac{\left\|\hat{\mathbf{X}}^t - \mathbf{X}^t\right\|_F}{\sqrt{N}} \tag{3.44}$$

where the large n limit of the first term of the r.h.s. is bounded and the second term vanishes from the induction hypothesis, which gives $\left\|\hat{\mathbf{X}}^t - \mathbf{X}^t\right\|_F/\sqrt{N} \xrightarrow[n\to+\infty]{} 0$. Combining these steps, we get $\left\|\hat{\mathbf{M}}^t - \mathbf{M}^t\right\|_F/\sqrt{N} \xrightarrow[n\to+\infty]{} 0$. Moving to $\hat{\mathbf{X}}^{t+1}$, we write :

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{X}}^{t+1} - \mathbf{X}^{t+1}\right\|_F \leqslant \|\mathbf{A}\|_{op}\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^t - \mathbf{M}^t\right\|_F + \frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^{t-1}\hat{b}_t^\top - \mathbf{M}^{t-1}b_t^\top\right\|_F \\
&\leqslant \|\mathbf{A}\|_{op}\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^t - \mathbf{M}^t\right\|_F + \frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^{t-1}\hat{b}_t^\top - \mathbf{M}^{t-1}\hat{b}_t^\top + \mathbf{M}^{t-1}\hat{b}_t^\top - \mathbf{M}^{t-1}b_t^\top\right\|_F \\
&\leqslant \|\mathbf{A}\|_{op}\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^t - \mathbf{M}^t\right\|_F + \frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^{t-1}\hat{b}_t^\top - \mathbf{M}^{t-1}\hat{b}_t^\top\right\|_F + \frac{1}{\sqrt{N}}\left\|\mathbf{M}^{t-1}\hat{b}_t^\top - \mathbf{M}^{t-1}b_t^\top\right\|_F \\
&\leqslant \|\mathbf{A}\|_{op}\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^t - \mathbf{M}^t\right\|_F + \frac{1}{\sqrt{N}}\left\|\hat{\mathbf{M}}^{t-1} - \mathbf{M}^{t-1}\right\|_F\|b_t\|_F + \frac{1}{\sqrt{N}}\left\|\hat{b}_t^\top - b_t^\top\right\|_F\left\|\mathbf{M}^{t-1}\right\|_F
\end{aligned} \tag{3.45}$$

and handle each quantity using similar arguments as before: the quantities $\|\mathbf{M}^{t-1}\|_F/\sqrt{N}$ and $\|b_t\|_F$ are bounded for large n using the state evolution from Theorem 5, the quantities $\left\|\hat{\mathbf{M}}^t - \mathbf{M}^t\right\|_F/\sqrt{N}$ and $\left\|\hat{\mathbf{M}}^{t-1} - \mathbf{M}^{t-1}\right\|_F/\sqrt{N}$ vanish for large n using the first part of this proof and the induction hypothesis. The operator norm of $\mathbf{A}$ may be bounded using Proposition (5). This proves the induction and concludes the proof of Theorem 6. □

3.3.3 Proof of intermediate lemmas

Those proofs which are too close to the ones appearing in [37] are not reminded.

Proof of Lemma 6. Recall the σ-algebra $\mathfrak{S}_t = \sigma(\mathbf{H}^1, \mathbf{H}^2, ..., \mathbf{H}^t)$. The LongAMP iteration verifies:

$$\mathbf{H}^{t+1} = (\mathbf{Id} - \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}})\mathbf{A}\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \tag{3.46}$$

$$= \mathbf{A}\mathbf{Q}^t_{\perp} - \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{Q}^t_{\perp} + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \tag{3.47}$$

where $\mathbf{Q}^t_{\perp} = \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^t$. We now show by an induction that conditioning on $\mathfrak{S}_t$ is equivalent to conditioning on the linear observations $\mathbf{A}\mathbf{Q}^0, \mathbf{A}\mathbf{Q}^1, ..., \mathbf{A}\mathbf{Q}^t$, and thus to conditioning on $\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1}$. Consider the first iteration which initializes the induction:

$$\mathbf{H}^1 = \mathbf{A}\mathbf{Q}^0 \tag{3.48}$$

thus $\mathbf{H}^1$ is $\sigma(\mathbf{A}\mathbf{Q}^0)$-measurable. Suppose now that $\boldsymbol{\mathcal{H}}_{t-1}$ is $\sigma(\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1})$-measurable. The LongAMP iteration then gives, remembering that $\mathbf{Q}^t_{\parallel} = P_{\mathbf{Q}_{t-1}}\mathbf{Q}^t$:

$$\mathbf{H}^{t+1} = \mathbf{A}\mathbf{Q}^t - \underbrace{\mathbf{A}\mathbf{Q}^t_{\parallel} - \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{Q}^t_{\perp} + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t}_{\sigma(\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1})-\text{measurable}} \tag{3.49}$$

where the highlighted term is $\sigma(\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1})-$measurable by definition of $\mathbf{Q}^t_{\parallel}$ and the induction hypothesis. This gives that $\boldsymbol{\mathcal{H}}_t$ is $\sigma(\mathbf{A}\mathbf{Q}_t)$-measurable. We can now condition on the linear observation $\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1}$ at each iteration. We thus have:

$$\mathbf{H}^{t+1}|_{\mathfrak{S}_t} \overset{d}{=} \mathbf{A}|_{\mathfrak{S}_t}\mathbf{Q}^t_{\perp} - \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{Q}^t_{\perp} + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \tag{3.50}$$

which amounts to condition the Gaussian space generated by the entries of $\mathbf{A}$ on its subspace defined by the linear combinations $\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1}$. Conditioning in Gaussian spaces amounts to doing orthogonal projections, which gives

$$\mathbf{A}|_{\mathfrak{S}_t} = \mathbb{E}\left[\mathbf{A}|\mathfrak{S}_t\right] + \mathcal{P}_t(\tilde{\mathbf{A}}) \tag{3.51}$$

as shown in [28],[135], where $\tilde{\mathbf{A}}$ is a copy of $\mathbf{A}$, independent of $\mathfrak{S}_t$ and $\mathcal{P}_t$ is the projector onto the subspace $\{\hat{\mathbf{A}} \in \mathbb{R}^{N\times N} | \hat{\mathbf{A}}\boldsymbol{\mathcal{Q}}_{t-1} = 0, \hat{\mathbf{A}} = \hat{\mathbf{A}}^\top\}$:

$$\mathbb{E}\left[\mathbf{A}|\mathfrak{S}_t\right] = \mathbf{A} - \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}} \tag{3.52}$$

$$\mathcal{P}_t(\tilde{A}) = \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}} \tag{3.53}$$

where $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$. Replacing in the original LongAMP iteration, we get :

$$\mathbf{H}^{t+1}|_{\mathfrak{S}_t} \overset{d}{=} \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \tag{3.54}$$

where we used $\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbb{E}\left[\mathbf{A}|\mathfrak{S}_t\right]\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}} = 0$. □

Proof of Lemma 7. We proceed by induction over t. Let S_t be the property at time t.

Initialization.

a) We have $\mathbf{H}^1 = \mathbf{A}\mathbf{Q}^0$. Then:

$$\begin{aligned}\frac{1}{N}(\mathbf{H}^1)^\top \mathbf{H}^1 &= \frac{1}{N}(\mathbf{A}\mathbf{Q}^0)^\top(\mathbf{A}\mathbf{Q}^0) \\ &\overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{Q}^0)^\top \mathbf{Q}^0 \end{aligned} \tag{3.55}$$

using Lemma 21. We then define $\boldsymbol{\kappa}^{1,1} = \frac{1}{N}(\mathbf{Q}^0)^\top \mathbf{Q}^0$.

b) We want to show that $\Phi_N(\mathbf{X}^0, \mathbf{H}^1) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\Phi_N(\mathbf{X}^0, \mathbf{Z}^1)\right]$ where $\mathbf{Z}^1 \sim \mathbf{N}(0, \boldsymbol{\kappa}^{1,1})$, where

$$\boldsymbol{\kappa}^{1,1} = \frac{1}{N}(\mathbf{Q}^0)^\top \mathbf{Q}^0 = \frac{1}{N}\left(f^0(\mathbf{X}^0)\right)^\top f^0(\mathbf{X}^0) \tag{3.56}$$

For any sequence $\{\Phi_N\}_{N\in\mathbb{N}}$ of order k pseudo-Lipschitz function

$$\begin{aligned}&\left\|\Phi_N(\mathbf{X}_0, \mathbf{A}\mathbf{Q}^0) - \mathbb{E}[\Phi_N(\mathbf{Z}^1)]\right\|_2 \leqslant \left\|\Phi_N(\mathbf{A}\mathbf{Q}^0) - \Phi_N(\mathbf{Z}^1)\right\|_2 + \left\|\Phi_N(\mathbf{Z}^1) - \mathbb{E}[\Phi_N(\mathbf{Z}^1)]\right\|_2 \\ &\leqslant L_n\left(1 + \left(\frac{\|\mathbf{A}\mathbf{Q}^0\|_2}{\sqrt{N}}\right)^{k-1} + \left(\frac{\|\mathbf{Z}^1\|}{\sqrt{N}}\right)^{k-1}\right)\frac{\|\mathbf{A}\mathbf{Q}^0 - \mathbf{Z}^1\|_2}{\sqrt{N}} + \left\|\Phi_N(\mathbf{Z}^1) - \mathbb{E}[\Phi_N(\mathbf{Z}^1)]\right\|_2\end{aligned} \tag{3.57}$$

where the large n limit of $\left(\frac{\|\mathbf{A}\mathbf{Q}^0\|_2}{\sqrt{N}}\right)^{k-1} + \left(\frac{\|\mathbf{Z}^1\|}{\sqrt{N}}\right)^{k-1}$ being bounded, $\frac{\|\mathbf{A}\mathbf{Q}^0 - \mathbf{Z}^1\|_2}{\sqrt{N}} \xrightarrow[n\to\infty]{a.s} 0$ and $\left\|\Phi_N(\mathbf{Z}^1) - \mathbb{E}[\Phi_N(\mathbf{Z}^1)]\right\|_2 \xrightarrow[n\to\infty]{P} 0$ follow from Lemmas 1 and 21 .

Induction. Here we assume that $S_0, S_1, ..., S_{t-1}$ are verified, and we prove S_t.

a) Consider the case $s < t$. Since $\mathbf{H}^{s+1}$ and $\langle \mathbf{Q}^s, \mathbf{Q}^r\rangle$ are $\mathfrak{S}_t$ measurable, using the conditioning lemma, we have :

$$\begin{aligned}\left((\mathbf{H}^{s+1})^\top \mathbf{H}^{t+1} - (\mathbf{Q}^s)^\top \mathbf{Q}^t\right)|_{\mathfrak{S}_t} &\overset{d}{=} \left((\mathbf{H}^{s+1})^\top \mathbf{H}^{t+1}|_{\mathfrak{S}_t} - (\mathbf{Q}^s)^\top \mathbf{Q}^t\right) \\ &= (\mathbf{H}^{s+1})^\top(\mathbf{P}^\perp_{\mathcal{Q}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathcal{Q}_{t-1}}\mathbf{Q}^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t) - (\mathbf{Q}^s)^\top \mathbf{Q}^t \\ &= (\mathbf{H}^{s+1})^\top \mathbf{P}^\perp_{\mathcal{Q}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp + (\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t - (\mathbf{Q}^s)^\top \mathbf{Q}^t\end{aligned} \tag{3.58}$$

We thus have :

$$\begin{aligned}\frac{1}{N}\left\|\left((\mathbf{H}^{s+1})^\top \mathbf{H}^{t+1} - (\mathbf{Q}^s)^\top \mathbf{Q}^t\right)|_{\mathfrak{S}_t}\right\|_F \leqslant &\frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \mathbf{P}^\perp_{\mathcal{Q}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F \\ &+ \frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t - (\mathbf{Q}^s)^\top \mathbf{Q}^t\right\|_F\end{aligned} \tag{3.59}$$

Starting with the term

$$\frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \mathbf{P}^\perp_{\mathcal{Q}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F = \frac{1}{N}\left\|(\mathbf{P}^\perp_{\mathcal{Q}_{t-1}}\mathbf{H}^{s+1})^\top \tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F \tag{3.60}$$

the induction ensires that $\frac{1}{\sqrt{N}}\|\mathbf{H}^{s+1}\|_F$, $\frac{1}{\sqrt{N}}\|\mathbf{Q}^t_\perp\|_F$ concentrate to finite values. Furthermore, $\left\|\mathbf{P}^\perp_{\mathcal{Q}_{t-1}}\mathbf{H}^{s+1}\right\|_F \leqslant \|\mathbf{H}^{s+1}\|_F$, so according to Lemma 21, the first term on the right-hand-side

will concentrate to zero.
Moving to the second term, since $s < t$, $\mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^s = \mathbf{Q}^s$. Then:

$$\begin{aligned}\frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t - (\mathbf{Q}^s)^\top \mathbf{Q}^t\right\|_F &= \frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t - (\mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^s)^\top \mathbf{Q}^t\right\|_F \\ &= \frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t - (\mathbf{Q}^s)^\top \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{Q}}_{t-1}^\top \mathbf{Q}^t\right\|_F \\ &= \frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t - (\mathbf{Q}^s)^\top \boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}_t\right\|_F \\ &\leqslant \frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1} - (\mathbf{Q}^s)^\top \boldsymbol{\mathcal{Q}}_{t-1}\right\|_F \|\boldsymbol{\alpha}_t\|_F \end{aligned} \tag{3.61}$$

Here we consider $s < t$ thus $s+1 \leqslant t$. Hence the induction hypothesis includes the concentration properties of $\mathbf{H}^{s+1}$ and $\boldsymbol{\alpha}_t$. We then have $\lim_{N\to\infty}\frac{1}{N}\left\|(\mathbf{H}^{s+1})^\top \boldsymbol{\mathcal{H}}_{t-1} - (\mathbf{Q}^s)^\top \boldsymbol{\mathcal{Q}}_{t-1}\right\|_F \to 0$ and $\|\boldsymbol{\alpha}_t\|_F$ has a finite and well-defined limit using the non-degeneracy assumption. Indeed:

$$\begin{aligned}\|\boldsymbol{\alpha}_t\|_F &= \left\|(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{Q}}_{t-1}^\top \mathbf{Q}^t\right\|_F \\ &\leqslant \frac{1}{Nc_t^2}\boldsymbol{\mathcal{Q}}_{t-1}^\top \mathbf{Q}^t \end{aligned} \tag{3.62}$$

using the induction hypothesis, $\lim_{n\to+\infty}\frac{1}{N}\boldsymbol{\mathcal{Q}}_{t-1}^\top\mathbf{Q}^t$ is finite. This proves the property for $s < t$. Now consider the case $s = t$. We then have:

$$\begin{aligned}&\left(\left\|\mathbf{H}^{t+1}\right\|_F^2 - \left\|\mathbf{Q}^t\right\|_F^2\right)|_{\mathfrak{S}_t} = \left(\left\|\mathbf{H}^{t+1}|_{\mathfrak{S}_t}\right\|_F^2 - \left\|\mathbf{Q}^t\right\|_F^2\right) \\ &= \left\|\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F^2 + 2\mathrm{Tr}\left(\left(\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right)^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right) + \left\|\boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right\|_F^2 - \left\|\mathbf{Q}^t\right\|_F^2 \end{aligned} \tag{3.63}$$

We then have

$$\frac{1}{N}\left\|\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F^2 = \frac{1}{N}\left\|\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F^2 - \frac{1}{N}\left\|\mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F^2 \overset{\mathrm{P}}{\simeq} \frac{1}{N}\left\|\mathbf{Q}^t_\perp\right\|_F^2 \tag{3.64}$$

where we used

$$\frac{1}{N}\left\|\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F^2 \overset{\mathrm{P}}{\simeq} \frac{1}{N}\left\|\mathbf{Q}^t_\perp\right\|_F^2 \quad \text{and} \quad \frac{1}{N}\left\|\mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right\|_F^2 \xrightarrow[n\to\infty]{P} 0 \tag{3.65}$$

which follows from Lemma 21 and the independence of $\tilde{\mathbf{A}}$. The second term then reads

$$\left(\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right)^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t = (\mathbf{Q}^t_\perp)^\top\tilde{\mathbf{A}}\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \tag{3.66}$$

From the induction hypothesis, we know that $\boldsymbol{\alpha}^t$ has finite norm when $N \to \infty$. Moreover, $\left\|\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right\|_F \leqslant \left\|\boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right\|_F$, and $\left\|\mathbf{Q}^t_\perp\right\|_F \leqslant \left\|\mathbf{Q}^t\right\|_F$. Also $\frac{1}{\sqrt{N}}\|\boldsymbol{\mathcal{H}}_{t-1}\|_F$ and $\frac{1}{\sqrt{N}}\|\mathbf{Q}^t\|_F$ converge to finite constants, again according to the induction hypothesis. Using Lemma 21, we get

$$\frac{1}{N}\mathrm{Tr}\left(\left(\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}^t_\perp\right)^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right) \xrightarrow[n\to\infty]{P} 0 \tag{3.67}$$

Finally the third term can be decomposed

$$\begin{aligned}\left\|\boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right\|_F^2 &= \mathrm{Tr}(\left(\boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right)^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t)\\ &= \mathrm{Tr}((\boldsymbol{\alpha}^t)^\top \boldsymbol{\mathcal{H}}_{t-1}^\top \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t)\\ &= \mathrm{Tr}((\boldsymbol{\alpha}^t)^\top \left(\boldsymbol{\mathcal{H}}_{t-1}^\top \boldsymbol{\mathcal{H}}_{t-1} - \boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1}\right)\boldsymbol{\alpha}^t) + \mathrm{Tr}\Big((\boldsymbol{\alpha}^t)^\top \boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}^t\Big)\\ &\leqslant \left\|\boldsymbol{\mathcal{H}}_{t-1}^\top \boldsymbol{\mathcal{H}}_{t-1} - \boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1}\right\|_F \|\boldsymbol{\alpha}_t\|_F + \|\boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}_t\|_F \end{aligned} \tag{3.68}$$

Using the induction hypothesis and the non-degeneracy assumption, $\lim_{N\to\infty}\left\|\boldsymbol{\alpha}^t\right\|_F$ is a finite constant, and $\frac{1}{N}\left\|\boldsymbol{\mathcal{H}}_{t-1}^\top \boldsymbol{\mathcal{H}}_{t-1} - \boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1}\right\|_F \xrightarrow[n\to\infty]{P} 0$. Furthermore, by definition of $\boldsymbol{\alpha}_t$, $\boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}_t = \mathbf{Q}_t^{\|}$.

Grouping all the terms, we get

$$\begin{aligned}\frac{1}{N}\left(\left\|\mathbf{H}^{t+1}\right\|_F^2 - \left\|\mathbf{Q}^t\right\|_F^2\right)|_{\mathfrak{S}_t} &\overset{\mathrm{P}}{\simeq} \frac{1}{N}\left\|\mathbf{Q}_\perp^t\right\|_F^2 + \frac{1}{N}\left\|\mathbf{Q}_{\|}^t\right\|_F^2 - \frac{1}{N}\left\|\mathbf{Q}^t\right\|_F^2\\ &= 0\end{aligned} \tag{3.69}$$

b) Using the conditioning lemma :

$$\begin{aligned}\Phi_N\left(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^t, \mathbf{H}^{t+1}\right)|_{\mathfrak{S}_t} &\overset{d}{=} \Phi_N\left(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^t, \mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right)\\ &= \Phi_N\left(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^t, \tilde{\mathbf{A}}\mathbf{Q}_\perp^t - P_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right)\end{aligned} \tag{3.70}$$

Let $\Phi_N'\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t - P_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right) = \Phi_N\left(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^t, \tilde{\mathbf{A}}\mathbf{Q}_\perp^t - P_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right)$ as a shorthand. Then, from the pseudo-Lipschitz property:

$$\begin{aligned}&\left|\Phi_N'\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t - P_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right) - \Phi_N'\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right)\right|\\ &\quad\leqslant L_N C(k,t)\Bigg(1 + \left(\frac{\|\mathbf{X}^0\|_F}{\sqrt{N}}\right)^{k-1} + \sum_{s=1}^{t}\left(\frac{\|\mathbf{H}^s\|_F}{\sqrt{N}}\right)^{k-1}\\ &\quad + \left(\frac{\|\mathbf{H}^{t+1}\|_F}{\sqrt{N}}\right)^{k-1} + \left(\frac{\left\|\tilde{\mathbf{A}}\mathbf{Q}_\perp^t\right\|_F}{\sqrt{N}}\right)^{k-1} + \left(\frac{\|\boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\|_F}{\sqrt{N}}\right)^{k-1}\Bigg)\frac{\left\|P_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}_\perp^t\right\|_F}{\sqrt{N}}\end{aligned} \tag{3.71}$$

where $C(k,t)$ is a constant depending only on k and t. The induction hypothesis ensures that $\left(\frac{\|\mathbf{X}^0\|_F}{\sqrt{N}}\right)^{k-1} + \sum_{s=1}^{t}\left(\frac{\|\mathbf{H}^s\|_F}{\sqrt{N}}\right)^{k-1}$ converges to a finite constant. Furthermore,

$$\frac{1}{\sqrt{N}}\left\|\tilde{\mathbf{A}}\right\|_F \leqslant \frac{1}{\sqrt{N}}\left\|\tilde{\mathbf{A}}\right\|_{op}\left\|\mathbf{Q}^t\right\|_F \tag{3.72}$$

which, using Proposition 5 and the induction hypothesis, converges to a finite constant. Also, using the fact that $\mathrm{rank}(\mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}) \leqslant tq$ with t, q finite, and the independence of $\tilde{\mathbf{A}}$, Lemma 21 gives

$$\frac{1}{\sqrt{N}}\left\|P_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}_\perp^t\right\|_F \xrightarrow[n\to\infty]{P} 0\,. \tag{3.73}$$

Ultimately, we obtain

$$\Phi_N^{'}\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t - P_{\boldsymbol{\mathcal{Q}}_{t-1}}\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right) \overset{\mathrm{P}}{\simeq} \Phi_N^{'}\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t\right)$$
$$\overset{\mathrm{P}}{\simeq} \Phi_N^{'}\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^{t,*}\right) \tag{3.74}$$

where $\boldsymbol{\alpha}_t^* = \lim_{N\to\infty}\boldsymbol{\alpha}_t$ which are finite matrices, and $\boldsymbol{\alpha}_t^* \in \mathbb{R}^{tq\times q}$. We write :

$$\begin{bmatrix}(\boldsymbol{\alpha}_t^*)_1\\ ...\\ (\boldsymbol{\alpha}_t^*)_t\end{bmatrix} \tag{3.75}$$

where $\forall 1 \leqslant i \leqslant t$, $(\boldsymbol{\alpha}_t^*)_i \in \mathbb{R}^{q\times q}$. Then

$$\Phi_N^{'}\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^{t,*}\right) \overset{\mathrm{P}}{\simeq} \Phi_N^{'}\left(\tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^{t,*}\right)$$
$$\overset{\mathrm{P}}{\simeq} \Phi(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^t, \tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^{t,*}) \tag{3.76}$$

Using Lemma 1, there exists $\mathbf{Z}_\perp^{t+1} \sim \mathbf{N}(0, \boldsymbol{\kappa}_\perp^{t+1}\otimes I_N)$ independent of $\mathfrak{S}_t$, where $\boldsymbol{\kappa}_\perp^{t+1} = \lim_{N\to\infty}\frac{1}{N}(\mathbf{Q}_\perp^t)^\top\mathbf{Q}_\perp^t$, such that:

$$\Phi(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^t, \tilde{\mathbf{A}}\mathbf{Q}_\perp^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^{t,*}) \overset{\mathrm{P}}{\simeq} \mathbb{E}_{\mathbf{Z}}\left[\Phi(\mathbf{X}^0, \mathbf{H}^1, ..., \mathbf{H}^t, \mathbf{Z}_\perp^{t+1} + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^{t,*})\right]$$
$$\overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\Phi_N(\mathbf{X}^0, \mathbf{Z}^1, ..., \mathbf{Z}^t, \mathbf{Z}_\perp^{t+1} + \sum_{i=1}^t \mathbf{Z}^i(\boldsymbol{\alpha}^{t,*})_i)\right] \tag{3.77}$$

We now need to match the covariance matrices defined by the prescription of $\mathbf{Z}^{t+1}$ we obtained with the ones from the state evolution. Let $\mathbf{Z}^{t+1} = \mathbf{Z}_\perp^{t+1} + \sum_{i=1}^t \mathbf{Z}^i(\boldsymbol{\alpha}^{t,*})_i) \in \mathbb{R}^{q\times q}$. We then write $\mathbf{Z}^{t+1} \sim \mathbf{N}(0, \boldsymbol{\kappa}^{t+1,t+1}\otimes\mathbf{I}_N)$ where $\boldsymbol{\kappa}^{t+1,t+1} = \lim_{N\to\infty}\frac{1}{N}(\mathbf{Z}^{t+1})^\top\mathbf{Z}^{t+1}$. Then, using the isometry proved above and remembering that, for any $1 \leqslant i \leqslant t, \mathbf{Q}^t = f^t(\mathbf{H}^t)$:

$$\frac{1}{N}(\mathbf{Z}^{t+1})^\top\mathbf{Z}^{t+1} \overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{H}^{t+1})^\top\mathbf{H}^{t+1} \overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{Q}^t)^\top\mathbf{Q}^t \xrightarrow[n\to\infty]{P} \boldsymbol{\kappa}^{t+1,t+1} \tag{3.78}$$

similarly, for $s \geqslant 2$:

$$\boldsymbol{\kappa}^s = \frac{1}{N}(\mathbf{Z}^s)^\top\mathbf{Z}^{t+1} \overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{H}^s)^\top\mathbf{H}^{t+1} \overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{Q}^{s-1})^\top\mathbf{Q}^t \xrightarrow[n\to\infty]{P} \boldsymbol{\kappa}^{s,t+1} \tag{3.79}$$

and for $s = 1$:

$$\boldsymbol{\kappa}^s = \frac{1}{N}(\mathbf{Z}^1)^\top\mathbf{Z}^{t+1} \overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{H}^1)^\top\mathbf{H}^{t+1} \overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{Q}^0)^\top\mathbf{Q}^t \xrightarrow[n\to\infty]{P} \boldsymbol{\kappa}^{1,t+1} \tag{3.80}$$

□

Proof of Lemma 8. This lemma is proven by induction.

Initialization. The first iterates read $\mathbf{H}^1 = \mathbf{A}\mathbf{Q}^0$ and $\hat{\mathbf{H}}^1 = \mathbf{A}\mathbf{Q}^0$. This concludes the initialization.

Induction. Assume the proposition is true up to time t. Define the $(t+1)q\times(t+1)q$ block-diagonal matrix $\boldsymbol{\mathcal{B}}_t = \text{diag}\left(0_{q\times q}, \mathbf{b}^1, ..., \mathbf{b}^t\right)$ and $\hat{\boldsymbol{\mathcal{H}}}_{t-1} = \left[\hat{\mathbf{H}}^1|\hat{\mathbf{H}}^2|...|\hat{\mathbf{H}}^t\right]$. We then have :

$$\begin{aligned}\mathbf{H}^{t+1} &= \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^t + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \\ &= \mathbf{A}\mathbf{Q}^t_{\perp} - \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{Q}^t_{\perp} + \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \end{aligned} \tag{3.81}$$

and

$$\begin{aligned}\hat{\mathbf{H}}^{t+1} &= \mathbf{A}\mathbf{Q}^t - \mathbf{Q}^{t-1}(\mathbf{b}^t)^\top \\ &= \mathbf{A}\mathbf{Q}^t_{\perp} + \mathbf{A}\mathbf{Q}^t_{\parallel} - \mathbf{Q}^{t-1}(\mathbf{b}^t)^\top \\ \text{where}\quad \mathbf{A}\mathbf{Q}^t_{\parallel} &= \mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{Q}}_{t-1}^\top\mathbf{Q}^t \\ &= \mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}^t \end{aligned} \tag{3.82}$$

which gives

$$\hat{\mathbf{H}}^{t+1} - \mathbf{H}^{t+1} = \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{Q}^t_{\perp} - \mathbf{Q}^{t-1}(\mathbf{b}^t)^\top + \mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}^t - \boldsymbol{\mathcal{H}}_{t-1}\boldsymbol{\alpha}^t \tag{3.83}$$

using the definition of iteration (3.29), we have:

$$\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1} = \hat{\boldsymbol{\mathcal{H}}}_{t-1} + \left[0_{N\times q}|\mathbf{Q}^0|...|\mathbf{Q}^{t-2}\right]\boldsymbol{\mathcal{B}}_{t-1}^\top \tag{3.84}$$

$$\begin{aligned}\hat{\mathbf{H}}^{t+1} - \mathbf{H}^{t+1} &= \mathbf{P}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{A}\mathbf{Q}^t_{\perp} - \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top\boldsymbol{\alpha}^t + \left(\hat{\boldsymbol{\mathcal{H}}}^{t-1} - \boldsymbol{\mathcal{H}}^{t-1}\right)\boldsymbol{\alpha}^t \\ &= \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{Q}}_{t-1}^\top\mathbf{A}\mathbf{Q}^t_{\perp} - \mathbf{Q}^{t-1}(\mathbf{b}_{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top\boldsymbol{\alpha}^t \\ &\qquad + \left(\hat{\boldsymbol{\mathcal{H}}}^{t-1} - \boldsymbol{\mathcal{H}}^{t-1}\right)\boldsymbol{\alpha}^t \end{aligned} \tag{3.85}$$

and

$$\begin{aligned}\boldsymbol{\mathcal{Q}}_{t-1}^\top\mathbf{A} &= (\mathbf{A}\boldsymbol{\mathcal{Q}}_{t-1})^\top \\ &= ((\hat{\boldsymbol{\mathcal{H}}}_{t-1} + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_t^\top))^\top \\ &= \hat{\boldsymbol{\mathcal{H}}}_{t-1}^\top + \boldsymbol{\mathcal{B}}_t\,[0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]^\top \end{aligned} \tag{3.86}$$

since $\mathbf{Q}^t_{\perp} = \mathbf{P}^{\perp}_{\mathbf{Q}_{t-1}}\mathbf{Q}^t$, it holds that:

$$\begin{aligned}\boldsymbol{\mathcal{Q}}_{t-1}^\top\mathbf{A}\mathbf{Q}^t_{\perp} &= \left(\hat{\boldsymbol{\mathcal{H}}}_{t-1}^\top + \boldsymbol{\mathcal{B}}_t\,[0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]^\top\right)\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^t \\ &\overset{\text{P}}{\sim} \hat{\boldsymbol{\mathcal{H}}}_{t-1}^\top\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}_{t-1}}\mathbf{Q}^t \end{aligned} \tag{3.87}$$

which in turn gives:

$$\begin{aligned}\hat{\mathbf{H}}^{t+1} - \mathbf{H}^{t+1} &\overset{\text{P}}{\sim} \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\hat{\boldsymbol{\mathcal{H}}}_{t-1}^\top\mathbf{Q}^t_{\perp} - \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top\boldsymbol{\alpha}^t \\ &\quad + \left(\hat{\boldsymbol{\mathcal{H}}}^{t-1} - \boldsymbol{\mathcal{H}}^{t-1}\right)\boldsymbol{\alpha}^t \\ &= \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{H}}_{t-1}^\top\mathbf{Q}^t_{\perp} - \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top\boldsymbol{\alpha}^t \\ &\quad + \left(\hat{\boldsymbol{\mathcal{H}}}^{t-1} - \boldsymbol{\mathcal{H}}^{t-1}\right)\boldsymbol{\alpha}^t + \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\left(\hat{\boldsymbol{\mathcal{H}}}_{t-1} - \boldsymbol{\mathcal{H}}_{t-1}\right)^\top\mathbf{Q}^t_{\perp} \end{aligned} \tag{3.88}$$

We now study the limiting behaviour of this quantity, starting with:

$$\mathbf{C} = \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{H}}_{t-1}^\top \mathbf{Q}_\perp^t - \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top \boldsymbol{\alpha}^t \tag{3.89}$$

We have :

$$\begin{aligned}\mathbf{Q}_\perp^t &= \mathbf{Q}^t - \mathbf{Q}_\parallel^t \\ &= \mathbf{Q}^t - \boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}_t\end{aligned} \tag{3.90}$$

and :

$$\mathbf{C} = \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{H}}_{t-1}^\top (\mathbf{Q}^t - \boldsymbol{\mathcal{Q}}_{t-1}\boldsymbol{\alpha}_t) - \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top \boldsymbol{\alpha}^t \tag{3.91}$$

Using Lemma 17, the state evolution, and the concentration properties of pseudo-Lipschitz functions Lemma 1, we get, for all $1 \leqslant j \leqslant t-1$ and $1 \leqslant i \leqslant t$:

$$\begin{aligned}\frac{1}{N}(\mathbf{H}^i)^\top f^j(\mathbf{H}^j) &\overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\frac{1}{N}(\mathbf{Z}^i)^\top f^j(\mathbf{Z}^j)\right] \\ &= \mathbf{K}_{i,j}\mathbb{E}\left[\frac{1}{N}\mathrm{div} f^j(\mathbf{Z}^j)\right] \\ &\overset{\mathrm{P}}{\simeq} \frac{1}{N}(\mathbf{Q}^{i-1})^\top \mathbf{Q}^{j-1}(\mathbf{b}^j)^\top\end{aligned} \tag{3.92}$$

and for $j = 0$:

$$\begin{aligned}\frac{1}{N}(\mathbf{H}^i)^\top f(\mathbf{X}^0) &\overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\frac{1}{N}(\mathbf{Z}^i)^\top f_0(\mathbf{X}^0)\right] \\ &= 0\end{aligned} \tag{3.93}$$

which in turn gives

$$\frac{1}{N}(\boldsymbol{\mathcal{H}}_{t-1}^\top \mathbf{Q}^t) = \frac{1}{N}\left[\mathbf{H}^1|...|\mathbf{H}^t\right]^\top f_t(\mathbf{H}^t) \overset{\mathrm{P}}{\simeq} \frac{1}{N}(\boldsymbol{\mathcal{Q}}_{t-1})^\top \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top \tag{3.94}$$

and

$$\frac{1}{N}\boldsymbol{\mathcal{H}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1} = \frac{1}{N}\left[\mathbf{H}^1|...|\mathbf{H}^t\right]^\top \left[\mathbf{Q}^0|f_1(\mathbf{H}^1)...|f_{t-1}(\mathbf{H}^{t-1})\right] \overset{\mathrm{P}}{\simeq} \frac{1}{N}\boldsymbol{\mathcal{Q}}_{t-1}^\top [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\boldsymbol{\mathcal{B}}_{t-1}^\top \tag{3.95}$$

Furthermore, note that

$$\boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{H}}_{t-1}^\top \mathbf{Q}_\perp^t = \boldsymbol{\mathcal{Q}}_{t-1}(\frac{1}{N}\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}\frac{1}{N}\boldsymbol{\mathcal{H}}_{t-1}^\top \mathbf{Q}_\perp^t \tag{3.96}$$

where the limit $\lim_{N\to\infty}(\frac{1}{N}\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}$ is well-defined owing to the non-degeneracy assumption. We can then write :

$$\begin{aligned}\mathbf{C} &\overset{\mathrm{P}}{\simeq} \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{Q}}_{t-1}^\top (\mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top - [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top \boldsymbol{\alpha}^t) && (3.97)\\ &\qquad - \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top \boldsymbol{\alpha}^t \\ &= \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top \boldsymbol{\mathcal{Q}}_{t-1})^{-1}\boldsymbol{\mathcal{Q}}_{t-1}^\top \underbrace{(\mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top - [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top \boldsymbol{\alpha}^t)}_{\in \mathrm{span}(\boldsymbol{\mathcal{Q}}_{t-1})} && (3.98)\\ &\qquad - \mathbf{Q}^{t-1}(\mathbf{b}^{t-1})^\top + [0_{N\times q}|\boldsymbol{\mathcal{Q}}_{t-2}]\,\boldsymbol{\mathcal{B}}_{t-1}^\top \boldsymbol{\alpha}^t \\ &= 0 && (3.99)\end{aligned}$$

At this point, we have :

$$\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{H}}^{t+1}-\mathbf{H}^{t+1}\right\|_F \leqslant \frac{1}{\sqrt{N}}\|\mathbf{C}\|_F + \frac{1}{\sqrt{N}}\|\left(\hat{\boldsymbol{\mathcal{H}}}_{t-1}-\boldsymbol{\mathcal{H}}_{t-1}\right)\boldsymbol{\alpha}^t + \boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\left(\hat{\boldsymbol{\mathcal{H}}}_{t-1}-\boldsymbol{\mathcal{H}}_{t-1}\right)^\top\mathbf{Q}_\perp^t\|_F \tag{3.100}$$

Where

$$\frac{1}{\sqrt{N}}\left\|\left(\hat{\boldsymbol{\mathcal{H}}}_{t-1}-\boldsymbol{\mathcal{H}}_{t-1}\right)\boldsymbol{\alpha}^t\right\|_F \leqslant \frac{1}{\sqrt{N}}\left\|\hat{\boldsymbol{\mathcal{H}}}_{t-1}-\boldsymbol{\mathcal{H}}_{t-1}\right\|_F\left\|\boldsymbol{\alpha}^t\right\|_F \tag{3.101}$$

As previously discussed, $\|\boldsymbol{\alpha}^t\|_F$ has a finite limit, and according to the induction hypothesis, $\frac{1}{\sqrt{N}}\left\|\hat{\boldsymbol{\mathcal{H}}}_{t-1}-\boldsymbol{\mathcal{H}}_{t-1}\right\|_F \xrightarrow[N\to\infty]{P} 0$. Then

$$\frac{1}{\sqrt{N}}\left\|\boldsymbol{\mathcal{Q}}_{t-1}(\boldsymbol{\mathcal{Q}}_{t-1}^\top\boldsymbol{\mathcal{Q}}_{t-1})^{-1}\left(\hat{\boldsymbol{\mathcal{H}}}_{t-1}-\boldsymbol{\mathcal{H}}_{t-1}\right)^\top\mathbf{Q}_\perp^t\right\|_F \leqslant \frac{1}{\sqrt{N}}\left\|\hat{\boldsymbol{\mathcal{H}}}_{t-1}-\boldsymbol{\mathcal{H}}_{t-1}\right\|_F\frac{1}{Nc_t^2}\|\boldsymbol{\mathcal{Q}}_{t-1}\|_F\left\|\mathbf{Q}^t\right\|_F \tag{3.102}$$

where $\frac{1}{Nc_t^2}\|\boldsymbol{\mathcal{Q}}_{t-1}\|_F\|\mathbf{Q}^t\|_F$ converges to a finite limit due to the state evolution proved above. This ultimately shows that

$$\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{H}}^{t+1}-\mathbf{H}^{t+1}\right\|_F \xrightarrow[N\to\infty]{P} 0 \tag{3.103}$$

and concludes the induction. □

Proof of Lemma 9. This one is another induction. Let S_t be the statement $\frac{1}{\sqrt{N}}\|\mathbf{Q}^t-\mathbf{M}^t\|_F \xrightarrow[N\to\infty]{P} 0$ and $\frac{1}{\sqrt{N}}\|\mathbf{H}^{t+1}-\mathbf{X}^{t+1}\|_F \xrightarrow[N\to\infty]{P} 0$.

Initialization. We have $\mathbf{Q}^0 = f^0(\mathbf{X}^0) = \mathbf{M}^0$ and $\mathbf{H}^1 = \mathbf{A}\mathbf{Q}^0, \mathbf{X}^1 = \mathbf{A}\mathbf{M}^0$.

Induction We assume S_{t-1} is true, and we prove S_t. We have

$$\frac{1}{\sqrt{N}}\left\|\mathbf{Q}^t-\mathbf{M}^t\right\|_F = \frac{1}{\sqrt{N}}\left\|f^t(\mathbf{H}^t)-f^t(\mathbf{X}^t)\right\|_F \leqslant L_t\left(1+\left(\frac{\|\mathbf{H}^t\|_F}{\sqrt{N}}\right)^{k-1}+\left(\frac{\|\mathbf{X}^t\|_F}{\sqrt{N}}\right)^{k-1}\right)\frac{\|\mathbf{H}^t-\mathbf{X}^t\|_F}{\sqrt{N}} \tag{3.104}$$

which goes to zero as n goes to infinity from the induction hypothesis. We then prove that $\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{H}}^{t+1}-\mathbf{X}^{t+1}\right\|_F \xrightarrow[N\to\infty]{P} 0$.

$$\hat{\mathbf{H}}^{t+1}-\mathbf{X}^{t+1} = \mathbf{A}\mathbf{Q}^t-\mathbf{Q}^{t-1}(\mathbf{b}^t)^\top-\mathbf{A}\mathbf{M}^t+\mathbf{M}^{t-1}(\mathbf{b}^t)^\top \tag{3.105}$$

and

$$\frac{1}{\sqrt{N}}\left\|\hat{\mathbf{H}}^{t+1}-\mathbf{X}^{t+1}\right\|_F \leqslant \|\mathbf{A}\|_{op}\frac{1}{\sqrt{N}}\left\|\mathbf{Q}^t-\mathbf{M}^t\right\|_F+\frac{1}{\sqrt{N}}\left\|\mathbf{Q}^{t-1}-\mathbf{M}^{t-1}\right\|_F\left\|\mathbf{b}^t\right\|_F \tag{3.106}$$

using Proposition 5, $\|\mathbf{A}\|_{op} \xrightarrow[N\to\infty]{P} 2$. Using the induction hypothesis, $\frac{1}{\sqrt{N}}\|\mathbf{Q}^t-\mathbf{M}^t\|_F \xrightarrow[N\to\infty]{P} 0$, $\frac{1}{\sqrt{N}}\|\mathbf{Q}^{t-1}-\mathbf{M}^{t-1}\|_F \xrightarrow[N\to\infty]{P} 0$, and $\|\mathbf{b}^t\|_F$ is finite. This concludes the induction step. □

Proof of Lemma 10. In this proof, we will consider the $2q \times 2q$ covariance matrix $\boldsymbol{\kappa} = \begin{bmatrix} \boldsymbol{\kappa}^{1,1} & \boldsymbol{\kappa}^{1,2} \\ \boldsymbol{\kappa}^{1,2} & \boldsymbol{\kappa}^{2,2} \end{bmatrix}$ and two matrices $\mathbf{Z}^1, \mathbf{Z}^2 \in (\mathbb{R}^{N\times q})^2$ following the distribution $\mathbf{N}(0, \boldsymbol{\kappa} \otimes \mathbf{I}_N)$, and we study the corresponding state evolution when the perturbed functions $f^t_{\epsilon\mathbf{Y}}$ are considered. We drop the ϵ exponent on the covariance matrices since we are just studying the well-definiteness of the perturbed SE as an induction. The link with the original SE will be studied in subsequent lemmas.

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}\left[\frac{1}{N}(f^s_{\epsilon\mathbf{Y}}(\mathbf{Z}^s)^\top f^t_{\epsilon\mathbf{Y}}(\mathbf{Z}^t)\right] &= \mathbb{E}_{\mathbf{Z}}\left[\frac{1}{N}(f^s(\mathbf{Z}^s)^\top f^t(\mathbf{Z}^t)\right] + \epsilon\mathbb{E}_{\mathbf{Z}}\left[\frac{1}{N}(f^s(\mathbf{Z}^s))^\top \mathbf{Y}^t\right] \\
&+ \epsilon\mathbb{E}_{\mathbf{Z}}\left[\frac{1}{N}(f^t(\mathbf{Z}^t))^\top \mathbf{Y}^s\right] + \epsilon^2\frac{1}{N}(\mathbf{Y}^s)^\top\mathbf{Y}^t \\
&= \mathbb{E}_{\mathbf{Z}}\left[\frac{1}{N}(f^s(\mathbf{Z}^s)^\top f^t(\mathbf{Z}^t)\right] + \frac{\epsilon}{N}\mathbb{E}_{\mathbf{Z}}\left[f^s(\mathbf{Z}^s)\right]^\top \mathbf{Y}^t \\
&+ \frac{\epsilon}{N}\mathbb{E}_{\mathbf{Z}}\left[f^t(\mathbf{Z}^t)\right]^\top \mathbf{Y}^s + \frac{\epsilon^2}{N}(\mathbf{Y}^s)^\top\mathbf{Y}^t
\end{aligned}$$

- the first term does not depend on the perturbation and is deterministic. Using assumptions (A6), this quantity has a finite limit.
- second term is a $q \times q$ matrix where each element have zero mean and variance

$$\mathrm{Var}\left[\frac{1}{N}\left(\mathbb{E}\left[f^s(\mathbf{Z}^s)^\top\mathbf{Y}^t\right]^i_j\right)\right] = \frac{1}{N^2}\|\mathbb{E}\left[f^s(\mathbf{Z}^s)\right]\|_2^2 \leqslant \frac{C}{N} \tag{3.107}$$

  Using the Gaussian tail and the Borel-Cantelli lemma, this term converges almost surely to zero.
- the third term is treated in the same way as the second one
- the last term follows from the strong law of large numbers:

$$\lim_{N\to\infty}\frac{1}{N}(\mathbf{Y}^s)^\top\mathbf{Y}^t \xrightarrow[n\to\infty]{a.s.} \mathbf{I}_{q\times q}\delta_{s=t} \tag{3.108}$$

Putting things together, we get, almost surely:

$$\lim_{N\to\infty}\mathbb{E}_{\mathbf{Z}}\left[\frac{1}{N}(f^s_{\epsilon\mathbf{Y}}(\mathbf{Z}^s))^\top f^t_{\epsilon\mathbf{Y}}(\mathbf{Z}^t)\right] = \lim_{N\to\infty}\mathbb{E}_{\mathbf{Z}}\left[\frac{1}{N}(f^s(\mathbf{Z}^s))^\top f^t(\mathbf{Z}^t)\right] + \epsilon^2\mathbf{I}_{q\times q}\delta_{s=t} \tag{3.109}$$

Verifying the initialization assumptions (A4-A5) is very similar to the previous steps, thus we directly give the result. The initialization reads:

$$\lim_{N\to\infty}\frac{1}{N}(f^0_{\epsilon\mathbf{Y}}(\mathbf{X}^0))^\top f^0_{\epsilon\mathbf{Y}}(\mathbf{X}^0) = \lim_{N\to\infty}\frac{1}{N}(f^0(\mathbf{X}^0))^\top f^0(\mathbf{X}^0) + \epsilon^2\mathbf{I}_{q\times q} \tag{3.110}$$

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\left[(f^0_{\epsilon\mathbf{Y}}(\mathbf{X}^0))^\top f^t_{\epsilon\mathbf{Y}}(\mathbf{Z}^t)\right] = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}\left[(f^0(\mathbf{X}^0))^\top f^t(\mathbf{Z}^t)\right] \tag{3.111}$$

It follows straightforwardly from these equations and a short induction that the resulting state evolution is almost surely non-random. □

Proof of Lemma 11. By definition, for any $t \in \mathbb{N}$:

$$\mathbf{Q}^{t,\epsilon\mathbf{Y}} = \mathbf{Q}^t + \epsilon \mathbf{Y}^t \tag{3.112}$$

Then

$$\mathbf{Q}_{\perp}^{\epsilon\mathbf{Y},t} = \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}^{\epsilon\mathbf{Y}}_{t-1}} f^t(\mathbf{H}^{\epsilon\mathbf{Y},t}) + \epsilon \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}^{\epsilon\mathbf{Y}}_{t-1}} \mathbf{Y}^t \tag{3.113}$$

with the parallel term a linear combination of the previous ones. Denote $\mathcal{F}_t$ the σ-algebra generated by $\mathbf{H}^{\epsilon\mathbf{Y},1}, ..., \mathbf{H}^{\epsilon\mathbf{Y},t}, \mathbf{Y}^1, ..., \mathbf{Y}^{t-1}$. Since $\mathbf{Y}^t$ is generated independently of $\mathcal{F}_t$, each column j of $\mathbf{Q}^{\epsilon\mathbf{Y},t}$ obeys the distribution:

$$(\mathbf{Q}_{\perp}^{\epsilon\mathbf{Y},t})_j|_{\mathcal{F}_t} \sim \mathbf{N}(\mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}^{\epsilon\mathbf{Y}}_{t-1}} (f^t(\mathbf{H}^{\epsilon\mathbf{Y},t}))_j, \epsilon^2 \mathbf{P}^{\perp}_{\boldsymbol{\mathcal{Q}}^{\epsilon\mathbf{Y}}_{t-1}}) \tag{3.114}$$

the variance of which is almost surely non-zero whenever $N \geqslant tq$. Thus, when $N \geqslant tq$, the matrix $\boldsymbol{\mathcal{Q}}_{t-1}$ has full column rank. We now need to control the minimal singular value of $\boldsymbol{\mathcal{Q}}_{t-1}$. Following [28], Lemma 9, we only need to check that, for any column j, almost surely, for N sufficiently large, there exists a constant $c_\epsilon > 0$ such that:

$$\frac{1}{N} \left\| (\mathbf{Q}_{\perp}^{\epsilon\mathbf{Y},t})_j \right\|^2 \geqslant c_\epsilon \tag{3.115}$$

which follows in almost identical fashion to [37], Lemma 9 using the moments of a $N - tq$ chi-square variable, instead of $N - t$ in the original proof, which extends straightforwardly since q is kept finite. □

Proof of Lemma 12. This result is proven for $q = 1$ in [37] and the proof for the case of finite, integer q is identical. □

Proof of Lemma 13. This lemma is proven by induction.

Initialization. From equation (3.110), it holds that

$$\mathbf{K}^{\epsilon}_{1,1} = \mathbf{K}_{1,1} + \epsilon^2 \xrightarrow[\epsilon \to 0]{} \mathbf{K}_{1,1} \tag{3.116}$$

Induction. Let t be a non-negative integer. Assume that, for any $r, s \leqslant t$, $\boldsymbol{\kappa}_\epsilon^{r,s} \to \boldsymbol{\kappa}^{r,s}$. Then:

$$\boldsymbol{\kappa}_\epsilon^{s+1,t+1} = \lim_{N \to \infty} \mathbb{E}\left[\frac{1}{N} (f^s_{\epsilon\mathbf{Y}}(\mathbf{Z}^s_{\epsilon\mathbf{Y}}))^\top f^t_{\epsilon\mathbf{Y}}(\mathbf{Z}^t_{\epsilon\mathbf{Y}}) \right] \tag{3.117}$$

where $\mathbf{Z}^s_{\epsilon\mathbf{Y}}, \mathbf{Z}^t_{\epsilon\mathbf{Y}}$ are $n \times q$ Gaussian random matrices whose distributions are specified by $\boldsymbol{\kappa}_\epsilon^{s,s}, \boldsymbol{\kappa}_\epsilon^{t,t}$ and $\boldsymbol{\kappa}_\epsilon^{s,t}$ which are $q \times q$ deterministic matrices. Then, from equation (3.109), we have

$$\boldsymbol{\kappa}_\epsilon^{s+1,t+1} = \lim_{N \to \infty} \frac{1}{N} \mathbb{E}_{\mathbf{Z}} \left[\frac{1}{N} (f_s(\mathbf{Z}^{\epsilon,s}))^\top f_t(\mathbf{Z}^{\epsilon,t}) \right] + \epsilon^2 \mathbf{I}_{q \times q} \delta_{s=t} \tag{3.118}$$

From Lemma 18, the function $(\mathbf{Z}^s, \mathbf{Z}^t) \to \frac{1}{N} f_s(\mathbf{Z}^s)^\top f_t(\mathbf{Z}^t)$ is uniformly pseudo-Lipschitz. Moreover, from the induction hypothesis, we have :

$$\lim_{\epsilon \to 0} \boldsymbol{\kappa}_\epsilon^{s,t} = \boldsymbol{\kappa}^{s,t} \tag{3.119}$$

thus, using the uniform convergence Lemma 12, we get :

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\left[f_s(\mathbf{Z}^{\epsilon,s})^\top f_t(\mathbf{Z}^{\epsilon,t})\right]=\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\left[f_s(\mathbf{Z}^{s})^\top f_t(\mathbf{Z}^{t})\right]=\boldsymbol{\kappa}^{s+1,t+1} \tag{3.120}$$

where $(\mathbf{Z}^s,\mathbf{Z}^t)\sim\mathbb{N}(0,\boldsymbol{\kappa}\otimes\mathbf{I}_n)$ and $\boldsymbol{\kappa}=\begin{bmatrix}\boldsymbol{\kappa}^{s,s},\boldsymbol{\kappa}^{s,t}\\ \boldsymbol{\kappa}^{t,s},\boldsymbol{\kappa}^{t,t}\end{bmatrix}$. This shows that

$$\boldsymbol{\kappa}_\epsilon^{s+1,t+1}\xrightarrow[\epsilon\to 0]{}\boldsymbol{\kappa}^{s+1,t+1} \tag{3.121}$$

which concludes the induction. Similar reasoning proves the convergence of correlations with the initial vector

$$\boldsymbol{\kappa}_\epsilon^{1,t+1}\xrightarrow[\epsilon\to 0]{}\boldsymbol{\kappa}^{1,t+1} \tag{3.122}$$

. □

Proof of Lemma 14. This Lemma is proven by induction.

Initialization.

$$\frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},0}-\mathbf{M}^0\right\|_F=f^0_{\epsilon\mathbf{Y}}(\mathbf{X}^0)-f^0(\mathbf{X}^0)=\frac{1}{\sqrt{N}}\epsilon\left\|\mathbf{Y}^0\right\|_F \tag{3.123}$$

Using the bound from Lemma 4, there exists an absolute constant $C_{\mathbf{Y}}$ independent of N such that, with high probability:

$$\frac{\epsilon}{\sqrt{N}}\left\|\mathbf{Y}^0\right\|_F\leqslant C_{\mathbf{Y}}\epsilon \tag{3.124}$$

Note that $C_{\mathbf{Y}}$ is the same for all $\mathbf{Y}^t$. We thus choose $h_0'(\epsilon)=C_{\mathbf{Y}}\epsilon$. Then

$$\frac{1}{\sqrt{N}}\left\|\mathbf{X}^{\epsilon\mathbf{Y},1}-\mathbf{X}^1\right\|_F\leqslant\|\mathbf{A}\|_{op}\frac{\epsilon}{\sqrt{N}}\left\|\mathbf{Y}^0\right\|_F\leqslant 2C_{\mathbf{Y}}\epsilon \tag{3.125}$$

using the bound on the operator norm of GOE matrices Proposition 5, and we can choose $h_0(\epsilon)=2C_{\mathbf{Y}}\epsilon$.

Induction Assume the property is verified up to time t, i.e., the functions $h_0(\epsilon),h_0'(\epsilon),...,h_{t-1}(\epsilon)$, $h_{t-1}'(\epsilon)$ exist and are known. We now need to show $h_t(\epsilon),h_t'(\epsilon)$ exist. By definition of the iteration:

$$\begin{aligned}
\frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t}-\mathbf{M}^t\right\|_F&=\frac{1}{\sqrt{N}}\left\|f^t_{\epsilon\mathbf{Y}}(\mathbf{X}^{\epsilon\mathbf{Y}})-f^t(\mathbf{X}^t)\right\|_F\\
&=\frac{1}{\sqrt{N}}\left\|f^t(\mathbf{X}^{\epsilon\mathbf{Y}})-f^t(\mathbf{X}^t)+\epsilon\mathbf{Y}^t\right\|_F\\
&\leqslant L_t\left(1+\left(\frac{\left\|\mathbf{X}^{\epsilon\mathbf{Y},t}\right\|_F}{\sqrt{N}}\right)^{k-1}+\left(\frac{\left\|\mathbf{X}^{t}\right\|_F}{\sqrt{N}}\right)^{k-1}\right)\frac{\left\|\mathbf{X}^{\epsilon\mathbf{Y},t}-\mathbf{X}^t\right\|_F}{\sqrt{N}}+\frac{1}{\sqrt{N}}\epsilon\left\|\mathbf{Y}^t\right\|_F\\
&\leqslant L_t\left(1+\left(\frac{\left\|\mathbf{X}^{\epsilon\mathbf{Y},t}\right\|_F}{\sqrt{N}}\right)^{k-1}+\left(\frac{\left\|\mathbf{X}^{t}\right\|_F}{\sqrt{N}}\right)^{k-1}\right)h_{t-1}(\epsilon)+C_{\mathbf{Y}}\epsilon\\
&\leqslant L_t\left(1+C_{\epsilon\mathbf{Y}}(k)+\left(\frac{\left\|\mathbf{X}^{\epsilon\mathbf{Y},t}\right\|_F}{\sqrt{N}}+\frac{\left\|\mathbf{X}^{\epsilon\mathbf{Y},t}-\mathbf{X}^t\right\|_F}{\sqrt{N}}\right)^{k-1}\right)h_{t-1}(\epsilon)+C_{\mathbf{Y}}\epsilon\\
&\leqslant L_t\left(1+C_{\epsilon\mathbf{Y}}(k)+2^{k-2}C_{\epsilon\mathbf{Y}}(k)^{k-1}+2^{k-2}h_{t-1}^{k-1}(\epsilon)\right)h_{t-1}(\epsilon)+C_{\mathbf{Y}}\epsilon
\end{aligned} \tag{3.126}$$

where we used the state evolution of the perturbed AMP orbit to show that $\frac{\|\mathbf{X}^{\epsilon\mathbf{Y},t}\|_F}{\sqrt{N}}$ has a finite limit and Hölder's inequality. We can thus choose

$$h'_t(\epsilon) = L_t\left(1 + C_{\epsilon\mathbf{Y}}(k) + 2^{k-2}C_{\epsilon\mathbf{Y}}(k)^{k-1} + 2^{k-2}h_{t-1}^{k-1}(\epsilon)\right)h_{t-1}(\epsilon) + C_{\mathbf{Y}}\epsilon \tag{3.127}$$

which goes to zero when ϵ goes to zero. Then

$$\begin{aligned}\frac{1}{\sqrt{N}}\left\|\mathbf{X}^{\epsilon\mathbf{Y},t+1} - \mathbf{X}^{t+1}\right\|_F &\leqslant \|\mathbf{A}\|_{op}\frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t} - \mathbf{M}^t\right\|_F + \frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t-1}(\mathbf{b}^t_{\epsilon\mathbf{Y}})^\top - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top\right\|_F \\ &\leqslant 2h'_t(\epsilon) + \frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t-1}(\mathbf{b}^{\epsilon\mathbf{Y}}_t)^\top - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top\right\|_F \\ &\leqslant 2h'_t(\epsilon) + \frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t-1} - \mathbf{M}^{t-1}\right\|_F\left\|\mathbf{b}^t\right\|_F + \frac{1}{\sqrt{N}}\left\|\mathbf{b}^t_{\epsilon\mathbf{Y}} - \mathbf{b}^t\right\|_F\left\|\mathbf{M}^{t-1}\right\|_F\end{aligned} \tag{3.128}$$

and

$$\begin{aligned}\left\|\mathbf{b}^t\right\|_F &= \left\|\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{Z}_i}(\mathbf{Z}^t)\right]\right\|_F \\ &\leqslant \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left\|\frac{\partial f_i^t}{\partial \mathbf{Z}_i}(\mathbf{Z}^t)\right\|_F\right]\end{aligned} \tag{3.129}$$

where $\mathbf{Z}^t \sim \mathbf{N}(0, \boldsymbol{\kappa}_{t,t} \otimes \mathbf{I}_n)$. Since the function $f^t : \mathbb{R}^{N\times q} \to \mathbb{R}^{N\times q}$ is pseudo-Lipschitz of order k, the components $f_i^t : \mathbb{R}^{N\times q} \to \mathbb{R}^q$ are pseudo-Lipschitz of order k as well. So are the functions $f_{i,j}^t : \mathbb{R}^{N\times q} \to \mathbb{R}$ for $1 \leqslant j \leqslant q$ generating each component of $f_i^t(\mathbf{Z}^t) \in \mathbb{R}^q$ and their $\mathbb{R}^q \to \mathbb{R}$ restrictions to the $i-th$ line of $\mathbf{Z}^t$. Then

$$\left\|\mathbf{b}^t\right\|_F \leqslant \frac{1}{N}\sum_{i=1}^N q \max_j\left\{\mathbb{E}\left\|\nabla_{\mathbf{Z}_i^t} f_{i,j}^t(\mathbf{Z}^t)\right\|_2\right\} \tag{3.130}$$

where $\max_j\left\{\mathbb{E}\left\|\nabla_{\mathbf{Z}_i^t} f_{i,j}^t(\mathbf{Z}^t)\right\|_2\right\}$ is bounded using the pseudo-Lipschitz property and a similar argument to the proof of lemma 1. Let C_J be this upper bound, then

$$\frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t-1}(\mathbf{b}^t_{\epsilon\mathbf{Y}})^\top - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top\right\|_F \leqslant qC_J h'_{t-1}(\epsilon) + \frac{1}{\sqrt{N}}\left\|\mathbf{M}^{t-1}\right\|_F\left\|\mathbf{b}^t_{\epsilon\mathbf{Y}} - \mathbf{b}^t\right\|_F \tag{3.131}$$

Using the same decomposition as before

$$\begin{aligned}\frac{1}{\sqrt{N}}\left\|\mathbf{M}^{t-1}\right\|_F\left\|\mathbf{b}^t_{\epsilon\mathbf{Y}} - \mathbf{b}^t\right\|_F &\leqslant \left(\frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t-1} - \mathbf{M}^{t-1}\right\| + \frac{1}{\sqrt{N}}\left\|\mathbf{M}^{\epsilon\mathbf{Y},t-1}\right\|\right)\left\|\mathbf{b}^t_{\epsilon\mathbf{Y}} - \mathbf{b}^t\right\|_F \\ &\leqslant \left(h'_{t-1}(\epsilon) + C_{\epsilon\mathbf{Y},t-1}\right)\left\|\mathbf{b}^t_{\epsilon\mathbf{Y}} - \mathbf{b}^t\right\|_F\end{aligned} \tag{3.132}$$

The definition of the Onsager correction terms gives

$$\left\|\mathbf{b}^t_{\epsilon\mathbf{Y}} - \mathbf{b}^t\right\|_F = \left\|\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial f_i^t}{\partial \tilde{\mathbf{Z}}_i^{\epsilon\mathbf{Y},t}}(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})\right] - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \frac{\partial f_i^t}{\partial \tilde{\mathbf{Z}}_i^t}(\tilde{\mathbf{Z}}^t)\right]\right\|_F \tag{3.133}$$

where $\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t} = \mathbf{Z}(\boldsymbol{\kappa}^{\epsilon\mathbf{Y}}_{t,t})^{1/2}$ where $\mathbf{Z} \in \mathbb{R}^{N\times q}$ is an i.i.d. standard normal matrix. Similarly $\tilde{\mathbf{Z}}^{t} = \mathbf{Z}(\boldsymbol{\kappa}_{t,t})^{1/2}$. Using the positive definiteness of $\boldsymbol{\kappa}_{t,t}$ along with Lemma 17, we can write, keeping in mind that the perturbation $\epsilon\mathbf{Y}$ doesn't change the derivatives in the Onsager correction:

$$\begin{aligned}\left\|\mathbf{b}^t_{\epsilon\mathbf{Y}} - \mathbf{b}^t\right\|_F &= \left\|(\boldsymbol{\kappa}^{t,t}_{\epsilon\mathbf{Y}})^{-1}\mathbb{E}\left[\frac{1}{N}(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})^\top f^t(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})\right] - (\boldsymbol{\kappa}^{t,t})^{-1}\mathbb{E}\left[\frac{1}{N}(\mathbf{Z}^t)^\top f^t(\mathbf{Z}^t)\right]\right\|_F \\ &\leqslant \left\|(\boldsymbol{\kappa}^{t,t}_{\epsilon\mathbf{Y}})^{-1} - (\boldsymbol{\kappa}^{t,t})^{-1}\right\|_F \mathbb{E}\left[\frac{1}{N}(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})^\top f^t(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})\right] + \\ &\qquad ((\boldsymbol{\kappa}^{t,t})^{-1})\left\|\mathbb{E}\left[\frac{1}{N}(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})^\top f^t(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})\right] - \mathbb{E}\left[\frac{1}{N}(\mathbf{Z}^t)^\top f^t(\mathbf{Z}^t)\right]\right\|_F\end{aligned}$$

The function $\mathbb{R}^{N\times q} \to \mathbb{R}^{q\times q}, \mathbf{Z} \to \mathbf{Z}^\top f^t(\mathbf{Z})$ is pseudo-Lipschitz of order $k+1$. Moreover, from Lemma 8, $\boldsymbol{\kappa}^{t,t}_{\epsilon\mathbf{Y}} \xrightarrow[\epsilon\to 0]{} \boldsymbol{\kappa}^{t,t}$. Thus using Lemma 12, we get

$$\lim_{\epsilon\to 0}\left\|\mathbb{E}\left[\frac{1}{N}(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})^\top f_t(\tilde{\mathbf{Z}}^{\epsilon\mathbf{Y},t})\right] - \mathbb{E}\left[\frac{1}{N}(\mathbf{Z}^t)^\top f_t(\mathbf{Z}^t)\right]\right\|_F = 0 \tag{3.134}$$

and Lemma 13 gives $\lim_{\epsilon\to 0}\left\|(\boldsymbol{\kappa}^{\epsilon\mathbf{Y}}_{t,t})^{-1} - (\boldsymbol{\kappa}_{t,t})^{-1}\right\|_F = 0$, which concludes the induction. □

3.4 Low-rank perturbations and projections

As mentioned in Section 2.2.3, AMP iterations associated to inference problems often present non-trivial dependencies between the non-linearities and the random matrices of the corresponding graph. These dependencies typically take the form of low-rank linear perturbations, or an additional argument in the non-linearities composed of a non-linear transform involving the random matrices of the graph, see the examples of Section 2.3. In this appendix, we propose a generic way of dealing with these dependencies by leveraging on the matrix-valued iteration Eq.(2.9-2.11), in the form of two lemmas.

3.4.1 Additive low-rank perturbation

Lemma 15. *Let $\mathbf{V}_0 \in \mathbb{R}^{N\times q}$ be a given matrix such that the quantity $\frac{1}{\sqrt{N}}\|\mathbf{V}_0\|_F$ converges to a finite constant as $N \to \infty$. Define the matrix*

$$\hat{\mathbf{A}} = \mathbf{A} + \frac{1}{N}\mathbf{V}_0\mathbf{V}_0^\top \quad \in \mathbb{R}^{N\times N}, \tag{3.135}$$

consider the AMP iteration initialized with $\mathbf{X}^0 \in \mathbb{R}^{N\times q}$

$$\mathbf{X}^{t+1} = \hat{\mathbf{A}}\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top \qquad \in \mathbb{R}^{N\times q}, \tag{3.136}$$

$$\mathbf{M}^t = f^t(\mathbf{X}^t) \qquad \in \mathbb{R}^{N\times q}, \tag{3.137}$$

$$\mathbf{b}^t = \frac{1}{N}\sum_{i=1}^N \frac{\partial f^t_i}{\partial \mathbf{X}_i}(\mathbf{X}^t) \qquad \in \mathbb{R}^{q\times q}. \tag{3.138}$$

and the following state evolution recursion, initialized with $\boldsymbol{\mu}_0 = 0_{q\times q}$,

$$\boldsymbol{\mu}_0,\ \boldsymbol{\kappa}^{1,1} = \lim_{N\to\infty} \frac{1}{N} f^0(\mathbf{V}_0\boldsymbol{\mu}_0 + \mathbf{X}^0)^\top f^0(\mathbf{V}_0\boldsymbol{\mu}_0 + \mathbf{X}^0) \tag{3.139}$$

$$\boldsymbol{\mu}^{s+1} = \lim_{N\to+\infty} \frac{1}{N}\mathbb{E}\left[(\mathbf{V}_0)^\top f^s\left(\mathbf{V}_0\boldsymbol{\mu}^s + \mathbf{Z}^s\right)\right] \tag{3.140}$$

$$\boldsymbol{\kappa}^{t+1,s+1} = \boldsymbol{\kappa}^{s+1,t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[f^s(\mathbf{V}_0\boldsymbol{\mu}^s + \mathbf{Z}^s)^\top f^t(\mathbf{V}_0\boldsymbol{\mu}^t + \mathbf{Z}^t)\right], \qquad s \in \{0,\dots,t\}. \tag{3.141}$$

where $(\mathbf{Z}^1,...,\mathbf{Z}^t) \sim \mathbf{N}\left(0, (\kappa^{s,r})_{s,r\leqslant t} \otimes \mathbf{I}_N\right)$. *Assume (B1)−(B6) and that for any* $t \in \mathbb{N}$, *any* $1 \leqslant i \leqslant N$, *the derivative* $\frac{\partial f_i^t}{\partial \mathbf{X}_i}$ *is pseudo-Lipschitz of order* k. *Then for any sequence* $\phi_N : (\mathbb{R}^{N\times q})^{t+1} \to \mathbb{R}$ *of pseudo-Lipschitz functions*

$$\phi_N\left(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t\right) \overset{\mathrm{P}}{\backsimeq} \mathbb{E}\left[\phi_N\left(\mathbf{V}_0\boldsymbol{\mu}^0 + \mathbf{Z}^0, \mathbf{V}_0\boldsymbol{\mu}^1 + \mathbf{Z}^1, ..., \mathbf{V}_0\boldsymbol{\mu}^t + \mathbf{Z}^t\right)\right] \tag{3.142}$$

Proof of Lemma 15. The proof follows a similar argument to that of Lemma 3.4 from [74]. Consider the following iteration

$$\mathbf{S}^{t+1} = \mathbf{A}\tilde{\mathbf{M}}^t - \tilde{\mathbf{m}}^{t-1}(\tilde{\mathbf{b}}^t)^\top \qquad \in \mathbb{R}^{N\times q}, \tag{3.143}$$

$$\tilde{\mathbf{M}}^t = f^t(\mathbf{V}_0\boldsymbol{\mu}^t + \mathbf{S}^t) \qquad \in \mathbb{R}^{N\times q}, \tag{3.144}$$

$$\tilde{\mathbf{b}}^t = \frac{1}{N}\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{S}_i}(\mathbf{V}_0\boldsymbol{\mu}^t + \mathbf{S}^t) \qquad \in \mathbb{R}^{q\times q}. \tag{3.145}$$

initialized with $\mathbf{S}^0 = \mathbf{X}^0 - \boldsymbol{\mu}_0\mathbf{V}_0$. Under assumptions $(B1) - (B6)$, the iterates $\mathbf{S}^t$ obey the state evolution equations Eq.(3.139) owing to Theorem 5. We now prove the following statement by induction.

$$\forall t \in \mathbb{N} \quad \frac{1}{\sqrt{N}}\left\|\mathbf{X}^t - \mathbf{S}^t - \mathbf{V}_0\boldsymbol{\mu}^t\right\|_F \xrightarrow[N\to\infty]{P} 0 \tag{3.146}$$

The statement is true at $t = 0$ owing to the initialization of the sequences. Assume the statement is true up to time t. We can then write

$$\mathbf{X}^{t+1} - \mathbf{S}^{t+1} - \mathbf{V}_0\boldsymbol{\mu}^{t+1} = \hat{\mathbf{A}}\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top - \mathbf{A}\tilde{\mathbf{M}}^t + \tilde{\mathbf{m}}^{t-1}(\tilde{\mathbf{b}}^t)^\top - \mathbf{V}_0\boldsymbol{\mu}^{t+1} \tag{3.147}$$

$$= \mathbf{A}\left(f^t(\mathbf{X}^t) - f^t\left(\mathbf{V}_0\boldsymbol{\mu}^t + \mathbf{S}^t\right)\right) + \frac{1}{N}\mathbf{V}_0\mathbf{V}_0^\top f^t\left(\mathbf{X}^t\right) - \mathbf{V}_0\boldsymbol{\mu}^{t+1}$$
$$+ \left(f^{t-1}(\mathbf{V}_0\boldsymbol{\mu}^{t-1} + \mathbf{S}^{t-1}) - f^{t-1}(\mathbf{X}^{t-1}))\right)(\tilde{\mathbf{b}}^t)^\top + f^{t-1}(\mathbf{X}^{t-1})(\tilde{\mathbf{b}}^t - \mathbf{b}^t)^\top \tag{3.148}$$

The triangle inequality then gives

$$\frac{1}{\sqrt{N}}\left\|\mathbf{X}^{t+1} - \mathbf{S}^{t+1} - \mathbf{V}_0\boldsymbol{\mu}^{t+1}\right\|_N \leqslant \frac{1}{\sqrt{N}}\|\mathbf{A}\|_{op}\left\|f^t(\mathbf{X}^t) - f^t\left(\mathbf{V}_0\boldsymbol{\mu}^t + \mathbf{S}^t\right)\right\|_F$$
$$+ \frac{1}{\sqrt{N}}\left\|\frac{1}{N}\mathbf{V}_0\mathbf{V}_0^\top f^t\left(\mathbf{X}^t\right) - \mathbf{V}_0\boldsymbol{\mu}^{t+1}\right\|_F$$
$$+ \frac{1}{\sqrt{N}}\left\|\left(f^{t-1}(\mathbf{V}_0\boldsymbol{\mu}^{t-1} + \mathbf{S}^{t-1}) - f^{t-1}(\mathbf{X}^{t-1}))\right)(\tilde{\mathbf{b}}^t)^\top\right\|_F + \frac{1}{\sqrt{N}}\left\|f^{t-1}(\mathbf{X}^{t-1})(\tilde{\mathbf{b}}^t - \mathbf{b}^t)^\top\right\|_F \tag{3.149}$$

and, owing to the pseudo-Lipschitz property

$$\frac{1}{\sqrt{N}}\left\|f^t(\mathbf{X}^t)-f^t\left(\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t\right)\right\|_F \leqslant L\left(1+\left(\frac{\|\mathbf{X}^t\|_F}{\sqrt{N}}\right)^{k-1}+\left(\frac{\|\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t\|_F}{\sqrt{N}}\right)^{k-1}\right)\frac{\|\mathbf{X}^t-\mathbf{V}^0\boldsymbol{\mu}^t-\mathbf{S}^t\|_F}{\sqrt{N}}, \tag{3.150}$$

where the state evolution verified by iteration Eq.(3.143) ensures that $\frac{\|\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t\|_F}{\sqrt{N}}$ is bounded with high probability. The induction hypothesis then gives that $\frac{\|\mathbf{X}^t-\mathbf{V}^0\boldsymbol{\mu}^t-\mathbf{S}^t\|_F}{\sqrt{N}} \xrightarrow[N\to\infty]{P} 0$, which, together with the previous statement ensures that $\frac{\|\mathbf{X}^t\|_F}{\sqrt{N}}$ is also bounded with high probability. Combining this with proposition 5 shows that

$$\frac{1}{\sqrt{N}}\|\mathbf{A}\|_{op}\left\|f^t(\mathbf{X}^t)-f^t\left(\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t\right)\right\|_F \xrightarrow[N\to\infty]{P} 0. \tag{3.151}$$

Then

$$\frac{1}{\sqrt{N}}\left\|\frac{1}{N}\mathbf{V}_0\mathbf{V}_0^\top f^t\left(\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t\right)-\mathbf{V}_0\boldsymbol{\mu}^{t+1}\right\|_F \leqslant \frac{\|\mathbf{V}_0\|_F}{\sqrt{N}}\left\|\frac{1}{N}\mathbf{V}_0^\top f^t\left(\mathbf{X}^t\right)-\boldsymbol{\mu}^{t+1}\right\|_F \tag{3.152}$$

where $\|\mathbf{V}_0\|_F/\sqrt{N}$ is bounded with high probability by assumption. Since the function $\mathbf{V}_0^\top f^t(.)$ is pseudo-Lipschitz, we can use the induction hypothesis and SE equations together with the definition of $\boldsymbol{\mu}^t$ show that the r.h.s. goes to zero with high probability. The third term of the sum in the r.h.s. of Eq.(3.149) can be bounded in similar fashion to the first one using the pseudo-Lipschitz property, the induction hypothesis and the boundedness of the norm of the Onsager term $\tilde{\mathbf{b}}^t$, which can be expressed as a pseudo-Lipschitz function of $\mathbf{S}^t$ using the SE property of iteration Eq.(3.143) and Lemma 17. The last term then verifies

$$\frac{1}{\sqrt{N}}\left\|f^{t-1}(\mathbf{X}^{t-1})(\tilde{\mathbf{b}}^t-\mathbf{b}^t)^\top\right\|_F \leqslant \frac{1}{\sqrt{N}}\left\|f^{t-1}(\mathbf{X}^{t-1})\right\|_F\left\|\tilde{\mathbf{b}}^t-\mathbf{b}^t\right\|_F \tag{3.153}$$

where $\frac{1}{\sqrt{N}}\|f^{t-1}(\mathbf{X}^{t-1})\|_F$ is bounded w.h.p. owing to the induction hypothesis, pseudo-Lipschitz property of f^{t-1} and the SE equations of iteration Eq.(3.143), and the difference in Onsager terms verifies

$$\begin{aligned}\left\|\tilde{\mathbf{b}}^t-\mathbf{b}^t\right\|_F &= \frac{1}{N}\left\|\sum_{i=1}^N\left(\frac{\partial f_i^t}{\partial \mathbf{S}_i}(\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t)-\frac{\partial f_i^t}{\partial \mathbf{X}_i}(\mathbf{X}^t)\right)\right\|_F \\ &\leqslant \sup_{1\leqslant i\leqslant N}\left\|\frac{\partial f_i^t}{\partial \mathbf{S}_i}(\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t)-\frac{\partial f_i^t}{\partial \mathbf{X}_i}(\mathbf{X}^t)\right\|_F\end{aligned} \tag{3.154}$$

where we remind that $f_i^t:\mathbb{R}^{N\times q}\to\mathbb{R}^q$ and is therefore a low-dimensional observable, for which the pseudo-Lipschitz assumption implies that there exists a constant L such that

$$\left\|\tilde{\mathbf{b}}^t-\mathbf{b}^t\right\|_F \leqslant L\left(1+\left(\frac{\|\mathbf{X}^t\|_F}{\sqrt{N}}\right)^{k-1}+\left(\frac{\|\mathbf{V}_0\boldsymbol{\mu}^t+\mathbf{S}^t\|_F}{\sqrt{N}}\right)^{k-1}\right)\frac{\|\mathbf{X}^t-\mathbf{V}^0\boldsymbol{\mu}^t-\mathbf{S}^t\|_F}{\sqrt{N}} \tag{3.155}$$

which converges to zero with high probability for large N using the induction hypthesis and the SE equations of iteration (3.143). This concludes the induction and proves the statement Eq.(3.146). The proof of Lemma 15 follows immediately from the pseudo-Lipschitz property, the property Eq.(3.146) and the SE equations of iteration Eq.(3.143). □

3.4.2 Dependence on an additional linear observation

Lemma 16. *Let $\mathbf{W}_0 \in \mathbb{R}^{N\times q}$ be a matrix such that $\frac{1}{N}\left\|\mathbf{W}_0^\top \mathbf{W}_0\right\|_F$ converges to a finite constant as $N \to \infty$, and a given pseudo-Lipschitz function $\varphi : \mathbb{R}^{N\times q} \to \mathbb{R}^N$. Consider the AMP iteration initialized with $\mathbf{X}^0 \in \mathbb{R}^{N\times q}$*

$$\mathbf{X}^{t+1} = \mathbf{A}\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top \qquad \in \mathbb{R}^{N\times q}, \tag{3.156}$$

$$\mathbf{M}^t = f^t(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t) \qquad \in \mathbb{R}^{N\times q}, \tag{3.157}$$

$$\mathbf{b}^t = \frac{1}{N}\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{X}_i}(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t) \qquad \in \mathbb{R}^{q\times q}. \tag{3.158}$$

where the functions $f^t : \mathbb{R}^{N\times(q+1)} \to \mathbb{R}^{N\times q}$ are pseudo-Lipschitz. Consider the following state evolution recursion, initialized with $\boldsymbol{\nu}^0, \hat{\boldsymbol{\nu}}^0 = 0_{q\times q}$,

$$\boldsymbol{\nu}^0, \hat{\boldsymbol{\nu}}^0, \boldsymbol{\kappa}^{1,1} = \frac{1}{N} f^0(\mathbf{X}^0)^\top f^0(\mathbf{X}^0) \tag{3.159}$$

$$\boldsymbol{\nu}^{t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\mathbf{W}_0^\top f^t\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right] \tag{3.160}$$

$$\hat{\boldsymbol{\nu}}^{t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{Z}_{\mathbf{W}_0,i,\varphi}}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right] \tag{3.161}$$

$$\boldsymbol{\kappa}^{t+1,s+1} = \boldsymbol{\kappa}^{s+1,t+1} =$$
$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Big[\left(f^s\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^s + \mathbf{W}_0\hat{\boldsymbol{\nu}}^s + \mathbf{Z}^s\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{s+1}\right)^\top$$
$$\left(f^t\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right)\Big] \tag{3.162}$$

where the notation $\partial \mathbf{Z}_{\mathbf{W}_{0,i,\varphi}}$ denotes a derivatives w.r.t. the argument of φ, $\rho_{\mathbf{W}_0} = \frac{1}{N}\mathbf{W}_0^\top\mathbf{W}_0$, and $\mathbf{Z}_{\mathbf{W}_0} \sim \mathbf{N}(0, \rho_{\mathbf{W}_0} \otimes \mathbf{I}_N)$ is independent from the $(\mathbf{Z}^1, ..., \mathbf{Z}^t) \sim \mathbf{N}\left(0, (\kappa^{s,r})_{s,r\leqslant t} \otimes \mathbf{I}_N\right)$. Assume (B1) – (B6) and that for any $t \in \mathbb{N}$, any $1 \leqslant i \leqslant N$, the derivative $\frac{\partial f_i^t}{\partial \mathbf{X}_i}$ is pseudo-Lipschitz of order k. Then for any sequence $\phi_N : (\mathbb{R}^{N\times q})^{t+1} \to \mathbb{R}$ of pseudo-Lipschitz functions

$$\phi_N\left(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_N\left(\mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^0 + \mathbf{W}_0\hat{\boldsymbol{\nu}}^0 + \mathbf{Z}^0, ..., \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right] \tag{3.163}$$

Proof of lemma 16. Consider the following iteration

$$\mathbf{S}^{t+1} = \tilde{\mathbf{A}}\tilde{\mathbf{M}}^t - \tilde{\mathbf{m}}^{t-1}(\tilde{\mathbf{b}}^t)^\top \qquad \in \mathbb{R}^{N\times q}, \tag{3.164}$$

$$\tilde{\mathbf{M}}^t = f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1} \qquad \in \mathbb{R}^{N\times q}, \tag{3.165}$$

$$\tilde{\mathbf{b}}^t = \frac{1}{N}\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{S}_i}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) \qquad \in \mathbb{R}^{q\times q}. \tag{3.166}$$

where $\tilde{\mathbf{A}}$ is a copy of $\mathbf{A}$ independent on $\mathbf{Z}_{\mathbf{W}_0}$. Under assumptions (B1) – (B6) and conditionally on $\mathbf{A}\mathbf{W}_0$, the iterates $\mathbf{S}^t$ obey the state evolution equations Eq.(3.159) where the $\mathbf{Z}_{\mathbf{W}_0}$ are replaced by fixed $\mathbf{A}\mathbf{W}_0$, owing to Theorem 5. For any t, the composition of f^t and φ is pseudo-Lipschitz

of order k, and owing to Lemma 21, $\frac{1}{\sqrt{N}}\|\mathbf{A}\mathbf{W}_0 - \mathbf{Z}_{\mathbf{W}_0}\|_F \xrightarrow[N\to+\infty]{P} 0$. Using the pseudo-Lipschitz property, the assumption on $\mathbf{W}_0$ to bound the norms of $\frac{1}{\sqrt{N}}\mathbf{A}\mathbf{W}_0$ and $\frac{1}{\sqrt{N}}\mathbf{W}_0$ w.h.p., and Lemma 1, we obtain that iteration Eq.(3.164) verifies the SE equations Eq.(3.159), where the expectations are taken w.r.t. $\mathbf{Z}_{\mathbf{W}_0}$ and all the $\mathbf{Z}^s$ for $0 \leqslant s \leqslant t$. We now prove the following statement by induction

$$\forall t \in \mathbb{N} \quad \frac{1}{\sqrt{N}}\left\|\mathbf{X}^t - \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t - \mathbf{W}_0\hat{\boldsymbol{\nu}}^t - \mathbf{S}^t\right\|_F \xrightarrow[N\to\infty]{P} 0 \tag{3.167}$$

The property is true at $t = 0$ owing to the initialization of both sequences. Assume the property is verified up to time t. Then, denoting the increment $\Delta^t = \mathbf{X}^t - \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t - \mathbf{W}_0\hat{\boldsymbol{\nu}}^t - \mathbf{S}^t$

$$\Delta^t = \mathbf{A}\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top - \left(\tilde{\mathbf{A}}\tilde{\mathbf{M}}^t - \tilde{\mathbf{m}}^{t-1}(\tilde{\mathbf{b}}^t)^\top\right) - \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1} - \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t+1} \tag{3.168}$$

Consider then the iteration Eq.(3.156), where we condition on the value of $\mathbf{A}\mathbf{W}_0$ at each iteration. A straightforward induction starting from the initialization then shows that, for any $t \in \mathbb{N}$

$$\begin{aligned}\mathbf{X}^{t+1}_{|\mathbf{A}\mathbf{W}_0} = \mathbf{A}_{|\mathbf{A}\mathbf{W}_0} f^t\left(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \mathbf{X}^t_{|\mathbf{A}\mathbf{W}_0}\right) \\ - f^{t-1}\left(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \mathbf{X}^{t-1}_{|\mathbf{A}\mathbf{W}_0}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial f_i^t}{\partial \mathbf{X}_i}\left(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \mathbf{X}^t_{|\mathbf{A}\mathbf{W}_0}\right)\right)^\top\end{aligned} \tag{3.169}$$

Using the same lemma from [28, 135] used in the proof of Lemma 6, we may write

$$\begin{aligned}\mathbf{A}_{|\mathbf{A}\mathbf{W}_0} &= \mathbf{A} - \mathbf{P}_{\mathbf{W}_0}\mathbf{A}\mathbf{P}_{\mathbf{W}_0} + \mathbf{P}^\perp_{\mathbf{W}_0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{W}_0} &(3.170)\\ &= \mathbf{A}\mathbf{P}_{\mathbf{W}_0} + \mathbf{P}_{\mathbf{W}_0}\mathbf{A} - \mathbf{P}_{\mathbf{W}_0}\mathbf{A}\mathbf{P}_{\mathbf{W}_0} + \mathbf{P}^\perp_{\mathbf{W}_0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{W}_0} &(3.171)\end{aligned}$$

where $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$ and $\mathbf{P}_{\mathbf{W}_0} = \mathbf{W}_0\left(\mathbf{W}_0^\top\mathbf{W}_0^\top\right)^{-1}\mathbf{W}_0^\top = \frac{1}{N}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\mathbf{W}_0^\top$ is always well-defined for $n \geqslant q$. We can then lift the conditioning by considering the distribution of $\mathbf{A}\mathbf{W}_0$ (which is straightforward since there is no correlation between $\mathbf{A}$ and $\mathbf{W}_0$) in all subsequent expressions. The increment Eq.(3.168) becomes

$$\begin{aligned}&\left(\mathbf{A}\mathbf{P}_{\mathbf{W}_0} + \mathbf{P}_{\mathbf{W}_0}\mathbf{A} - \mathbf{P}_{\mathbf{W}_0}\mathbf{A}\mathbf{P}_{\mathbf{W}_0} + \mathbf{P}^\perp_{\mathbf{W}_0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{W}_0}\right)\mathbf{M}^t - \mathbf{M}^{t-1}(\mathbf{b}^t)^\top - \left(\tilde{\mathbf{A}}\tilde{\mathbf{M}}^t - \tilde{\mathbf{m}}^{t-1}(\tilde{\mathbf{b}}^t)^\top\right)\\ &- \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1} - \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t+1}\end{aligned} \tag{3.172}$$

where we chose the matrix $\tilde{\mathbf{A}}$ coming from the decomposition of $\mathbf{A}$ to define the iteration Eq.(3.164), and

$$\begin{aligned}&\Delta^t = \mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) + \mathbf{P}_{\mathbf{W}_0}\mathbf{A} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \mathbf{P}_{\mathbf{W}_0}\mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\\ &+ \mathbf{P}^\perp_{\mathbf{W}_0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^{t-1}\right)(\mathbf{b}^t)^\top - \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1} - \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t+1}\\ &- \tilde{\mathbf{A}}\left(f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right)\\ &+ \left(f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t-1} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t-1} + \mathbf{S}^{t-1}\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t\right)\left(\tilde{\mathbf{b}}^t\right)^\top\end{aligned} \tag{3.173}$$

$$\begin{aligned}&= \mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1} + \mathbf{P}_{\mathbf{W}_0}\mathbf{A} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t+1} - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t(\tilde{\mathbf{b}}^t)^\top\\ &- f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^{t-1}\right)(\mathbf{b}^t)^\top + f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t-1} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t-1} + \mathbf{S}^{t-1}\right)\left(\tilde{\mathbf{b}}^t\right)^\top\\ &- \tilde{\mathbf{A}}\left(f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right) + \mathbf{P}^\perp_{\mathbf{W}_0}\tilde{\mathbf{A}}\mathbf{P}^\perp_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\\ &- \mathbf{P}_{\mathbf{W}_0}\mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\end{aligned} \tag{3.174}$$

where, the second equality is only a reorganization of the terms. We now study the asymptotic behaviour of each component of the previous sum. We have

$$\frac{1}{\sqrt{N}}\left\|\mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right\|_F \leqslant \\ \|\mathbf{A}\|_{op}\frac{1}{\sqrt{N}}\left\|\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\right\|_F\left\|\frac{1}{N}\mathbf{W}_0^\top f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \boldsymbol{\nu}^{t+1}\right\|_F \quad (3.175)$$

where $\|\mathbf{A}\|_{op}$ is bounded w.h.p. owing to lemma 5 and $\frac{1}{\sqrt{N}}\left\|\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\right\|_F$ is bounded w.h.p. by assumption. Then, using the pseudo-Lipschitz property, the induction hypothesis and Lemma 1, it holds that

$$\frac{1}{\sqrt{N}}\left\|f^t\left(\phi(\mathbf{A}\mathbf{W}_0, \mathbf{X}^t)\right) - f^t\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right)\right\|_F \xrightarrow[N\to\infty]{P} 0 \quad (3.176)$$

The triangle inequality then gives

$$\left\|\frac{1}{N}\mathbf{W}_0^\top f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \boldsymbol{\nu}^{t+1}\right\|_F \leqslant \left\|\frac{1}{N}\mathbf{W}_0^\top f^t\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) - \boldsymbol{\nu}^{t+1}\right\|_F \times \\ \frac{1}{\sqrt{N}}\|\mathbf{W}_0\|_F\frac{1}{\sqrt{N}}\left\|f^t\left(\varphi(\mathbf{A}\mathbf{W}_0, \mathbf{X}^t)\right) - f^t\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right)\right\|_F. \quad (3.177)$$

Using the definition of $\boldsymbol{\mu}^{t+1}$, the assumption on $\mathbf{W}_0$ and Eq.(3.176), we conclude that, with high probability

$$\frac{1}{\sqrt{N}}\left\|\mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right\|_F \xrightarrow[N\to\infty]{} 0 \quad (3.178)$$

The term

$$\frac{1}{\sqrt{N}}\left\|f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t-1} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t-1} + \mathbf{S}^{t-1}\right)\left(\tilde{\mathbf{b}}^t\right)^\top - f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^{t-1}\right)(\mathbf{b}^t)^\top\right\|_F \\ \leqslant \frac{1}{\sqrt{N}}\left\|\left(f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t-1} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t-1} + \mathbf{S}^{t-1}\right) - f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^{t-1}\right)\right)\left(\tilde{\mathbf{b}}^t\right)^\top\right\|_F \\ + \frac{1}{\sqrt{N}}\left\|f^{t-1}\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^{t-1}\right)\left(\tilde{\mathbf{b}}^t - \mathbf{b}^t\right)\right\|_F, \quad (3.179)$$

is similar to the third term of Eq.(3.149) in the proof of Lemma 15 and converges to zero with high probability for large N using similar arguments. Then, letting

$$\Delta_1^t = \tilde{\mathbf{A}}\left(f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right) - \mathbf{P}_{\mathbf{W}_0}^{\perp}\tilde{\mathbf{A}}\mathbf{P}_{\mathbf{W}_0}^{\perp} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right), \quad (3.180)$$

the defintion of $\mathbf{P}_{\mathbf{W}_0}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{W}_0}$ and the triangle inequality yield

$$\frac{1}{\sqrt{N}}\left\|\Delta_1^t\right\|_F \leqslant \left\|\tilde{\mathbf{A}}\right\|_{op}\frac{1}{\sqrt{N}}\left\|\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right\|_F \\ + \left\|\tilde{\mathbf{A}}\right\|_{op}\frac{1}{\sqrt{N}}\left\|f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{A}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) - f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\right\|_F \\ + \frac{1}{\sqrt{N}}\left\|\mathbf{P}_{\mathbf{W}_0}\tilde{\mathbf{A}}\mathbf{P}_{\mathbf{W}_0}^{\perp} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\right\|_F \quad (3.181)$$

where the first term converges to zero w.h.p. using the same argument as the one used for Eq.(3.175). For the second term, the operator norm of $\tilde{\mathbf{A}}$ is bounded w.h.p. using Lemma 5, and the diffence goes to zero w.h.p. using the pseudo-Lipschitz property, the induction hypothesis and the SE equations Eq.(3.159) of iteration Eq.(3.164). Finally, since $\mathbf{P}_{\mathbf{W}_0}$ has finite rank and $\frac{1}{\sqrt{N}}\left\|\mathbf{P}^{\perp}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\right\|_F$ is bounded w.h.p. using the induction hypothesis and SE equations of iteration Eq.(3.164),the last term goes to zero w.h.p. using Lemma 21.
Moving to the term $\mathbf{P}_{\mathbf{W}_0}\mathbf{A} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) - \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t+1} - \mathbf{W}_0\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t(\tilde{\mathbf{b}}^t)^\top$, which we denote Δ_2^t, we may write

$$\mathbf{P}_{\mathbf{W}_0}\mathbf{A} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right) = \frac{1}{N}\mathbf{W}_0\rho^{-1}_{\mathbf{W}_0}(\mathbf{A}\mathbf{W}_0)^\top f^t\left(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \mathbf{X}^t\right) \tag{3.182}$$

since the function $\mathbf{A}\mathbf{W}_0, \mathbf{X}^t \to (\mathbf{A}\mathbf{W}_0)^\top f^t\left(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \mathbf{X}^t\right)$ is pseudo-Lipschitz, Lemma 21 and the induction hypothesis give

$$\left\|\frac{1}{N}\left(\mathbf{A}\mathbf{W}_0\right)^\top f^t\left(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \mathbf{X}^t\right) - \frac{1}{N}\mathbf{Z}^\top_{\mathbf{W}_0} f^t\left(\varphi\left(\mathbf{Z}_{\mathbf{W}_0}\right), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right)\right\|_F \xrightarrow[N\to\infty]{P} 0, \tag{3.183}$$

where the SE equations for iteration Eq.(3.164) yield

$$\begin{aligned}\frac{1}{N}\mathbf{Z}^\top_{\mathbf{W}_0} f^t\left(\varphi\left(\mathbf{Z}_{\mathbf{W}_0}\right), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{S}^t\right) \overset{\mathrm{P}}{\simeq} \\ \frac{1}{N}\mathbb{E}\left[\mathbf{Z}^\top_{\mathbf{W}_0} f^t\left(\varphi\left(\mathbf{Z}_{\mathbf{W}_0}\right), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right]\end{aligned} \tag{3.184}$$

An application of Lemma 17 and the chain rule gives

$$\begin{aligned}&\frac{1}{N}\mathbb{E}\left[\mathbf{Z}^\top_{\mathbf{W}_0} f^t\left(\varphi\left(\mathbf{Z}_{\mathbf{W}_0}\right), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right] = \\ &\quad\frac{1}{N}\rho_{\mathbf{W}_0}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{Z}_{\mathbf{W}_0,i,\varphi}}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right] \\ &\qquad+ \frac{1}{N}\mathbf{m}^t\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{Z}_i}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right].\end{aligned} \tag{3.185}$$

The SE equations of iteration Eq.(3.164) and the pseudo-Lipschitz assumptions on the Jacobians of the f^t then show that

$$\tilde{\mathbf{b}}^\top \overset{\mathrm{P}}{\simeq} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \frac{\partial f_i^t}{\partial \mathbf{Z}_i}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right], \tag{3.186}$$

which, combined with the definition of $\hat{\boldsymbol{\nu}}^t$, shows that

$$\frac{1}{N}\mathbb{E}\left[\mathbf{Z}^\top_{\mathbf{W}_0} f^t\left(\varphi\left(\mathbf{Z}_{\mathbf{W}_0}\right), \mathbf{Z}_{\mathbf{W}_0}\rho^{-1}_{\mathbf{W}_0}\boldsymbol{\nu}^t + \mathbf{W}_0\hat{\boldsymbol{\nu}}^t + \mathbf{Z}^t\right)\right] \overset{\mathrm{P}}{\simeq} \rho_{\mathbf{W}_0}\hat{\boldsymbol{\nu}}^{t+1} + \boldsymbol{\nu}^t\left(\tilde{\mathbf{b}}^t\right)^\top \tag{3.187}$$

combining this with Eq.(3.182) and Eq.(3.183), a straightforward application of the triangle inequality allows to show that

$$\frac{1}{\sqrt{N}}\left\|\Delta_2^t\right\|_F \xrightarrow[N\to\infty]{P} 0. \tag{3.188}$$

The only remaining term in Δ^t is $\mathbf{P}_{\mathbf{W}_0}\mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)$. Expanding the projectors, we obtain

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left\|\mathbf{P}_{\mathbf{W}_0}\mathbf{A}\mathbf{P}_{\mathbf{W}_0} f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\right\|_F = \frac{1}{\sqrt{N}}\left\|\frac{1}{N}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\mathbf{W}_0^\top\mathbf{A}\frac{1}{N}\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\mathbf{W}_0^\top f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\right\|_F \\
&\leqslant \frac{1}{\sqrt{N}}\left\|\mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\right\|_F\frac{1}{N}\left\|\mathbf{W}_0^\top\mathbf{A}\mathbf{W}_0\right\|_F\frac{1}{N}\left\|\mathbf{W}_0^\top f^t\left(\varphi(\mathbf{A}\mathbf{W}_0), \mathbf{X}^t\right)\right\|_F\left\|\rho_{\mathbf{W}_0}^{-1}\right\|_F. \qquad (3.189)
\end{aligned}$$

Lemma 21 then shows that $\frac{1}{N}\left\|\mathbf{W}_0^\top\mathbf{A}\mathbf{W}_0\right\|_F \xrightarrow[N\to\infty]{P} 0$, and the other terms are bounded w.h.p. We have now treated all the terms in Δ^t, and the triangle inequality gives

$$\frac{1}{\sqrt{N}}\left\|\Delta^t\right\|_F \xrightarrow[N\to\infty]{P} 0. \qquad (3.190)$$

which concludes the induction. Combining this with the pseudo-Lipschitz property and the SE equations to ensure all iterates have bounded scaled norms, we conclude the proof of Lemma 16. □

Application to graph-based AMP iterations : proof of Lemma 4 Consider the AMP iteration (2.15)-(2.16).To obtain the SE equations for this iteration, we follow a similar argument as the proof of Theorem 4 and embed the iteration indexed on the graph $G = (V, E)$ into a large, symmetric iteration of the form of that of Lemma 15 and Lemma 16. We may then write the $N \times N$ GOE matrix corresponding to the symmetric AMP iteration

$$\hat{\mathbf{A}} = \begin{pmatrix}
\hat{\mathbf{A}}_{\overrightarrow{e}_1} & & & & & & & & \\
& \ddots & & & & & & * & \\
& & \hat{\mathbf{A}}_{\overrightarrow{e}_l} & & & & & & \\
& & & * & \hat{\mathbf{A}}_{\overrightarrow{e}_{l+1}} & & & & \\
& & & \hat{\mathbf{A}}_{\overleftarrow{e}_{l+1}} & * & & & & \\
& & & & & \ddots & & & \\
& * & & & & & * & \hat{\mathbf{A}}_{\overrightarrow{e}_m} & \\
& & & & & & \hat{\mathbf{A}}_{\overleftarrow{e}_m} & * &
\end{pmatrix}$$

where, using the definition of each $\hat{\mathbf{A}}_{\vec{e}}$, we may write

$$
\begin{aligned}
&= \begin{pmatrix}
\mathbf{A}_{\vec{e}_1} & & & & & & & & \\
& \ddots & & & & & * & & \\
& & \mathbf{A}_{\vec{e}_l} & & & & & & \\
& & & * & \mathbf{A}_{\vec{e}_{l+1}} & & & & \\
& & & \mathbf{A}_{\overleftarrow{e}_{l+1}} & * & & & & \\
& & & & & \ddots & & & \\
& * & & & & & * & \mathbf{A}_{\vec{e}_m} & \\
& & & & & & \mathbf{A}_{\overleftarrow{e}_m} & * &
\end{pmatrix} \\
&\quad + \begin{pmatrix}
\frac{1}{N}\mathbf{v}_{\vec{e}_1}\mathbf{v}_{\vec{e}_1}^\top & & & & & & & & \\
& \ddots & & & & & & 0 & \\
& & \frac{1}{N}\mathbf{v}_{\vec{e}_l}\mathbf{v}_{\vec{e}_l}^\top & & & & & & \\
& & & 0 & 0 & & & & \\
& & & 0 & 0 & & & & \\
& & & & & \ddots & & & \\
& 0 & & & & & 0 & 0 & \\
& & & & & & 0 & 0 &
\end{pmatrix}
\end{aligned}
$$

where the second term gives the form of the matrix $\mathbf{V}_0$ from Lemma 15, i.e.

$$
\mathbf{V}_0 = \begin{pmatrix}
\mathbf{v}_{\vec{e}_1} & & & & & & \\
& \ddots & & & & 0 & \\
& & \mathbf{v}_{\vec{e}_l} & & & & \\
& & & 0 & 0 & & \\
& & & 0 & 0 & & \\
& & & & & \ddots & \\
& 0 & & & & 0 & 0 \\
& & & & & 0 & 0
\end{pmatrix} \tag{3.191}
$$

.

Furthermore, we may write the update function of the symmetric AMP iteration as

$$\tilde{f}^t\begin{pmatrix}\mathbf{x}_{\overrightarrow{e}_1} & & & & & & & & \\ & \ddots & & & & & * & & \\ & & \mathbf{x}_{\overrightarrow{e}_l} & & & & & & \\ & & & \mathbf{x}_{\overrightarrow{e}_{l+1}} & & & & & \\ & & & & \mathbf{x}_{\overleftarrow{e}_{l+1}} & & & & \\ & & & & & \ddots & & & \\ & * & & & & & \mathbf{x}_{\overrightarrow{e}_m} & & \\ & & & & & & & \mathbf{x}_{\overleftarrow{e}_m}\end{pmatrix} \tag{3.192}$$

$$= \begin{pmatrix}\tilde{f}^t_{\overrightarrow{e}_1}\left((\mathbf{x}_{\overrightarrow{e}})_{\overrightarrow{e}:\overrightarrow{e}\to\overrightarrow{e}_1}\right) & & & & & & & & \\ & \ddots & & & & & & 0 & \\ & & \tilde{f}^t_{\overrightarrow{e}_l}(\dots) & & & & & & \\ & & & 0 & \tilde{f}^t_{\overleftarrow{e}_{l+1}}(\dots) & & & & \\ & & & \tilde{f}^t_{\overrightarrow{e}_{l+1}}(\dots) & 0 & & & & \\ & & & & & \ddots & & & \\ & 0 & & & & & 0 & \tilde{f}^t_{\overleftarrow{e}_m}(\dots) & \\ & & & & & & \tilde{f}^t_{\overrightarrow{e}_m}(\dots) & 0 & \end{pmatrix}$$

$$= f^t\left(\Phi\left(\mathbf{A}\mathbf{W}_0\right), \begin{pmatrix}\mathbf{x}_{\overrightarrow{e}_1} & & & & & & & \\ & \ddots & & & & & * & \\ & & \mathbf{x}_{\overrightarrow{e}_l} & & & & & \\ & & & \mathbf{x}_{\overrightarrow{e}_{l+1}} & & & & \\ & & & & \mathbf{x}_{\overleftarrow{e}_{l+1}} & & & \\ & & & & & \ddots & & \\ & * & & & & & \mathbf{x}_{\overrightarrow{e}_m} & \\ & & & & & & & \mathbf{x}_{\overleftarrow{e}_m}\end{pmatrix}\right) \tag{3.193}$$

where

$$\mathbf{W}_0 = \begin{pmatrix}0 & & & & & & & \\ & \ddots & & & & & 0 & \\ & & 0 & & & & & \\ & & & 0 & \mathbf{w}_{\overleftarrow{e}_{l+1}} & & & \\ & & & \mathbf{w}_{\overrightarrow{e}_{l+1}} & 0 & & & \\ & & & & & \ddots & & \\ & 0 & & & & & 0 & \mathbf{w}_{\overleftarrow{e}_m} \\ & & & & & & \mathbf{w}_{\overrightarrow{e}_m} & 0\end{pmatrix} \tag{3.194}$$

and the function Φ contains the functions $\varphi_{\vec{e}}$

$$\Phi\begin{pmatrix} 0 & & & & & & & \\ & \ddots & & & & 0 & & \\ & & 0 & & & & & \\ & & & \mathbf{A}_{\vec{e}_{l+1}}\mathbf{w}_{\vec{e}_{l+1}} & & & & \\ & & & & \mathbf{A}_{\overleftarrow{e}_{l+1}}\mathbf{w}_{\overleftarrow{e}_{l+1}} & & & \\ & & & & & \ddots & & \\ & 0 & & & & & \mathbf{A}_{\vec{e}_m}\mathbf{w}_{\vec{e}_m} & \\ & & & & & & & \mathbf{A}_{\overleftarrow{e}_m}\mathbf{w}_{\overleftarrow{e}_m} \end{pmatrix} =$$

$$\begin{pmatrix} 0 & & & & & & & \\ & \ddots & & & & 0 & & \\ & & 0 & & & & & \\ & & & \varphi_{\vec{e}_{l+1}}\left(\mathbf{A}_{\vec{e}_{l+1}}\mathbf{w}_{\vec{e}_{l+1}}\right) & & & & \\ & & & & \varphi_{\overleftarrow{e}_{l+1}}\left(\mathbf{A}_{\overleftarrow{e}_{l+1}}\mathbf{w}_{\overleftarrow{e}_{l+1}}\right) & & & \\ & & & & & \ddots & & \\ & 0 & & & & & \varphi_{\vec{e}_m}\left(\mathbf{A}_{\vec{e}_m}\mathbf{w}_{\vec{e}_m}\right) & \\ & & & & & & & \varphi_{\overleftarrow{e}_m}\left(\mathbf{A}_{\overleftarrow{e}_m}\mathbf{w}_{\overleftarrow{e}_m}\right) \end{pmatrix} \tag{3.195}$$

Under the condition that the matrices $\mathbf{V}_0, \mathbf{W}_0$ and the function Φ verify the assumptions of Lemma 15 and Lemma 16, we may use those results to obtain the SE equations for the iteration Eq.(2.15)-(2.16). Evaluating the matrix products defining the parameters $\boldsymbol{\mu}^t, \boldsymbol{\nu}^t, \hat{\boldsymbol{\nu}}^t$ then leads to the SE equations of Lemma 4.

3.5 Useful definitions and probability lemmas

In this section, we compile useful definitions and lemmas that appear throughout the proof. Most of those results are finite-width matrix generalizations of those appearing in [37] and some are the same.

Proposition 4. *(Norm of matrices with Gaussian entries [288]) Let $\mathbf{Y}$ be an $M \times N$ random matrix with independent $\mathbf{N}(0,1)$ entries. Then, for any $t > 0$, we have:*

$$\mathbb{P}\left(\|\mathbf{Y}\|_F \leqslant C\left(\sqrt{M} + \sqrt{N} + t\right)\right) \geqslant 1 - 2\exp\left(-t^2\right) \tag{3.196}$$

where C is an absolute constant.

Proposition 5. *(Operator norm of GOE(N) [47])*
Consider a sequence of matrices $\mathbf{A} \sim$ GOE(N). Then $\|\mathbf{A}\|_{op} \to 2$ almost surely as $N \to \infty$.

Proposition 6. *(Gaussian Poincaré inequality [47])*
Let $\mathbf{Z} \in \mathbb{R}^N$ be a $\mathbf{N}(0, \mathbf{I}_N)$ random vector. Then for any continuous, weakly differentiable φ, there exists a constant $c \geqslant 0$ such that:

$$Var[\varphi(\mathbf{Z})] \leqslant c\mathbb{E}\left[\|\nabla\varphi(\mathbf{Z})\|_2^2\right] \tag{3.197}$$

The next result is a matrix version of Gaussian integration by parts, or Stein's lemma.

Lemma 17. *(Stein's lemma, matrix version) Let* $(\mathbf{Z}_1, \mathbf{Z}_2) \in \left(\mathbb{R}^{N\times q}\right)^2$ *be two* $\mathbf{N}(0, \boldsymbol{\kappa} \otimes \mathbf{I}_N)$ *random vectors, where* $\boldsymbol{\kappa} \in \mathbb{R}^{(2q)\times(2q)}$.

$$\boldsymbol{\kappa} = \begin{bmatrix} \boldsymbol{\kappa}_{11} & \boldsymbol{\kappa}_{12} \\ \boldsymbol{\kappa}_{12} & \boldsymbol{\kappa}_{22} \end{bmatrix} \tag{3.198}$$

Consider an almost everywhere differentiable function $f : \mathbb{R}^{N\times q} \to \mathbb{R}^{N\times q}$. *For any* $\mathbf{Z} \in \mathbb{R}^{N\times q}$ *we can write:*

$$f\left(\begin{bmatrix} \mathbf{Z}_{11}, ..., \mathbf{Z}_{1q} \\ ... \\ \mathbf{Z}_{n1}, ..., \mathbf{Z}_{nq} \end{bmatrix}\right) = \begin{bmatrix} f_1(\mathbf{Z}) \\ ... \\ f_n(\mathbf{Z}) \end{bmatrix} = \begin{bmatrix} f_1^1(\mathbf{Z}), ...f_1^q(\mathbf{Z}) \\ ... \\ f_n^1(\mathbf{Z}), ..., f_n^q(\mathbf{Z}) \end{bmatrix} \tag{3.199}$$

Then

$$\mathbb{E}\left[(\mathbf{Z}_1)^\top f(\mathbf{Z}_2)\right] = \boldsymbol{\kappa}_{1,2}\left(\sum_{k=1}^{N} \mathbb{E}\left[\frac{\partial f_k(\mathbf{Z}_2)}{\partial \mathbf{Z}_k}\right]\right)^\top \tag{3.200}$$

where $\frac{\partial f_k(\mathbf{Z}_2)}{\partial \mathbf{Z}_k} \in \mathbb{R}^{q\times q}$ *is the Jacobian containing the partial derivatives of* f_k *w.r.t. the line* $\mathbf{Z}_k \in \mathbb{R}^q$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{Z}_1)^\top f(\mathbf{Z}_2)\right]_{ij} &= \sum_{k=1}^{N} \mathbb{E}\left[((\mathbf{Z}_1)_{ki} f_{kj}(\mathbf{Z}_2)\right] \\
&= \sum_{k=1}^{N}\sum_{l=1}^{q} \mathbb{E}[\mathbf{Z}_{ki}^1 \mathbf{Z}_{kl}^2]\mathbb{E}\left[\frac{\partial f_{kj}}{\partial (\mathbf{Z}_2)_{kl}}(\mathbf{Z}_2)\right] \quad \text{since } (\mathbf{Z}_1, \mathbf{Z}_2) \sim \mathbf{N}(0, \boldsymbol{\kappa} \otimes I_N) \\
&= \sum_{l=1}^{q} (\boldsymbol{\kappa}_{12})_{il} \sum_{k=1}^{N} \mathbb{E}\left[\frac{\partial f_{kj}}{\partial (\mathbf{Z}_2)_{kl}}(\mathbf{Z}_2)\right] \\
&= \sum_{l=1}^{q} (\boldsymbol{\kappa}_{12})_{il} \left(\sum_{k=1}^{N} \mathbb{E}\left[\frac{\partial f_k(\mathbf{Z}_2)}{\partial \mathbf{Z}_k}\right]\right)_{jl} \\
&= \left(\boldsymbol{\kappa}_{12}\left(\sum_{k=1}^{N} \mathbb{E}\left[\frac{\partial f_k(\mathbf{Z}_2)}{\partial \mathbf{Z}_k}\right]\right)^\top\right)_{ij}
\end{aligned} \tag{3.201}$$

where the second step is obtained by iteratively conditioning on the entries of $\mathbf{Z}_2$ and applying one dimensional Gaussiaan integration by parts, see e.g. [288] Lemma 7.2.5. □

Definition 7 (pseudo-Lipschitz function)**.** *For* $k \in \mathbb{N}^*$ *and any* $N, m \in \mathbb{N}^*$, *a function* $\Phi : \mathbb{R}^{N\times q} \to \mathbb{R}^{m\times q}$ *is said to be* pseudo-Lipschitz of order k *if there exists a constant* L *such that for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{N\times q}$,

$$\frac{\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|_F}{\sqrt{m}} \leqslant L\left(1 + \left(\frac{\|\mathbf{x}\|_F}{\sqrt{N}}\right)^{k-1} + \left(\frac{\|\mathbf{y}\|_F}{\sqrt{N}}\right)^{k-1}\right)\frac{\|\mathbf{x} - \mathbf{y}\|_F}{\sqrt{N}} \tag{3.202}$$

A family of pseudo-Lipschitz functions is said to be *uniformly* pseudo-Lipschitz if all functions of the family are pseudo-Lipschitz with the same order k and the same constant L. We now remind useful properties of pseudo-Lipschitz functions from [37].

Lemma 18. *Let* k *be any positive integer. Consider two sequences* $f : \mathbb{R}^N \to \mathbb{R}^N, N \geqslant 1$ *and* $g : \mathbb{R}^N \to \mathbb{R}^N, N \geqslant 1$ *of uniformly pseudo-Lipschitz functions of order* k. *The sequence of functions* $\Phi_N : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}, N \geqslant 1$ *such that* $\Phi_N(\mathbf{x}, \mathbf{y}) = \langle f(\mathbf{x}), g(\mathbf{y})\rangle$ *is uniformly pseudo-Lipschitz of order* $2k$.

Lemma 19. *Let t,s and k be any three positive integers. Consider a sequence (in N) of* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s \in \mathbb{R}^N$ *such that* $\frac{1}{\sqrt{N}}\|\mathbf{x}_j\| \leqslant c_j$ *for some constant* c_j *independent of N, for* $j = 1, ..., s$ *and a sequence of order-k uniformly pseudo-Lipschitz functions* $\varphi_N : (\mathbb{R}^N)^{t+s} \to \mathbb{R}$. *The sequence of functions* $\phi_N(.) = \varphi_N(., \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s)$ *is also uniformly pseudo-Lipschitz of order k.*

Lemma 20. *Let t be any positive integer. Consider a sequence of uniformly pseudo-Lipschitz functions* $\varphi_N : (\mathbb{R}^N)^t \to \mathbb{R}$ *of order k. The sequence of functions* $\Phi_N : (\mathbb{R}^N)^t \to \mathbb{R}$ *such that* $\Phi_N(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{t)} = \mathbb{E}\left[\varphi_N(\mathbf{x}_1, ..., \mathbf{x}_{t-1}, \mathbf{x}_t + \mathbf{Z})\right]$, *in which* $\mathbf{Z} \sim \mathbf{N}(0, a\mathbf{I}_N)$ *and* $a \leqslant 0$, *is also uniformly pseudo-Lipschitz of order k.*

We now state a result on Gaussian concentration of matrix-valued pseudo-Lipschitz functions. This is an extension to the matrix case (of finite width) of Lemma C.8 from [37].

The next lemmas are matrix generalizations of the ones used in [37].

Lemma 21. *Consider a sequence of matrices* $\mathbf{A} \sim GOE(N)$ *and two sequences of non-random matrices,* $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{N \times q}$ *such that the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *verify* $\|\mathbf{U}^i\|_2 = \|\mathbf{V}^i\|_2 = \sqrt{N}$. *Under this hypothesis, define the finite quantity* $\mathbf{G} = \lim_{N\to\infty} \frac{1}{N}\mathbf{U}^\top\mathbf{U}$, *the limiting Gram matrix of the columns of* $\mathbf{U}$. *We then have:*

a) $\frac{1}{N}\mathbf{V}^\top\mathbf{A}\mathbf{U} \xrightarrow[N\to\infty]{P} 0_{q\times q}$ *and* $\frac{1}{N}\left\|\mathbf{V}^\top\mathbf{A}\mathbf{U}\right\|_F \xrightarrow[N\to\infty]{P} 0$.

b) Let $\mathbf{P} \in \mathbb{R}^{N\times N}$ *be a sequence of non-random projection matrices such that there exists a constant t that satisfies, for all N, k=rank(*$\mathbf{P}$*)*$\leqslant t$. *Then* $\frac{1}{N}\|\mathbf{P}\mathbf{A}\mathbf{U}\|_F^2 \xrightarrow[N\to\infty]{P} 0$.

c) There exists a sequence of random matrices $\mathbf{Z} \in \mathbb{R}^{N\times q}$, *such that* $\frac{1}{N}\|\mathbf{A}\mathbf{U} - \mathbf{Z}\|_F^2 \xrightarrow[N\to\infty]{P} 0$ *where* $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{G} \otimes \mathbf{I}_N)$.

d) $\frac{1}{N}(\mathbf{A}\mathbf{U})^\top\mathbf{A}\mathbf{U} \xrightarrow[N\to\infty]{P} \mathbf{G}$.

Proof. In this proof, the i-th line of a given matrix $\mathbf{Z}$ is denoted $\mathbf{Z}_i$ and its j-th column $\mathbf{Z}^j$.

a) For any $1 \leqslant i, j \leqslant q$, the i-th element of the j-th column verifies:

$$\begin{aligned}\frac{1}{N}(\mathbf{V}^\top\mathbf{A}\mathbf{U})_i^j &= \frac{1}{N}(\mathbf{V}^i)^\top\mathbf{A}\mathbf{U}^j \\ &= \frac{1}{N}(\mathbf{V}^i)^\top\mathbf{H}\mathbf{U}^j + \frac{1}{N}(\mathbf{V}^i)^\top\mathbf{H}^\top\mathbf{U}^j \end{aligned} \tag{3.203}$$

where $\mathbf{H}$ is a matrix with i.i.d. $\mathbf{N}(0, \frac{1}{2N})$ elements. The random variable $\frac{1}{N}(\mathbf{V}^i)^\top\mathbf{H}\mathbf{U}^j$ is centered Gaussian with variance

$$\frac{1}{N^2}\sum_{k,l=1}^{N}(\mathbf{V}_k^i)^2(\mathbf{U}_l^j)^2\frac{1}{2N} = \frac{\|\mathbf{V}^i\|_2^2\|\mathbf{U}^j\|_2^2}{2N^3} = \frac{1}{2N} \to 0 \tag{3.204}$$

which shows that $\frac{1}{N}(\mathbf{V}^i)^\top\mathbf{H}\mathbf{U}^j$ converges in probability to zero. A similar argument shows that $\frac{1}{N}(\mathbf{V}^i)^\top\mathbf{H}^\top\mathbf{U}^j$ also converges in probability to zero. The union bound then immediately gives that $\frac{1}{N}(\mathbf{V}^\top\mathbf{A}\mathbf{U})_i^j \xrightarrow[N\to\infty]{P} 0$. Thus each element of the finite size $q \times q$ matrix $\frac{1}{N}\mathbf{V}^\top\mathbf{A}\mathbf{U}$ goes to zero. Since q is finite, the union bound then gives the desired result on the Frobenius norm.

b) For any $1 \leqslant i \leqslant q$:

$$\frac{1}{N}(\mathbf{PAU})^i = \frac{1}{N}(\mathbf{PAU}^i) \tag{3.205}$$

Now let $\mathbf{v}_1, ..., \mathbf{v}_k$ be an orthogonal basis of the image of $\mathbf{P}$, such that $\|\mathbf{v}_1\| = ... = \|\mathbf{v}_k\| = \sqrt{N}$, and $\mathbf{V} \in \mathbb{R}^{N \times t}$ the matrix of concatenated $\mathbf{v}$. Note that k can depend on N, but k is uniformly bounded by t. Then, using point (a) and the fact that q and k are finite for all N:

$$\frac{1}{N}\|\mathbf{PAU}\|_F^2 = \frac{1}{N}\left\|\mathbf{V}^\top \mathbf{AU}\right\| \xrightarrow[N\to\infty]{P} 0 \tag{3.206}$$

This proves point (b).

c) The matrix $\mathbf{AU}$ is a $\mathbb{R}^{N \times q}$ correlated Gaussian matrix. For any two columns $\mathbf{U}^l, \mathbf{U}^m$, the vector $(\mathbf{AU}^l, \mathbf{AU}^m)$ is a Gaussian vector with zero mean, whose covariance matrix has elements:

$$\begin{aligned}
\mathbb{E}\left[\left(\mathbf{AU}^l\right)\left(\mathbf{AU}^m\right)^\top\right]_i^j &= \mathbb{E}\left[\left(\mathbf{AU}^l\right)_i \left(\mathbf{AU}^m\right)_j\right] \\
&= \mathbb{E}\left[\sum_{k=1}^N \mathbf{A}_i^k \mathbf{U}_k^l \sum_{k'=1}^N \mathbf{A}_j^{k'} \mathbf{U}_{k'}^m\right] \\
&= \mathbb{E}\Big[\sum_{k,k'} \mathbf{H}_i^k \mathbf{H}_j^{k'} \mathbf{U}_k^l \mathbf{U}_{k'}^m + \mathbf{H}_i^k \mathbf{H}_{k'}^j \mathbf{U}_k^l \mathbf{U}_{k'}^m + \mathbf{H}_k^i \mathbf{H}_j^{k'} \mathbf{U}_k^l \mathbf{U}_{k'}^m + \mathbf{H}_k^i \mathbf{H}_{k'}^j \mathbf{U}_k^l \mathbf{U}_{k'}^m\Big] \\
&= \frac{1}{N}\left(\delta_{ij} \sum_k \mathbf{U}_k^l \mathbf{U}_k^m + \mathbf{U}_i^l \mathbf{U}_j^m\right)
\end{aligned} \tag{3.207}$$

which gives the block

$$\mathbb{E}\left[\left(\mathbf{AU}^l\right)\left(\mathbf{AU}^m\right)^\top\right] = \frac{1}{N}(\mathbf{U}^l)^\top \mathbf{U}^m \mathbf{I}_N + \frac{1}{N}\mathbf{U}^l(\mathbf{U}^m)^\top \tag{3.208}$$

and the covariance matrix

$$\Sigma = \begin{bmatrix} \mathbf{I}_N + \frac{1}{N}\mathbf{U}^l(\mathbf{U}^l)^\top & \frac{(\mathbf{U}^l)^\top \mathbf{U}^m}{N}\mathbf{I}_N + \frac{1}{N}\mathbf{U}^l(\mathbf{U}^m)^\top \\ \frac{(\mathbf{U}^l)^\top \mathbf{U}^m}{N}\mathbf{I}_N + \frac{1}{N}\mathbf{U}^m(\mathbf{U}^l)^\top & \mathbf{I}_N + \frac{1}{N}\mathbf{U}^m(\mathbf{U}^m)^\top \end{bmatrix} \tag{3.209}$$

and in turn the following covariance matrix for the joint law of the q vectors $\mathbf{AU}^1, ..., \mathbf{AU}^q$.

$$\begin{aligned}
\Sigma &= \frac{1}{N}\mathbf{U}^\top \mathbf{U} \otimes \mathbf{I}_N + \frac{1}{N}\begin{bmatrix} \mathbf{U}^1(\mathbf{U}^1)^\top & ... & ... & ... & \mathbf{U}^1(\mathbf{U}^q)^\top \\ ... & ... & ... & ... & ... \\ ... & \mathbf{U}^i(\mathbf{U}^{i-1})^\top & \mathbf{U}^i(\mathbf{U}^i)^\top & \mathbf{U}^i(\mathbf{U}^{i+1})^\top & ... \\ ... & ... & ... & ... & ... \\ \mathbf{U}^q(\mathbf{U}^1)^\top & ... & ... & ... & \mathbf{U}^q(\mathbf{U}^q)^\top \end{bmatrix} \\
&= \frac{1}{N}\mathbf{U}^\top \mathbf{U} \otimes \mathbf{I}_N + \frac{1}{N}\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top
\end{aligned} \tag{3.210}$$

where $\tilde{\mathbf{U}} \in \mathbb{R}^{Nq}$ is the vector of vertically concatenated columns of $\mathbf{U}$. Now consider two independent $\mathbf{N}(0, \mathbf{I}_{Nq})$ vectors $\tilde{\mathbf{Z}}^1, \tilde{\mathbf{Z}}^2$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{Nq}$ the vector of vertically concatenated columns of $\mathbf{AU}$. We can write that the quantity:

$$\frac{\left\|\tilde{\mathbf{V}} - \left(\frac{1}{N}\mathbf{U}^\top \mathbf{U} \otimes \mathbf{I}_N\right)^{1/2} \tilde{\mathbf{Z}}^1\right\|_2}{\sqrt{N}} \tag{3.211}$$

is distributed as

$$\frac{\left\|(\frac{1}{N}\mathbf{U}^\top\mathbf{U}\otimes\mathbf{I}_N)^{1/2}\tilde{\mathbf{Z}}^1+(\frac{1}{N}\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top)^{1/2}\tilde{\mathbf{Z}}^2-(\frac{1}{N}\mathbf{U}^\top\mathbf{U}\otimes\mathbf{I}_N)^{1/2}\tilde{\mathbf{Z}}^1\right\|_2}{\sqrt{N}}=\frac{1}{N\sqrt{N}}\left\|\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\tilde{\mathbf{Z}}^2\right\|_2$$
$$=\frac{\sqrt{q}}{N}\left|\tilde{\mathbf{U}}^\top\tilde{\mathbf{Z}}^2\right|\xrightarrow[N\to\infty]{P}0 \tag{3.212}$$

where the last convergence follows from the fact that $\frac{1}{N}\tilde{\mathbf{U}}^\top\tilde{\mathbf{Z}}^2$ is a centered Gaussian random variable with variance $\left\|\tilde{\mathbf{U}}\right\|_2^2/N^2=q/N$, where q is kept finite. This concludes the proof of point (c).

d) The function $\Phi:\mathbb{R}^{N\times q}\to\mathbb{R},\mathbf{X}\to\frac{1}{N}\mathbf{X}^\top\mathbf{X}$ is pseudo-Lipschitz of order 2. A straightforward calculation shows that, for any $\mathbf{Z}\sim\mathbf{N}(0,\mathbf{G}\otimes\mathbf{I}_N)$, we have $\mathbb{E}[\phi(\mathbf{Z})]=\mathbf{G}$. Then :

$$\mathbb{P}\left(\|\Phi(\mathbf{AU})-\mathbb{E}[\Phi(\mathbf{Z})]\|_F\geqslant\epsilon\right)\leqslant\mathbb{P}\left(\|\Phi(A\mathbf{U})-\Phi(\mathbf{Z})\|_F\geqslant\epsilon\right)+\mathbb{P}\left(\|\Phi(\mathbf{Z})-\mathbb{E}[\Phi(\mathbf{Z})]\|_F\geqslant\epsilon\right) \tag{3.213}$$

the second term on the right-hand side vanishes as $N\to\infty$ using the Gaussian concentration of matrix-valued pseudo-Lipschitz functions Lemma 1, and the first term vanishes using the definition of pseudo-Lipschitz function and the statement (c) proven above. This concludes the proof of statement (d).

□

Chapter 4

Multi-layer State Evolution Under Random Convolutional Design

The results presented in this chapter were published in [70].

Motivated by the multilayer iteration -MLAMP- proposed in [188], we seek further models of deep neural networks with random weights for which marginals can be computed using AMP iterations, and for which SE equations can be made rigorous using the framework proposed in Chapters 2 and 3. We show that the MLAMP iteration corresponding to multilayer neural networks with random convolutional matrices, which we define in 4.1, admit rigorous SE equations that exactly match those of the usual case with dense matrices, up to a rescaling. Further discussions on the litterature of generative models in deep learning, computational benefits of random convolutional matrices over dense ones and future directions can be found in the original paper [70].

In a typical signal recovery problem, one seeks to recover a data signal x_0 given access to measurements $y_0 = G_\theta(x_0)$, where the parameters θ of the signal model are known. In many problems, it is natural to view the measurement generation process as a composition of simple forward operators, or 'layers.' In this work, we are concerned with multi-layer signal models of the form

$$G_\theta(h) = \phi^{(1)}(W^{(1)}\phi^{(2)}(W^{(2)}\dots\phi^{(L)}(W^{(L)}h))). \tag{4.1}$$

where $W^{(l)} \in \mathbb{R}^{n_{l-1}\times n_l}$ are linear sensing matrices and where $\phi^{(l)}(z)$ are separable, possibly non-linear channel functions. In the $L = 1$ case, this signal model naturally generalizes problems such as phase retrieval $\phi(z) = |z|$ or compressive sensing $\phi(z) = z$, and for multi-layer models $L > 1$, $G_\theta(h)$ may be viewed as a deep neural network.

Recently, convolutional Generative Neural Networks (GNNs) have shown promise as generalizations of sparsity priors for a variety of signal processing applications [44]. Motivated by this success, we take interest in a variant of the recovery problem (4.1) in which some of the sensing matrices $W^{(l)}$ may be *multi-channel convolutional* (MCC) matrices, having a certain block-sparse circulant structure which captures the convolutional layers used by many modern generative neural network architectures [142, 143].

In this work, we develop an asymptotic analysis of the performance of an *Approximate Message Passing* (AMP) algorithm [83] for recovery from multichannel convolutional signal models. This family of algorithms originates in statistical physics [195, 300] and allows to compute the marginals of an elaborate posterior distribution defined by an inference problem involving dense random

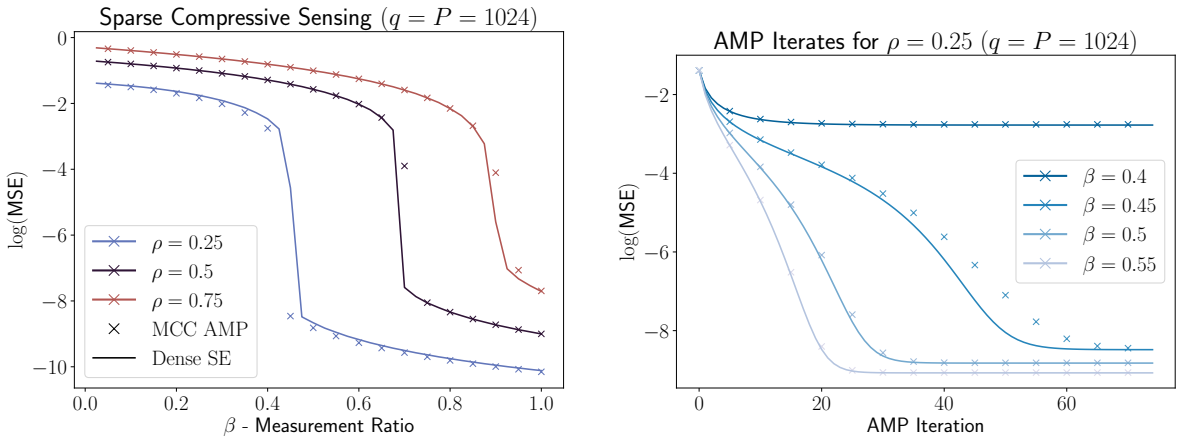


Figure 4.1: Agreement between the performance of the AMP algorithm run with random multi-channel convolutional matrices and its state evolution as proven in this paper. (**left**) Compressive sensing $y_0 = Wx_0 + \zeta$ for noise $\zeta_i \sim \mathcal{N}(0, 10^{-4})$ and signal prior $x_0 \sim \rho\mathcal{N}(0,1) + (1-\rho)\delta(x)$, where $W \in \mathbb{R}^{Dq \times Pq}$ has varying aspect ratio $\beta = D/P$. Crosses correspond to AMP evaluations for $W \sim \mathrm{MCC}(D, P, q, k)$ according to Definition 9, averaged over 10 independent trials. Lines show the state evolution predictions when $W_{ij} \sim \mathcal{N}(0, 1/Pq)$. The system size is $P = 1024$, $q = 1024$, $k = 3$, where β and $D = \beta P$ vary. While our theorem treats the limit $P, D \to \infty$, $q, k = O(1)$, we observe strong empirical agreement even when $q \sim P$. (**right**) AMP iterates at $\rho = 0.25$ and β near the recovery transition.

matrices. A number of AMP iterations have been proposed for various inference problems, such as compressed sensing [83], low-rank matrix recovery [241] or generalized linear modeling [240]. More recently, composite AMP iterations (ML-AMP) have been proposed to study multilayer inference problems [188, 13]. Here we consider the ML-AMP proposed in [188] to compute marginals of a multilayer generalized linear model, however the usual dense Gaussian matrices will be replaced by random convolutional ones. A major benefit of AMP lies in the fact that the asymptotic distribution of their iterates can be exactly determined by a low-dimensional recursion: the state evolution equations. This enables to obtain precise theoretical results for the reconstruction performance of the proposed algorithm. Another benefit of such iterations is their low computational complexity, as they only involve matrix-multiplication and, in the separable case, pointwise non-linearities.

Previous works on AMP suggest that the state evolution is not readily applicable to our setting because its derivation requires strong independence assumptions on the coordinates of the $\{W^{(l)}\}$ which are violated by structured multi-channel convolution matrices. Despite this, we use AMP for our setting and rigorously prove its state evolution. Our main contributions are:

1. We rigorously prove state evolution equations for models of the form (4.1), where weights are allowed to be either i.i.d. Gaussian or random structured MCC matrices, as in Definition 9.

2. For separable channel functions $\phi^{(l)}$ and separable signal priors, we show that the original ML-AMP of [188] used with dense Gaussian matrices or random convolutional ones admits the same state evolution equations, up to a rescaling. Multi-layer MCC signal models can therefore simulate dense signal models while making use of fast structured matrix operations for convolutions.

3. The core of our proof shows how an AMP iteration involving random convolutional matrices may be reduced to another one with dense Gaussian matrices. We first show that random convolutional matrices are equivalent, through permutation matrices, to dense Gaussian ones with a (sparse) block-circulant structure. We then show how the block-circulant structure can be embedded in a new, matrix-valued, multilayer AMP with dense Gaussian matrices, the state evolution equations of which are proven using the results of [110], with techniques involving spatially coupled matrices [154, 135].

4. We validate our theory numerically and observe close agreement between convolutional AMP iterations and its state evolution predictions, as shown in Figure 4.1 and in Section 4.3. Our code can be used as a general purpose library to build compositional models and evaluate AMP and its state evolution. We make this code publically available on Github.

Further discussion on related works can be found in the original paper [70].

4.1 Definition of the problem

4.1.1 Multi-channel Convolutional Matrices

We consider block structured signal vectors $x \in \mathbb{R}^{Pq}$ of the form $x = [x^{(i)}]_{i=1}^{P}$, and we refer to the blocks $x^{(i)} \in \mathbb{R}^q$ as 'channels.' For any vector of dimension d, we denote by $\mathcal{P}_d \in \mathbb{R}^{d\times d}$ the cyclic coordinate permutation matrix of order d, whose coordinates are $\langle e_i, \mathcal{P}_d e_j\rangle = \mathbf{1}[i = j+1]$. For a block-structured vector $x \in \mathbb{R}^{Pq}$, we denote by $\mathcal{P}_{P,q} \in \mathbb{R}^{Pq\times Pq}$ the block cyclic permutation matrix satisfying $(\mathcal{P}_{P,q}x)^{(i)} = x^{(i+1)}$ for $1 \leqslant i < P$, and $(\mathcal{P}_{P,q}x)^{(P)} = x^{(1)}$. Similarly, we denote by $\mathcal{S}_{i,j} \in \mathbb{R}^{Pq\times Pq}$ the swap permutation matrix which exchanges blocks i, j: $[\mathcal{S}_{i,j}x]^{(i)} = x^{(j)}$, $[\mathcal{S}_{i,j}x]^{(j)} = x^{(i)}$, and $[\mathcal{S}_{i,j}x]^{(k)} = x^{(k)}$ for $k \neq i, j$. Last, given a vector $\omega \in \mathbb{R}^k$ for $k \leqslant q$, denote by $\texttt{Zero-Pad}_{q,k}(\omega)$ the vector whose first k coordinates are ω, and whose other coordinates are zero.

$$\texttt{Zero-Pad}_{q,k}(\omega) = \begin{bmatrix}\omega_1 & \omega_2 & \dots & \omega_k & 0 & \dots & 0\end{bmatrix} \in \mathbb{R}^q.$$

We define the following ensemble for random multi-channel convolution matrices.

Definition 8 (Gaussian i.i.d. Convolution)**.** *Let $q \geqslant k$ be integers. The convolutional ensemble $\mathcal{C}(q,k)$ contains random circulant matrices $C \in \mathbb{R}^{q\times q}$ whose first row is given by $C_1 = \texttt{Zero-pad}_{q,k}[\omega]$ where $\omega \in \mathbb{R}^k$ has i.i.d. Gaussian coordinates $\omega_i \sim \mathcal{N}(0, 1/k)$. The remaining rows C_i are determined by circulant structure, ie. $C_i = \mathcal{P}_q^{i-1}\texttt{Zero-pad}_{q,k}[\omega]$.*

Random multi-channel convolutions are block-dense matrices with independent $\mathcal{C}(q,k)$ blocks.

Definition 9 (Multi-channel Gaussian i.i.d. Convolution)**.** *Let $D, P \geqslant 1$ and $q \geqslant k \geqslant 1$ be integers. The random multi-channel convolution ensemble $\mathcal{M}(D, P, k, q)$ contains random block matrices $M \in \mathbb{R}^{Dq\times Pq}$ of the form*

$$M = \frac{1}{\sqrt{P}}\begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,P} \\ C_{2,1} & \ddots & & \vdots \\ \vdots & & & \\ C_{D,1} & \dots & & C_{D,P}\end{bmatrix}$$

where each $C_{i,j} \sim \mathcal{C}(q,k)$ is sampled independently.

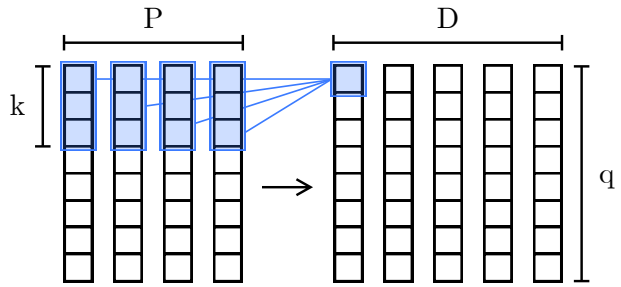


Figure 4.2: MCC matrices operate on Pq dimensional input data, composed of q-dimensional signals for each of P separate channels. The i-th output channel is a linear combination of convolutional features extracted from input channels, where k is the convolutional filter size: $y^{(i)} = \sum_{j=1\dots P} C_{ij} x^{(j)}$. Blue boxes show linear dependencies between signal coordinates.

Fig. 4.2 gives a graphical explanation of the link between these matrices and the convolutional layers. The parameter P (D) is the number of input (output) channels, q is the dimension of the input and k the filter size.

4.1.2 Multi-layer AMP

In this section, we define a class of probabilistic graphical models (PGMs) that captures the inference problems of interest, and we state the Multi-layer Approximate Message Passing (ML-AMP) [188] iterations, which can be used for inference on these PGMs. We consider the following signal model.

Definition 10 (Multi-layer Signal Model)**.** *Let $\{W^{(l)}\}_{1\leqslant l\leqslant L}$ be matrices of dimension $W^{(l)} \in \mathbb{R}^{n_{l-1}\times n_l}$. Let $\{\phi_\zeta^{(l)}(z)\}_{1\leqslant l\leqslant L}$ be scalar channel functions $\phi_\zeta^{(l)} : \mathbb{R} \to \mathbb{R}$ for which z is the estimation quantity and ζ represents channel noise. We write $\phi_\zeta^{(l)}(z)$ for vectors $z \in \mathbb{R}^{n_{l-1}}$ to indicate the coordinatewise application of $\phi^{(l)}$. The multi-layer GLM signal model is given by*

$$y = \phi_\zeta^{(1)}(W^{(1)}\phi_\zeta^{(2)}(W^{(2)}(\dots\phi_\zeta^{(L)}W^{(L)}x))).$$

We assume $x \in \mathbb{R}^{n_L}$ follows a known separable prior, $x_i \sim P_X(x)$ i.i.d., and that $\zeta \sim \mathcal{N}(0,1)$.

The full estimation quantities of the model are the coordinates of the vectors $\{h^{(l)}\}_{1\leqslant l\leqslant L}$, $\{z^{(l)}\}_{1\leqslant l\leqslant L}$, which are related by

$$\begin{aligned} y_\mu &= \phi_\zeta^{(1)}(z^{(1)}) & z_\mu^{(1)} &= \sum_i W_{\mu i}^{(1)} h_i^{(1)}\,, \\ h_i^{(1)} &= \phi_\zeta^{(2)}(z^{(2)}) & z_\mu^{(2)} &= \sum_i W_{\mu i}^{(2)} h_i^{(2)}\,, \\ &\vdots \\ h_i^{(L-1)} &= \phi_\zeta^{(L)}(z^{(L)}) & z_\mu^{(L)} &= \sum_i W_{\mu i}^{(L)} x_i \end{aligned} \tag{4.2}$$

and the corresponding conditional probabilities, which define the factor nodes of the underlying PGM, are given by

$$P^{(l)}(h \mid z) = \int d\zeta\, e^{-\frac{1}{2}\zeta^2}\delta(h - \phi_\zeta(z))\,.$$

To compute the posterior marginals, ML-AMP iteratively updates the parameters of independent 1D Gaussian approximations to each marginal. Each coordinate $h_i^{(l)}(t)$ has corresponding parameters $\{A_i^{(l)}(t), B_i^{(l)}(t)\}$ and each $z_\mu^{(l)}(t)$ has corresponding $\{V_\mu^{(l)}(t), \omega_\mu^{(l)}(t)\}$, where $t \geqslant 1$ indexes the ML-AMP iterations. The recursive relationship between these parameters is defined in terms of scalar *denoising functions*, $\hat{h}^{(l)}$ and $g^{(l)}$, which compute posterior averages of the estimation quantities given their prior parameters.

In general, these denoising functions can be chosen (up to regularity assumptions) to adjust ML-AMP's performance in applied settings, such as in [192], and in these cases the denoisers may be nonseparable vector valued functions. However, in the separable, Bayes-optimal regime where $P_x(x)$ and $P^{(l)}(h \mid z)$ are known, the optimal denoisers are given by,

$$\begin{aligned}
\hat{h}_i^{(l)}(t+1) &:= \partial_B \log \mathcal{Z}^{(l+1)}(A_i^{(l)}, B_i^{(l)}, V_i^{(l+1)}, \omega_i^{(l+1)}) \\
\sigma_i^{(l)}(t+1) &:= \partial_B \hat{h}_i^{(l)}(t+1) \\
g_\mu^{(l)}(t) &:= \partial_\omega \log \mathcal{Z}^{(l)}(A_\mu^{(l-1)}, B_\mu^{(l-1)}, V_\mu^{(l)}, \omega_\mu^{(l)}) \\
\eta_\mu^{(l)}(t) &:= \partial_\omega g_\mu^{(l)}(t) \\
\mathcal{Z}^{(l)}(A, B, V, \omega) &:= \frac{1}{\sqrt{2\pi V}} \int P^{(l)}(h \mid z) \exp\left(Bh - \frac{1}{2}Ah^2 - \frac{(z-\omega)^2}{2V} \right) dh\, dz
\end{aligned} \tag{4.3}$$

where $2 \leqslant L \leqslant L-1$, $t \geqslant 2$ and the prior parameters on the right hand side are taken at iteration $t \geqslant 2$. The corresponding ML-AMP iterations are given by,

$$\begin{aligned}
V_\mu^{(l)}(t) &= \sum_i [W_{\mu i}^{(l)}]^2 \, \sigma_i^{(l)}(t) & \omega_\mu^{(l)}(t) &= \sum_i W_{\mu i}^{(l)} \, \hat{h}_i^{(l)}(t) - V_\mu^{(l)}(t) \, g_\mu^{(l)}(t-1) \\
A_i^{(l)}(t) &= -\sum_\mu [W_{\mu i}^{(l)}]^2 \, \eta_\mu^{(l)}(t) & B_i^{(l)}(t) &= \sum_\mu W_{\mu i}^{(l)} \, g_\mu^{(l)}(t) + A_i^{(l)}(t) \hat{h}_i^{(l)}(t).
\end{aligned} \tag{4.4}$$

For the boundary cases $t = 1$, $l = 1$, and $l = L$, the iterations (4.3), (4.4) are modified as follows.

1. At $t = 1$, we initialize $B_i^{(l)} \sim P_{B_0}^{(l)}$ and $\omega_\mu^{(l)} \sim P_{\omega_0}^{(l)}$, where $P_{B_0}^{(l)}$, $P_{\omega 0}^{(l)}$ are the distributions of the signal model parameters (4.2) when $x_i \sim P_X$. We take $(A_i^{(l)})^{-1} = \mathrm{Var}(B_i^{(l)})$ and $V_\mu^{(l)} = \mathrm{Var}(\omega_\mu^{(l)})$.

2. At $l = 1$, the denoiser $g_\mu^{(1)}(t) = \partial_\omega \log \mathcal{Z}^{(1)}(y, V_\mu^{(1)}, \omega_\mu^{(1)})$, where

$$\mathcal{Z}^{(1)}(y, V_\mu^{(1)}, \omega_\mu^{(1)}) = \frac{1}{\sqrt{2\pi V}} \int P^{(1)}(y \mid z) \exp\left(-\frac{(z-\omega_\mu^{(1)})^2}{2V_\mu^{(1)}} \right) dz.$$

3. At $l = L$, the denoiser $\hat{h}^{(L)}(t) = \partial_B \log \mathcal{Z}^{(L)}(A_i^{(L)}, B_i^{(L)})$, where

$$\mathcal{Z}^{(L)}(A_i^{(L)}, B_i^{(L)}) = \int P_X(h) \exp\left(B_\mu^{(L)} h - \frac{1}{2} A_\mu^{(L)} h^2 \right) dh\,.$$

4.2 Main result

We now state our main technical result, starting with the set of required assumptions.

(A1) for any $1 \leqslant l \leqslant L$, the function ϕ^l is continuous and there exists a polynomial $b^{(l)}$ of finite order such that, for any $x \in \mathbb{R}$, $|\phi^{(l)}(x)| \leqslant |b^{(l)}(x)|$

(A2) for any $1 \leqslant l \leqslant L$, the matrix $\mathbf{W}^{(l)}$ is sampled from the ensemble $\mathcal{M}(D^l, P^l, k^l, q^l)$ where $P^l q^l = D^{l-1} q^{l-1}$

(A3) the iteration 4.4 is initialized with a random vector independent of the mixing matrices verifying $\frac{1}{N}\|\mathbf{h}_0\|_2^2 < +\infty$ almost surely

(A4) for any $1 \leqslant l \leqslant L$, $D_l, P_l \to \infty$ with constant ratio $\beta_l = D_l/P_l$, with finite q_l.

Under these assumptions, we may define the following *state evolution* recursion

Definition 11 (State Evolution)**.** *Consider the following recursion,*

$$\hat{m}^{(l)}(t) = -\beta^{(l)} \mathbb{E}^{(l)}[\partial_\omega g(\hat{m}^{(l-1)}, \hat{m}b, \tau_1 - m^{(l)}, h)] \tag{4.5}$$

$$m^{(l-1)}(t+1) = \mathbb{E}^{(l)}[h\, \hat{h}^{(l-1)}(\hat{m}^{(l-1)}, \hat{m}b, \tau_1 - m^{(l)}, h)], \tag{4.6}$$

where $\tau^{(l)}$ is the second moment of $P_{B_0}^{(l)}$, where the right hand side parameters are taken at time t, and the expectations $\mathbb{E}^{(l)}$ are taken with respect to

$$P^{(l)}(w, z, h, b) = P_{out}^{(l)}(h \mid z)\, \mathcal{N}(z; w, \tau^{(l)} - m^{(l)})\, \mathcal{N}(w; 0, m^{(l)})\, \mathcal{N}(b; \hat{m}^{(l-1)} h, \hat{m}^{(l-1)}).$$

At $t = 1$, the state evolution is initialized at $\kappa^{(l)} = 0$ and $(\hat{\kappa}^{(l)})^{-1} = \tau^{(l)}$. At the boundaries $l = 1, L$, the expectations are modified analogously to the ML-AMP iterations as described by [188]. We then have the following asymptotic characterization of the iterates from the convolutional ML-AMP algorithm

Theorem 7. *Under the set of assumptions (A1)-(A4), for any sequences of uniformly pseudo-Lipschitz functions ψ_1^N, ψ_2^N of order k, for any $1 \leqslant l \leqslant L$ and any $t \in \mathbb{N}$, the following holds*

$$\frac{1}{D_l q_l} \sum_{i=1}^{D_l q_l} \psi_1(\omega_i^{(l)}(t)) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\psi_1\left(Z^l(t)\right)\right] \tag{4.7}$$

$$\frac{1}{P_l q_l} \sum_{i=1}^{P_l q_l} \psi_2(B_i^{(l)}(t)) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\psi_2\left(\hat{Z}^l(t)\right)\right] \tag{4.8}$$

where $Z^l(t) \sim \mathcal{N}(0, \kappa^l(t))$, $\hat{Z}^l(t) \sim \mathcal{N}(0, \hat{\kappa}^l(t))$ are independent random variables.

4.2.1 Proof Sketch

The proof of Theorem 7, which is given in Appendix 5.1, has two key steps. First, we construct permutation matrices $U, \tilde{U}$ such that for $W \sim \mathrm{MCC}(D, P, q, k)$, the matrix $\tilde{W} = UW\tilde{U}^T$ is a block matrix whose blocks either have i.i.d. Gaussian elements or are zero valued, and has a block-circulant structure. The effect of the permutation is that entries of $\tilde{W}$ which are correlated due to circulant structure of W are relocated to different blocks. Once these permutation matrices are defined, we define a new, matrix-valued AMP iteration involving the dense Gaussian matrices obtained from the permutations, and whose non-linearities account for the block-circulant structures and the permutation matrices. The state evolution of this new iteration is proven using the results of [110].

This provides an explicit example of how the aforementioned results can be used to obtain rigorous, non Bayes-optimal SE equations on a composite AMP iteration. The separability assumption is key in showing that the AMP iterates obtained with the convolutional matrices can be *exactly* embedded in a larger one. Note that this is a stronger result than proving SE equations for an algorithm that computes marginals of a random convolutional posterior: we show the SE equations are the same as in the dense case. We finally invoke the Nishimori conditions, see e.g. [154], to simplify the generic, non Bayes-optimal SE equations to the Bayes-optimal ones.

The idea of embedding a non-separable effect such as a block-circulant structure or different variances in a mixing matrix is the core idea in the proofs of SE equations for spatially coupled systems, notably as done in [135, 80]. We note that in the numerical experiments shown at Figure 4.1, the parameter q, considered finite in the proof, is actually comparable to the number of channel, considered to be extensive. Empirically we observe that this does not hinder the validity of the result, something that was also observed in the spatial coupling literature, e.g. [154], where large number of different blocks in spatially coupled matrices were considered, with convincing numerical agreement.

The existence of permutations matrices verifying the property described above is formalized in the following lemma:

Lemma 22 (Permutation Lemma)**.** *Let $W \sim \mathcal{M}(D, P, k, q)$ be a multi-channel convolution matrix. There exist row and column permutation matrices $U \in \mathbb{R}^{Dq \times Dq}$, $\tilde{U} \in \mathbb{R}^{Pq \times Pq}$ such that $\tilde{W} = UW\tilde{U}^T$ is a block-convolutional matrix with dense, Gaussian i.i.d. blocks. That is,*

$$\tilde{W} = \frac{1}{\sqrt{k}} \begin{bmatrix} A^{(1)} & A^{(2)} & \dots & A^{(k)} & & & \\ & A^{(1)} & A^{(2)} & \dots & A^{(k)} & & \\ \vdots & & A^{(2)} & \dots & A^{(k)} & & \\ & & & & \ddots & \vdots & \\ A^{(2)} & A^{(3)} & \dots & A^{(k)} & & & A^{(1)} \end{bmatrix}$$

where each $A^{(s)} \in \mathbb{R}^{D,P}$, $1 \leqslant s \leqslant k$ has i.i.d. $\mathcal{N}(0, 1/P)$ coordinates.

Proof. Consider the elements of the matrix M which are non-zero and sampled i.i.d. as opposed to exact copies of other variables. They are positioned on the first line of each block of size $q \times q$, and thus the indexing for their lines and columns can be written as $M_{aq+1,bq+c}$ where a, b, c are integers such that $0 \leqslant a \leqslant D-1$, $0 \leqslant b \leqslant P-1$ and $1 \leqslant c \leqslant k$. The integers a, b describe the position of the $q \times q$ block the variable is in, and c describes, for each block, the position in the initial random Gaussian vector of size k that is zero-padded and circulated to generate the block. The goal is to find the mapping that groups these variables into k dense blocks of extensive size $D \times P$. To do so, one can use the following bijection $\tilde{M}_{\gamma,\alpha P+\beta} = M_{aq+1,bq+c}$ where $\gamma = a+1$, $\alpha = c-1$ and $\beta = b+1$. By doing this, c becomes the block index and a, b become the position in the dense block. This mapping can be represented by left and right permutation matrices which also prescribe the permutation for the rest of the elements of M. A graphical sketch of this coordinate permutation is shown in Figure 4.3. □

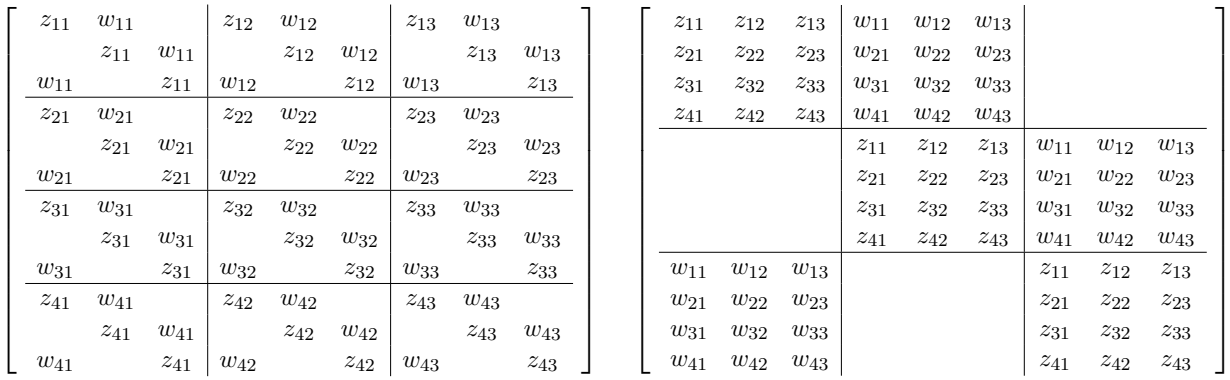


Figure 4.3: A sketch of the permutation lemma applied to matrix $W \sim \mathrm{MCC}(4,3,3,2)$. Left: W before permutation. Right: after permutation, $UW\tilde{U}^T$.

.

4.3 Numerical Experiments

In this section, we compare state evolution predictions from Theorem 7 with a numerical implementation of the ML-AMP algorithm described in Section 4.1.2.

Our first experiment, shown in Figure 4.1, is a noisy compressive sensing task under a sparsity prior $P_X(x) = \rho\mathcal{N}(x;0,1)+(1-\rho)\delta(x)$, where ρ is the expected fraction of nonzero components of x_0. Measuremements are generated $y_0 = Wx_0+\eta$ for noise $\eta \sim \mathcal{N}(0,10^{-4})$, where $W \sim \mathrm{MCC}(D,P,q,k)$. We show recovery performance at sparsity levels $\rho \in \{0.25,0.5,0.75\}$ as the measurement ratio $\beta = D/P$ varies, averaged over 10 independent AMP iterates. Additionally, we show convergence of the (averaged) AMP iterates for sparsity $\rho = 0.25$ at a range of β near the recovery threshold. We observe strong agreement between AMP empirical performance and the state evolution prediction. The system sizes are $P = 1024$, $q = 1024$, with $D = \beta P$ varying.

In Figure 4.4, we show two examples of $L = 2,3,4$ layer models following Equation (4.2). In both, the output channel $l = 1$ generates noisy, compressive linear measurements $y = z^{(1)} + \zeta$ for $\zeta_i \sim \mathcal{N}(0,\sigma^2)$ and for dense couplings $W_{ij}^{(1)} \sim \mathcal{N}(0,1/n^{(1)})$. Layers $2 \leqslant l \leqslant 4$ use MCC couplings $W^{(l)} \sim \mathrm{MCC}(D_l,P_l,q,k)$, where $qP_l = n_l$ and $D_l = \beta P_l = qn_{l-1}$. Channel functions $\{\phi^{(l)}\}$ vary across the two experiments. The input prior is $P_X(x) = \mathcal{N}(x;0,1)$ and model has $q = 10$ channels, filter size $k = 3$, noise level $\sigma^2 = 10^{-4}$, input dimension $n^{(L)} = 5000$, layerwise aspect ratios $\beta^{(L)} = 2$ and $\beta^{(l)} = 1$ for $2 \leqslant l < L$. The channel aspect ratio $\beta^{(1)}$ varies in each experiment.

We compare the state evolution equations to empirical AMP results in two cases. In the left panel, we show multilayer models with identity channel functions, and in the right panel, we show models with ReLU channel functions. The latter model captures a simple but accurate example of a convolutional generative neural network.

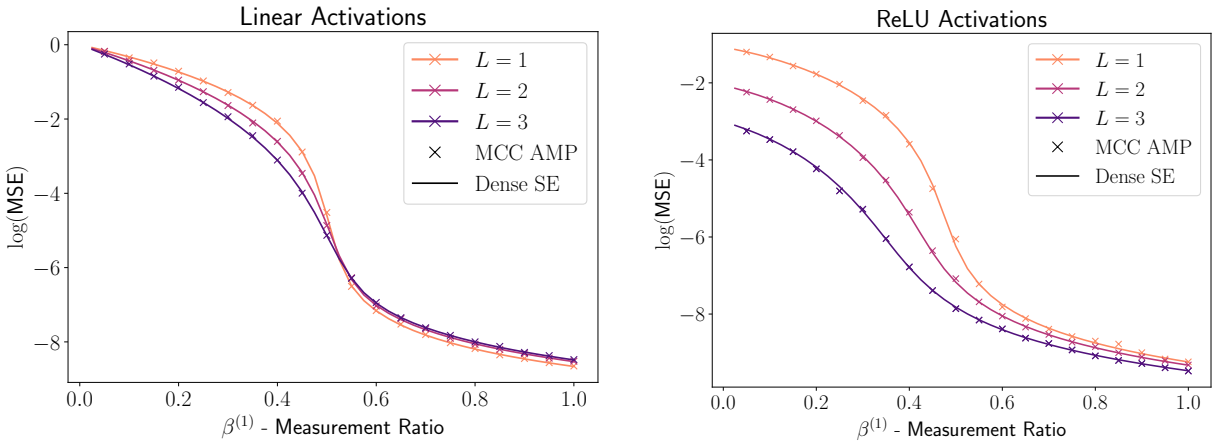


Figure 4.4: ML-AMP compressive sensing recovery under multichannel convolutional designs (crossed) and the corresponding state evolution for the corresponding fully connected model (lined). Left: For $2 \leqslant l \leqslant L$, the channel functions are $\phi^{(l)}(z) = z + \zeta$ where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$. Right: For $2 \leqslant l \leqslant L$, the channel functions are $\phi^{(l)}(z) = \max(z, 0)$ where the maximum is applied coordinate-wise.

Chapter 5

Proofs for the multi-layer random convolutional model

5.1 Proof of the main theorem

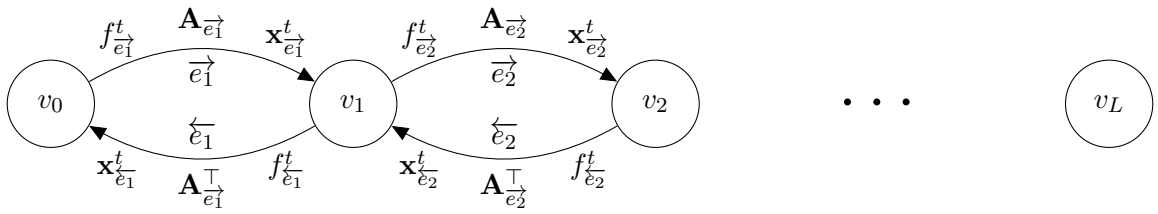
The proof of the main theorem is presented in this section. We start with a generic result on a family of AMP iterations including the (non Bayes-optimal) MLAMP one, using the framework of [110], from which we remind the required notions.

5.1.1 State evolution for generic multilayer AMP iterations with matrix valued variables and dense Gaussian matrices

In the notations of [110], consider the AMP iteration indexed by the following directed graph $G = (V, \overrightarrow{E})$, where the set of vertices is denoted $V = \{v_0, v_1, ..., v_L\}$, and the set of edges $\overrightarrow{E} = \{\overrightarrow{e}_1, ..., \overrightarrow{e}_l, \overleftarrow{e}_1, ..., \overleftarrow{e}_L\}$. For any edge $\overrightarrow{e}_l$, the corresponding matrix $\mathbf{A}_{\overrightarrow{e}_l}$ has dimensions $\mathbb{R}^{n_l \times n_{l-1}}$ with $\mathbf{A}_{\overleftarrow{e}_l} = \mathbf{A}_{\overrightarrow{e}_l}^\top$, and the variables $\mathbf{x}_{\overrightarrow{e}_l} \in \mathbb{R}^{n_l \times q}, \mathbf{x}_{\overleftarrow{e}_l} \in \mathbb{R}^{n_{l-1} \times q}$ for some finite $q \in \mathbb{N}$, with $N = \sum_{l=1}^L n_l$. Finally, we define the non-linearities of the iteration by specifying the variables they are acting on as follows:

- $f^t_{\overrightarrow{e}_1} : \mathbb{R}^{n_0 \times q} \to \mathbb{R}^{n_0 \times q}, \mathbf{x}^t_{\overleftarrow{e}_1} \mapsto f^t_{\overrightarrow{e}_1}\left(\mathbf{x}^t_{\overleftarrow{e}_1}\right)$,
- for any $2 \leqslant l \leqslant L$, $f^t_{\overrightarrow{e}_l} : (\mathbb{R}^{n_{l-1} \times q})^2 \to \mathbb{R}^{n_{l-1} \times q}$, $(\mathbf{x}^t_{\overrightarrow{e}_{l-1}}, \mathbf{x}^t_{\overleftarrow{e}_l}) \mapsto f^t_{\overrightarrow{e}_l}(\mathbf{x}^t_{\overrightarrow{e}_{l-1}}, \mathbf{x}^t_{\overleftarrow{e}_l})$,
- for any $1 \leqslant l \leqslant L-1$, $f^t_{\overleftarrow{e}_l} : (\mathbb{R}^{n_l \times q})^3 \to \mathbb{R}^{n_l \times q}$, $(\mathbf{x}^t_{\overrightarrow{e}_l}, \mathbf{x}^t_{\overleftarrow{e}_{l+1}}) \mapsto f^t_{\overleftarrow{e}_l}(\mathbf{A}_{\overrightarrow{e}_l}\mathbf{w}_{\overrightarrow{e}_l}, \mathbf{x}^t_{\overrightarrow{e}_l}, \mathbf{x}^t_{\overleftarrow{e}_{l+1}})$
- $f^t_{\overleftarrow{e}_L} : (\mathbb{R}^{n_L \times q})^2 \to \mathbb{R}^{n_L \times q}, \mathbf{x}^t_{\overleftarrow{e}_L} \to f^t_{\overleftarrow{e}_L}\left(\mathbf{A}_{\overrightarrow{e}_L}\mathbf{w}_{\overrightarrow{e}_L}, \mathbf{x}^t_{\overleftarrow{e}_L}\right)$

where $\mathbf{w}_{\overrightarrow{e}_1}, ..., \mathbf{w}_{\overrightarrow{e}_L}$ are low-rank matrices respectively in $\mathbb{R}^{n_0 \times q}, ..., \mathbb{R}^{n_{L-1} \times q}$, whose rows are sampled i.i.d. from subgaussian probability distributions in $\mathbb{R}^q$. The graph indexing the iteration then reads:



with the corresponding iteration:

$$
\begin{aligned}
\mathbf{x}^{t+1}_{\overrightarrow{e_1}} &= \mathbf{A}_{\overrightarrow{e_1}}\mathbf{m}^t_{\overrightarrow{e_1}} - \mathbf{m}^{t-1}_{\overleftarrow{e_1}}\left(\mathbf{b}^t_{\overrightarrow{e_1}}\right)^\top, \\
\mathbf{m}^t_{\overrightarrow{e_1}} &= f^t_{\overrightarrow{e}_1}\left(\mathbf{x}^t_{\overleftarrow{e_1}}\right), \\
\mathbf{x}^{t+1}_{\overleftarrow{e_1}} &= \mathbf{A}^\top_{\overrightarrow{e_1}}\mathbf{m}^t_{\overleftarrow{e_1}} - \mathbf{m}^{t-1}_{\overrightarrow{e_1}}\left(\mathbf{b}^t_{\overleftarrow{e_1}}\right)^\top, \\
\mathbf{m}^t_{\overleftarrow{e_1}} &= f^t_{\overleftarrow{e_1}}\left(\mathbf{A}_{\overrightarrow{e}_1}\mathbf{w}_{\overrightarrow{e}_1}, \mathbf{x}^t_{\overrightarrow{e_1}}, \mathbf{x}^t_{\overleftarrow{e_2}}\right), \\
\\
\mathbf{x}^{t+1}_{\overrightarrow{e_2}} &= \mathbf{A}_{\overrightarrow{e_2}}\mathbf{m}^t_{\overrightarrow{e_2}} - \mathbf{m}^{t-1}_{\overleftarrow{e_2}}\left(\mathbf{b}^t_{\overrightarrow{e_2}}\right)^\top, \\
\mathbf{m}^t_{\overrightarrow{e_2}} &= f^t_{\overrightarrow{e}_2}\left(\mathbf{x}^t_{\overrightarrow{e_1}}, \mathbf{x}^t_{\overleftarrow{e_2}}\right), \\
\mathbf{x}^{t+1}_{\overleftarrow{e_2}} &= \mathbf{A}^\top_{\overrightarrow{e_2}}\mathbf{m}^t_{\overleftarrow{e_2}} - \mathbf{m}^{t-1}_{\overrightarrow{e_2}}(\mathbf{b}^t_{\overleftarrow{e_2}})^\top, \\
\mathbf{m}^t_{\overleftarrow{e_2}} &= f^t_{\overleftarrow{e_2}}\left(\mathbf{A}_{\overrightarrow{e}_2}\mathbf{w}_{\overrightarrow{e}_2}, \mathbf{x}^t_{\overrightarrow{e_2}}, \mathbf{x}^t_{\overleftarrow{e_3}}\right), \\
\\
&\vdots \\
\\
\mathbf{x}^{t+1}_{\overrightarrow{e_L}} &= \mathbf{A}_{\overrightarrow{e_L}}\mathbf{m}^t_{\overrightarrow{e_L}} - \mathbf{m}^{t-1}_{\overleftarrow{e_L}}\left(\mathbf{b}^t_{\overrightarrow{e_L}}\right)^\top, \\
\mathbf{m}^t_{\overrightarrow{e_L}} &= f^t_{\overrightarrow{e}_L}\left(\mathbf{x}^t_{\overrightarrow{e}_{L-1}}, \mathbf{x}^t_{\overleftarrow{e_L}}\right), \\
\mathbf{x}^{t+1}_{\overleftarrow{e_L}} &= \mathbf{A}^\top_{\overrightarrow{e_L}}\mathbf{m}^t_{\overleftarrow{e_L}} - \mathbf{m}^{t-1}_{\overrightarrow{e_L}}(\mathbf{b}^t_{\overleftarrow{e_L}})^\top, \\
\mathbf{m}^t_{\overleftarrow{e_L}} &= f^t_{\overleftarrow{e_L}}\left(\mathbf{A}_{\overrightarrow{e}_L}\mathbf{w}_{\overrightarrow{e}_L}, \mathbf{x}^t_{\overrightarrow{e_L}}\right)
\end{aligned}
\tag{5.1}
$$

and Onsager terms, for the right oriented edges

$$
\mathbf{b}^t_{\overrightarrow{e}_l} = \frac{1}{N}\sum_{i=1}^{n_{l-1}} \frac{\partial f^t_{\overrightarrow{e}_l,i}}{\partial \mathbf{x}_{\overleftarrow{e}_l,i}}\left(\left(\mathbf{x}^t_{\overrightarrow{e}'_l}\right)_{\overrightarrow{e}'_l:\overrightarrow{e}'_l\to\overrightarrow{e}_l}\right) \qquad \in \mathbb{R}^{q\times q}.
$$

and left oriented edges

$$
\mathbf{b}^t_{\overleftarrow{e}_l} = \frac{1}{N}\sum_{i=1}^{n_l} \frac{\partial f^t_{\overleftarrow{e}_l,i}}{\partial \mathbf{x}_{\overrightarrow{e}_l,i}}\left(\mathbf{A}_{\overrightarrow{e}_l}\mathbf{w}_{\overrightarrow{e}_l}, \left(\mathbf{x}^t_{\overleftarrow{e}'_l}\right)_{\overleftarrow{e}'_l:\overleftarrow{e}'_l\to\overleftarrow{e}_l}\right) \qquad \in \mathbb{R}^{q\times q}.
$$

We now make the following assumptions

(A1) The matrices $(\mathbf{A}_{\overrightarrow{e}})_{\overrightarrow{e}\in\overrightarrow{E}}$ are random and independent, up to the symmetry condition $\mathbf{A}_{\overleftarrow{e}} = \mathbf{A}^\top_{\overrightarrow{e}}$. Moreover $\mathbf{A}_{\overrightarrow{e}}$ has independent centered Gaussian entries with variance $1/N$.

(A2) For all $1 \leqslant l \leqslant L$, $n_l \to \infty$ and n_l/N converges to a well-defined limit $\delta_l \in [0,1]$. We denote by $n \to \infty$ the limit under this scaling.

(A3) For all $t \in \mathbb{N}$ and $\overrightarrow{e} \in \overrightarrow{E}$, the non-linearity $f^t_{\overrightarrow{e}}$ is pseudo-Lipschitz of finite order, uniformly with respect to the problem dimensions $(n_l)_{0\leqslant l\leqslant L}$

(A4) For all $\overrightarrow{e} \in E$, the lines of $\mathbf{x}^0_{\overrightarrow{e}}, \mathbf{w}_{\overrightarrow{e}}$ are sampled from subgaussian probability distributions in $\mathbf{R}^q$.

(A5) For all $\overrightarrow{e} \in E$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{N} \left\langle f^0_{\overrightarrow{e}} \left(\left(\mathbf{x}^0_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right), f^0_{\overrightarrow{e}} \left(\left(\mathbf{x}^0_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\rangle$$

(A6) Let $(\kappa_{\overrightarrow{e}})_{\overrightarrow{e}\in E}$ be an array of bounded non-negative reals and $\mathbf{Z}_{\overrightarrow{e}} \sim \mathbf{N}(0, \kappa_{\overrightarrow{e}} \mathbf{I}_{n_w})$ independent random variables for all $\overrightarrow{e}$. For all $\overrightarrow{e} \in E$, for any $t \in \mathbb{N}_{>0}$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{N} \mathbb{E} \left[\left\langle f^0_{\overrightarrow{e}} \left(\left(\mathbf{x}^0_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right), f^t_{\overrightarrow{e}} \left(\left(\mathbf{Z}^t_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\rangle \right].$$

(A7) Consider any array of 2×2 positive definite matrices $(\boldsymbol{S}_{\overrightarrow{e}})_{\overrightarrow{e}\in E}$ and the collection of random variables $(\mathbf{Z}_{\overrightarrow{e}}, \mathbf{Z}'_{\overrightarrow{e}}) \sim \mathbf{N}(0, \boldsymbol{S}_{\overrightarrow{e}} \otimes \mathbf{I}_{n_w}))$ defined independently for each edge $\overrightarrow{e}$. Then for any $\overrightarrow{e} \in E$ and $s, t > 0$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{N} \mathbb{E} \left[\left\langle f^s_{\overrightarrow{e}} \left((\mathbf{Z}^s_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right), f^t_{\overrightarrow{e}} \left(\left(\tilde{\mathbf{Z}}^t_{\overrightarrow{e}'} \right)_{\overrightarrow{e}':\overrightarrow{e}'\to\overrightarrow{e}} \right) \right\rangle \right].$$

Under these assumptions, we define the following state evolution recursion:

- for $l = 1$:

$$\boldsymbol{\nu}^0_{\overrightarrow{e}_1} = \lim_{N\to\infty} \frac{1}{N} \mathbf{w}^\top_{\overrightarrow{e}_1} f^0_{\overrightarrow{e}_1}(\mathbf{x}^0_{\overleftarrow{e}_1}),\ \boldsymbol{\kappa}^{1,1}_{\overrightarrow{e}_1} = \lim_{N\to\infty} \frac{1}{N} f^0_{\overrightarrow{e}_1}(\mathbf{x}^0_{\overleftarrow{e}_1})^\top f^0_{\overrightarrow{e}_1}(\mathbf{x}^0_{\overleftarrow{e}_1}) \tag{5.2}$$

$$\boldsymbol{\nu}^{t+1}_{\overrightarrow{e}_1} = \lim_{N\to+\infty} \frac{1}{N} \mathbb{E} \left[\mathbf{w}^\top_{\overrightarrow{e}_1} f^t_{\overrightarrow{e}_1} \left(\mathbf{w}_{\overrightarrow{e}_1} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_1} + \mathbf{Z}^t_{\overleftarrow{e}_1} \right) \right] \tag{5.3}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\overrightarrow{e}_1} = \boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}_1} = \lim_{N\to+\infty} \frac{1}{N} \mathbb{E} \Big[\left(f^s_{\overrightarrow{e}_1} \left(\mathbf{w}_{\overrightarrow{e}_1} \hat{\boldsymbol{\nu}}^s_{\overleftarrow{e}_1} + \mathbf{Z}^s_{\overleftarrow{e}_1} \right) - \mathbf{w}_{\overrightarrow{e}_1} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_1}} \boldsymbol{\nu}^{s+1}_{\overrightarrow{e}_1} \right)^\top$$
$$\left(f^t_{\overrightarrow{e}_1} \left(\mathbf{w}_{\overrightarrow{e}_1} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_1} + \mathbf{Z}^t_{\overleftarrow{e}_1} \right) - \mathbf{w}_{\overrightarrow{e}_1} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_1}} \boldsymbol{\nu}^{t+1}_{\overrightarrow{e}_1} \right) \Big] \tag{5.4}$$

$$\hat{\boldsymbol{\nu}}^0_{\overleftarrow{e}_1}, \boldsymbol{\kappa}^{1,1}_{\overleftarrow{e}_1} = \lim_{n\to\infty} \frac{1}{N} f^0_{\overleftarrow{e}_1} \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}}, \mathbf{x}^0_{\overrightarrow{e}_1}, \mathbf{x}^0_{\overleftarrow{e}_2} \right)^\top f^0_{\overleftarrow{e}_1} \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}}, \mathbf{x}^0_{\overrightarrow{e}_1}, \mathbf{x}^0_{\overleftarrow{e}_2} \right) \tag{5.5}$$

$$\hat{\boldsymbol{\nu}}^{t+1}_{\overleftarrow{e}_1} = \lim_{N\to\infty} \frac{1}{N} \mathbb{E} \Big[\sum_{i=1}^N \frac{\partial f^t_{\overleftarrow{e}_1,i}}{\partial \mathbf{z}_{\mathbf{w}_{\overleftarrow{e}_1},i}, \varphi_{\overleftarrow{e}_1}} \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_1}} \boldsymbol{\nu}^t_{\overrightarrow{e}_1} + \mathbf{Z}^t_{\overrightarrow{e}_1}, \mathbf{w}_{\overrightarrow{e}_2} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_2} + \mathbf{Z}^t_{\overleftarrow{e}_2} \right) \Big] \tag{5.6}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\overleftarrow{e}_1} = \lim_{n\to\infty} \frac{1}{N} \mathbb{E} \Big[f^s_{\overleftarrow{e}_1} \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_1}} \boldsymbol{\nu}^s_{\overrightarrow{e}_1} + \mathbf{Z}^s_{\overrightarrow{e}_1}, \mathbf{w}_{\overrightarrow{e}_2} \hat{\boldsymbol{\nu}}^s_{\overleftarrow{e}_2} + \mathbf{Z}^s_{\overleftarrow{e}_2} \right)^\top$$
$$f^t_{\overleftarrow{e}_1} \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_1}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_1}} \boldsymbol{\nu}^t_{\overrightarrow{e}_1} + \mathbf{Z}^t_{\overrightarrow{e}_1}, \mathbf{w}_{\overrightarrow{e}_2} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_2} + \mathbf{Z}^t_{\overleftarrow{e}_2} \right) \Big] \tag{5.7}$$

- for any $2 \leqslant l \leqslant L-1$

$$\boldsymbol{\nu}^0_{\overrightarrow{e}_l} = \lim_{N\to\infty} \frac{1}{N} \mathbf{w}^\top_{\overrightarrow{e}_l} f^0_{\overrightarrow{e}_l}(\mathbf{x}^0_{\overleftarrow{e}_l}),\ \boldsymbol{\kappa}^{1,1}_{\overrightarrow{e}_l} = \lim_{N\to\infty} \frac{1}{N} f^0_{\overrightarrow{e}_l}(\mathbf{x}^0_{\overleftarrow{e}_l})^\top f^0_{\overrightarrow{e}_l}(\mathbf{x}^0_{\overleftarrow{e}_l}) \tag{5.8}$$

$$\boldsymbol{\nu}^{t+1}_{\overrightarrow{e}_l} = \lim_{N\to+\infty} \frac{1}{N} \mathbb{E}\left[\mathbf{w}^\top_{\overrightarrow{e}_l} f^t_{\overrightarrow{e}_l}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_{l-1}}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_{l-1}}} \boldsymbol{\nu}^t_{\overrightarrow{e}_{l-1}} + \mathbf{Z}^t_{\overrightarrow{e}_{l-1}}, \mathbf{w}_{\overrightarrow{e}_l} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_l} + \mathbf{Z}^t_{\overleftarrow{e}_l} \right) \right] \tag{5.9}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\overrightarrow{e}_l} = \boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}_l} = \lim_{N\to+\infty} \tag{5.10}$$

$$\frac{1}{N}\mathbb{E}\Big[\left(f^s_{\overrightarrow{e}_l}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_{l-1}}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_{l-1}}} \boldsymbol{\nu}^s_{\overrightarrow{e}_{l-1}} + \mathbf{Z}^s_{\overrightarrow{e}_{l-1}}, \mathbf{w}_{\overrightarrow{e}_l} \hat{\boldsymbol{\nu}}^s_{\overleftarrow{e}_l} + \mathbf{Z}^s_{\overleftarrow{e}_l} \right) - \mathbf{w}_{\overrightarrow{e}_l} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_l}} \boldsymbol{\nu}^{s+1}_{\overrightarrow{e}_l} \right)^\top$$
$$\left(f^t_{\overrightarrow{e}_l}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_{l-1}}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_{l-1}}} \boldsymbol{\nu}^t_{\overrightarrow{e}_{l-1}} + \mathbf{Z}^t_{\overrightarrow{e}_{l-1}}, \mathbf{w}_{\overrightarrow{e}_l} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_l} + \mathbf{Z}^t_{\overleftarrow{e}_l} \right) - \mathbf{w}_{\overrightarrow{e}_l} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_l}} \boldsymbol{\nu}^{t+1}_{\overrightarrow{e}_l} \right) \Big] \tag{5.11}$$

$$\hat{\boldsymbol{\nu}}^0_{\overleftarrow{e}_l}, \boldsymbol{\kappa}^{1,1}_{\overleftarrow{e}_l} = \lim_{n\to\infty} \frac{1}{N} f^0_{\overleftarrow{e}_l}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}}, \mathbf{x}^0_{\overrightarrow{e}_l}, \mathbf{x}^0_{\overleftarrow{e}_{l+1}} \right)^\top f^0_{\overleftarrow{e}_l}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}}, \mathbf{x}^0_{\overrightarrow{e}_l}, \mathbf{x}^0_{\overleftarrow{e}_{l+1}} \right) \tag{5.12}$$

$$\hat{\boldsymbol{\nu}}^{t+1}_{\overleftarrow{e}_l} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Big[\sum_{i=1}^N \frac{\partial f^t_{\overleftarrow{e}_l,i}}{\partial \mathbf{z}_{\mathbf{w}_{\overleftarrow{e}_l}},i, \varphi_{\overleftarrow{e}_l}} \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_l}} \boldsymbol{\nu}^t_{\overrightarrow{e}_l} + \mathbf{Z}^t_{\overrightarrow{e}_l}, \mathbf{w}_{\overrightarrow{e}_{l+1}} \hat{\boldsymbol{\nu}}^t_{\overrightarrow{e}_{l+1}} \mathbf{Z}^t_{\overleftarrow{e}_{l+1}} \right) \Big] \tag{5.13}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\overleftarrow{e}_l} = \lim_{n\to\infty} \frac{1}{N}\mathbb{E}\Big[f^s_{\overleftarrow{e}_l}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_l}} \boldsymbol{\nu}^s_{\overrightarrow{e}_l} + \mathbf{Z}^s_{\overrightarrow{e}_l}, \mathbf{w}_{\overrightarrow{e}_{l+1}} \hat{\boldsymbol{\nu}}^s_{\overrightarrow{e}_{l+1}} \mathbf{Z}^s_{\overleftarrow{e}_{l+1}} \right)^\top$$
$$f^t_{\overleftarrow{e}_l}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_l}} \boldsymbol{\nu}^t_{\overrightarrow{e}_l} + \mathbf{Z}^t_{\overrightarrow{e}_l}, \mathbf{w}_{\overrightarrow{e}_{l+1}} \hat{\boldsymbol{\nu}}^t_{\overrightarrow{e}_{l+1}} \mathbf{Z}^t_{\overleftarrow{e}_{l+1}} \right) \Big] \tag{5.14}$$

- for l=L

$$\boldsymbol{\nu}^0_{\overrightarrow{e}_L} = \lim_{N\to\infty} \frac{1}{N} \mathbf{w}^\top_{\overrightarrow{e}_l} f^0_{\overrightarrow{e}_L}(\mathbf{x}^0_{\overleftarrow{e}_L}),\ \boldsymbol{\kappa}^{1,1}_{\overrightarrow{e}_L} = \lim_{N\to\infty} \frac{1}{N} f^0_{\overrightarrow{e}_L}(\mathbf{x}^0_{\overleftarrow{e}_L})^\top f^0_{\overrightarrow{e}_L}(\mathbf{x}^0_{\overleftarrow{e}_L}) \tag{5.15}$$

$$\boldsymbol{\nu}^{t+1}_{\overrightarrow{e}_L} = \lim_{N\to+\infty} \frac{1}{N} \mathbb{E}\left[\mathbf{w}^\top_{\overrightarrow{e}_L} f^t_{\overrightarrow{e}_L}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_{L-1}}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_{L-1}}} \boldsymbol{\nu}^t_{\overrightarrow{e}_{L-1}} + \mathbf{Z}^t_{\overrightarrow{e}_{L-1}}, \mathbf{w}_{\overrightarrow{e}_L} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_L} + \mathbf{Z}^t_{\overleftarrow{e}_L} \right) \right] \tag{5.16}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\overrightarrow{e}_L} = \boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}_L} = \lim_{N\to+\infty} \tag{5.17}$$

$$\frac{1}{N}\mathbb{E}\Big[\left(f^s_{\overrightarrow{e}_L}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_{L-1}}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_{L-1}}} \boldsymbol{\nu}^s_{\overrightarrow{e}_{L-1}} + \mathbf{Z}^s_{\overrightarrow{e}_{L-1}}, \mathbf{w}_{\overrightarrow{e}_L} \hat{\boldsymbol{\nu}}^s_{\overleftarrow{e}_L} + \mathbf{Z}^s_{\overleftarrow{e}_L} \right) - \mathbf{w}_{\overrightarrow{e}_L} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_L}} \boldsymbol{\nu}^{s+1}_{\overrightarrow{e}_L} \right)^\top$$
$$\left(f^t_{\overrightarrow{e}_L}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_{L-1}}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_{L-1}}} \boldsymbol{\nu}^t_{\overrightarrow{e}_{L-1}} + \mathbf{Z}^t_{\overrightarrow{e}_{L-1}}, \mathbf{w}_{\overrightarrow{e}_L} \hat{\boldsymbol{\nu}}^t_{\overleftarrow{e}_L} + \mathbf{Z}^t_{\overleftarrow{e}_L} \right) - \mathbf{w}_{\overrightarrow{e}_L} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_L}} \boldsymbol{\nu}^{t+1}_{\overrightarrow{e}_L} \right) \Big] \tag{5.18}$$

$$\hat{\boldsymbol{\nu}}^0_{\overleftarrow{e}_L}, \boldsymbol{\kappa}^{1,1}_{\overleftarrow{e}_L} = \lim_{n\to\infty} \frac{1}{N} f^0_{\overleftarrow{e}_L}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}}, \mathbf{x}^0_{\overrightarrow{e}_L} \right)^\top f^0_{\overleftarrow{e}_L}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}} \right) \tag{5.19}$$

$$\hat{\boldsymbol{\nu}}^{t+1}_{\overleftarrow{e}_L} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Big[\sum_{i=1}^N \frac{\partial f^t_{\overleftarrow{e}_L,i}}{\partial \mathbf{z}_{\mathbf{w}_{\overleftarrow{e}_L}},i, \varphi_{\overleftarrow{e}_L}} \left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_L}} \boldsymbol{\nu}^t_{\overrightarrow{e}_L} + \mathbf{Z}^t_{\overrightarrow{e}_L} \right) \Big] \tag{5.20}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\overleftarrow{e}_L} = \lim_{n\to\infty} \frac{1}{N}\mathbb{E}\Big[f^s_{\overleftarrow{e}_L}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_L}} \boldsymbol{\nu}^s_{\overrightarrow{e}_L} + \mathbf{Z}^s_{\overrightarrow{e}_L} \right)^\top$$
$$f^t_{\overleftarrow{e}_L}\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}}, \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_L}} \rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_L}} \boldsymbol{\nu}^t_{\overrightarrow{e}_L} + \mathbf{Z}^t_{\overrightarrow{e}_L} \right) \Big] \tag{5.21}$$

where, for any $1 \leqslant l \leqslant L$, the symbol $\partial \mathbf{z}_{\mathbf{w}_{\overrightarrow{e}}},i, \varphi_{\overrightarrow{e}}$ denotes the partial derivative w.r.t. the argument of $\varphi_{\overrightarrow{e}}$, $(\mathbf{Z}^1_{\overrightarrow{e}}, \dots, \mathbf{Z}^t_{\overrightarrow{e}})$ is a centered Gaussian random vector with covariance $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}})_{r,s\leqslant t} \otimes \mathbf{I}_{n_w}$ (and similarly for left-oriented edges), and $z_{\mathbf{w}_{\overrightarrow{e}}}$ is distributed according to $\mathbf{N}(0, \boldsymbol{\rho}_{\mathbf{w}_{\overrightarrow{e}}})$.

Theorem 8. *Assume (A1)-(A7). Define, as above, independently for each* $\overrightarrow{e}_l$, $\mathbf{Z}^0_{\overrightarrow{e}_l} = \mathbf{x}^0_{\overrightarrow{e}_l}$ *and* $(\mathbf{Z}^1_{\overrightarrow{e}_l}, \dots, \mathbf{Z}^t_{\overrightarrow{e}_l})$ *a centered Gaussian random vector of covariance* $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}_l})_{r,s\leqslant t} \otimes \mathbf{I}_{n_{l-1}}$. *Then for any sequence of uniformly (in n) pseudo-Lipschitz function* $\Phi : (\mathbb{R}^{n_{l-1}\times(t+1)q})^2 \to \mathbb{R}$, *for any* $1 \leqslant l \leqslant L$

$$\Phi\left(\left(\mathbf{x}^s_{\overrightarrow{e}_l}\right)_{0\leqslant s\leqslant t}, \left(\mathbf{x}^s_{\overleftarrow{e}_{l-1}}\right)_{0\leqslant s\leqslant t}\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\Phi\left(\left(\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}}\rho^{-1}_{\mathbf{w}_{\overrightarrow{e}_l}}\boldsymbol{\nu}^s_{\overrightarrow{e}_l} + \mathbf{Z}^s_{\overrightarrow{e}_{l-1}}\right)_{0\leqslant s\leqslant t}, \left(\mathbf{w}_{\overrightarrow{e}_{l-1}}\hat{\boldsymbol{\nu}}^s_{\overleftarrow{e}_{l-1}} + \mathbf{Z}^s_{\overleftarrow{e}_{l-1}}\right)_{0\leqslant s\leqslant t}\right)\right]$$

In summary, at each time step, the variables associated with right oriented edges $\mathbf{x}_{\overrightarrow{e}_l}$ asymptotically behave as the sum of the ground truth $\mathbf{w}_{\overrightarrow{e}_l}$ reweighted by a $q \times q$ matrix coefficient $\hat{\boldsymbol{\nu}}_{\overleftarrow{e}_l}$ and a $n_{l-1} \times q$ random matrix with i.i.d. lines $\mathbf{Z}_{\overrightarrow{e}_l}$ with $q \times q$ covariance $\boldsymbol{\kappa}_{\overleftarrow{e}_l}$ determined by the function associated to the corresponding left-oriented arrow $f^t_{\overleftarrow{e}_l}$. Similarly, the variables associated with left oriented edges $\mathbf{x}_{\overleftarrow{e}_l}$ asymptotically behave as the sum of the linear response to the ground truth $\mathbf{z}_{\mathbf{w}_{\overrightarrow{e}_l}}$ (asymptotic equivalent of $\mathbf{A}_{\overrightarrow{e}_l}\mathbf{w}_{\overrightarrow{e}_l}$) reweighted by a $q \times q$ matrix coefficient $\boldsymbol{\nu}_{\overleftarrow{e}_l}$ and a $n_l \times q$ random matrix with i.i.d. lines $\mathbf{Z}_{\overleftarrow{e}_l}$ with $q \times q$ covariance $\boldsymbol{\kappa}_{\overrightarrow{e}_l}$ determined by the function associated to the corresponding right-oriented arrow $f^t_{\overrightarrow{e}_l}$.

Proof. This result is a special case of Lemma 2 from [110], with a perturbation where only the left-oriented edges involve an additional dependence on $\mathbf{A}_{\overrightarrow{e}}\mathbf{w}_{\overrightarrow{e}}$. The required conditions are the same as in [110], barring the subgaussian assumption (A3) which ensures the scaled norm of the $\mathbf{x}^0_{\overrightarrow{e}}, \mathbf{w}_{\overrightarrow{e}}$ are finite with high-probability as $n \to \infty$. □

5.1.2 State evolution for multilayer AMP iterations with random convolutional matrices

The following lemma proves the state evolution equations for a multilayer AMP iteration where the dense Gaussian matrices are replaced with random convolutional ones (MCC from Def.9) with variance $\frac{1}{N}$, with a vector valued variables, i.e. q=1, and separables non-linearities. We choose the variance as $\frac{1}{N}$ to follow the notations of [110] for more convenience, recovering the variances of iteration Eq.(4.4) is a straightforward rescaling as done in [37] and will be discussed in the next section. Assume $q = 1$ and that, for any $t \in \mathbb{N}$ and $1 \leqslant l \leqslant L$, the functions $f^t_{\overrightarrow{e}_l}, f^t_{\overleftarrow{e}_l}$ are separable in all their arguments, i.e there exists scalar valued, pseudo-Lipschitz functions $\sigma^t_{\overrightarrow{e}_l} : \mathbb{R}^2 \to \mathbb{R}, \sigma^t_{\overleftarrow{e}_l} : \mathbb{R}^3 \to \mathbb{R}$ (where $\sigma^t_{\overrightarrow{e}_1} : \mathbb{R} \to \mathbb{R}, \sigma^t_{\overleftarrow{e}_L} : \mathbb{R}^2 \to \mathbb{R}$) such that:

for $l = 1$, for any $1 \leqslant i \leqslant n_0$:

$$f^t_{\overleftarrow{e}_1}(\mathbf{x}^t_{\overleftarrow{e}_1})_i = \sigma^t_{\overleftarrow{e}_1}(x^t_{\overleftarrow{e}_1,i})$$

for any $1 \leqslant l \leqslant L-1$, for any $1 \leqslant i \leqslant n_l$:

$$f^t_{\overleftarrow{e}_l}\left(\mathbf{A}_{\overrightarrow{e}_l}\mathbf{w}_{\overrightarrow{e}_l}, \mathbf{x}^t_{\overrightarrow{e}_l}, \mathbf{x}^t_{\overleftarrow{e}_{l+1}}\right)_i = \sigma^t_{\overleftarrow{e}_l}\left((\mathbf{A}_{\overrightarrow{e}_l}\mathbf{w}_{\overrightarrow{e}_l})_i, x^t_{\overrightarrow{e}_l,i}, x^t_{\overleftarrow{e}_{l+1},i}\right)$$

for any $2 \leqslant l \leqslant L$, $1 \leqslant i \leqslant n_{l-1}$:

$$f^t_{\overrightarrow{e}_l}\left(\mathbf{x}^t_{\overrightarrow{e}_{l-1}}, \mathbf{x}^t_{\overleftarrow{e}_l}\right)_i = \sigma^t_{\overrightarrow{e}_l}\left(x^t_{\overrightarrow{e}_{l-1},i}, x^t_{\overleftarrow{e}_l,i}\right)$$

for l=L, any $1 \leqslant i \leqslant n_L$:

$$f^t_{\overleftarrow{e}_L}(\mathbf{A}_{\overrightarrow{e}_L}\mathbf{w}_{\overleftarrow{e}_L}, \mathbf{x}^t_{\overrightarrow{e}_L})_i = \sigma^t_{\overleftarrow{e}_L}((\mathbf{A}_{\overrightarrow{e}_L}\mathbf{w}_{\overleftarrow{e}_L})_i, x^t_{\overrightarrow{e}_L,i})$$

Define the following scalar SE equations

- for $l = 1$:

$$\nu^0_{\overrightarrow{e}_1} = \delta_0 \mathbb{E}\left[w_{\overrightarrow{e}_1}\sigma^0_{\overrightarrow{e}_1}(x^0_{\overleftarrow{e}_1})\right],\ \kappa^{1,1}_{\overrightarrow{e}_1} = \delta_0 \mathbb{E}\left[\sigma^0_{\overrightarrow{e}_1}(x^0_{\overleftarrow{e}_1})\sigma^0_{\overrightarrow{e}_1}(x^0_{\overleftarrow{e}_1})\right] \tag{5.22}$$

$$\nu^{t+1}_{\overrightarrow{e}_1} = \delta_0 \mathbb{E}\left[w_{\overrightarrow{e}_1}\sigma^t_{\overrightarrow{e}_1}\left(w_{\overrightarrow{e}_1}\hat{\nu}^t_{\overleftarrow{e}_1} + Z^t_{\overleftarrow{e}_1}\right)\right] \tag{5.23}$$

$$\kappa^{s+1,t+1}_{\overrightarrow{e}_1} = \kappa^{t+1,s+1}_{\overrightarrow{e}_1} = \delta_0 \mathbb{E}\Big[\left(\sigma^s_{\overrightarrow{e}_1}\left(w_{\overrightarrow{e}_1}\hat{\nu}^s_{\overleftarrow{e}_1} + Z^s_{\overleftarrow{e}_1}\right) - w_{\overrightarrow{e}_1}\rho^{-1}_{w_{\overrightarrow{e}_1}}\nu^{s+1}_{\overrightarrow{e}_1}\right)$$
$$\left(\sigma^t_{\overrightarrow{e}_1}\left(w_{\overrightarrow{e}_1}\hat{\nu}^t_{\overleftarrow{e}_1} + Z^t_{\overleftarrow{e}_1}\right) - w_{\overrightarrow{e}_1}\rho^{-1}_{w_{\overrightarrow{e}_1}}\nu^{t+1}_{\overrightarrow{e}_1}\right)\Big] \tag{5.24}$$

$$\hat{\nu}^0_{\overleftarrow{e}_1}, \kappa^{1,1}_{\overleftarrow{e}_1} = \delta_1 \mathbb{E}\Big[\sigma^0_{\overleftarrow{e}_1}\left(z_{w_{\overrightarrow{e}_1}}, x^0_{\overrightarrow{e}_1}, x^0_{\overleftarrow{e}_2}\right)\sigma^0_{\overleftarrow{e}_1}\left(z_{w_{\overrightarrow{e}_1}}, x^0_{\overrightarrow{e}_1}, x^0_{\overleftarrow{e}_2}\right)\Big] \tag{5.25}$$

$$\hat{\nu}^{t+1}_{\overleftarrow{e}_1} = \delta_1 \mathbb{E}\Big[\frac{\partial \sigma^t_{\overleftarrow{e}_1,i}}{\partial z_{w_{\overleftarrow{e}_1}}, i, \varphi_{\overleftarrow{e}_1}}\left(z_{w_{\overrightarrow{e}_1}}, z_{w_{\overrightarrow{e}_1}}\rho^{-1}_{w_{\overrightarrow{e}_1}}\nu^t_{\overrightarrow{e}_1} + Z^t_{\overrightarrow{e}_1}, w_{\overrightarrow{e}_2}\hat{\nu}^t_{\overleftarrow{e}_2} + Z^t_{\overleftarrow{e}_2}\right)\Big] \tag{5.26}$$

$$\kappa^{s+1,t+1}_{\overleftarrow{e}_1} = \delta_1 \mathbb{E}\Big[\sigma^s_{\overleftarrow{e}_1}\left(z_{w_{\overrightarrow{e}_1}}, z_{w_{\overrightarrow{e}_1}}\rho^{-1}_{w_{\overrightarrow{e}_1}}\nu^s_{\overrightarrow{e}_1} + Z^s_{\overrightarrow{e}_1}, w_{\overrightarrow{e}_2}\hat{\nu}^s_{\overleftarrow{e}_2} + Z^s_{\overleftarrow{e}_2}\right)$$
$$\sigma^t_{\overleftarrow{e}_1}\left(z_{w_{\overrightarrow{e}_1}}, z_{w_{\overrightarrow{e}_1}}\rho^{-1}_{w_{\overrightarrow{e}_1}}\nu^t_{\overrightarrow{e}_1} + Z^t_{\overrightarrow{e}_1}, w_{\overrightarrow{e}_2}\hat{\nu}^t_{\overleftarrow{e}_2} + Z^t_{\overleftarrow{e}_2}\right)\Big] \tag{5.27}$$

- for any $2 \leqslant l \leqslant L-1$

$$\nu^0_{\overrightarrow{e}_l} = \delta_{n_{l-1}} \mathbb{E}\Big[w_{\overrightarrow{e}_l}\sigma^0_{\overrightarrow{e}_l}(x^0_{\overleftarrow{e}_l})\Big],\ \kappa^{1,1}_{\overrightarrow{e}_l} = \delta_{n_{l-1}} \mathbb{E}\left[\sigma^0_{\overrightarrow{e}_l}(x^0_{\overleftarrow{e}_l})\sigma^0_{\overrightarrow{e}_l}(x^0_{\overleftarrow{e}_l})\right] \tag{5.28}$$

$$\nu^{t+1}_{\overrightarrow{e}_l} = \delta_{n_{l-1}} \mathbb{E}\left[w_{\overrightarrow{e}_l}\sigma^t_{\overrightarrow{e}_l}\left(z_{w_{\overrightarrow{e}_{l-1}}}\rho^{-1}_{w_{\overrightarrow{e}_{l-1}}}\nu^t_{\overrightarrow{e}_{l-1}} + Z^t_{\overrightarrow{e}_{l-1}}, w_{\overrightarrow{e}_l}\hat{\nu}^t_{\overleftarrow{e}_l} + Z^t_{\overleftarrow{e}_l}\right)\right] \tag{5.29}$$

$$\kappa^{s+1,t+1}_{\overrightarrow{e}_l} = \kappa^{t+1,s+1}_{\overrightarrow{e}_l} = \tag{5.30}$$
$$\delta_{n_{l-1}} \mathbb{E}\Big[\left(\sigma^s_{\overrightarrow{e}_l}\left(z_{w_{\overrightarrow{e}_{l-1}}}\rho^{-1}_{w_{\overrightarrow{e}_{l-1}}}\nu^s_{\overrightarrow{e}_{l-1}} + Z^s_{\overrightarrow{e}_{l-1}}, w_{\overrightarrow{e}_l}\hat{\nu}^s_{\overleftarrow{e}_l} + Z^s_{\overleftarrow{e}_l}\right) - w_{\overrightarrow{e}_l}\rho^{-1}_{w_{\overrightarrow{e}_l}}\nu^{s+1}_{\overrightarrow{e}_l}\right)$$
$$\left(\sigma^t_{\overrightarrow{e}_l}\left(z_{w_{\overrightarrow{e}_{l-1}}}\rho^{-1}_{w_{\overrightarrow{e}_{l-1}}}\nu^t_{\overrightarrow{e}_{l-1}} + Z^t_{\overrightarrow{e}_{l-1}}, w_{\overrightarrow{e}_l}\hat{\nu}^t_{\overleftarrow{e}_l} + Z^t_{\overleftarrow{e}_l}\right) - w_{\overrightarrow{e}_l}\rho^{-1}_{w_{\overrightarrow{e}_l}}\nu^{t+1}_{\overrightarrow{e}_l}\right)\Big] \tag{5.31}$$

$$\hat{\nu}^0_{\overleftarrow{e}_l}, \kappa^{1,1}_{\overleftarrow{e}_l} = \delta_{n_l} \mathbb{E}\Big[\sigma^0_{\overleftarrow{e}_l}\left(z_{w_{\overrightarrow{e}_l}}, x^0_{\overrightarrow{e}_l}, x^0_{\overleftarrow{e}_{l+1}}\right)\sigma^0_{\overleftarrow{e}_l}\left(z_{w_{\overrightarrow{e}_l}}, x^0_{\overrightarrow{e}_l}, x^0_{\overleftarrow{e}_{l+1}}\right)\Big] \tag{5.32}$$

$$\hat{\nu}^{t+1}_{\overleftarrow{e}_l} = \delta_{n_l} \mathbb{E}\Big[\frac{\partial \sigma^t_{\overleftarrow{e}_l,i}}{\partial z_{w_{\overleftarrow{e}_l}}, i, \varphi_{\overleftarrow{e}_l}}\left(z_{w_{\overrightarrow{e}_l}}, z_{w_{\overrightarrow{e}_l}}\rho^{-1}_{w_{\overrightarrow{e}_l}}\nu^t_{\overrightarrow{e}_l} + Z^t_{\overrightarrow{e}_l}, w_{\overrightarrow{e}_{l+1}}\hat{\nu}^t_{\overrightarrow{e}_{l+1}}Z^t_{\overleftarrow{e}_{l+1}}\right)\Big] \tag{5.33}$$

$$\kappa^{s+1,t+1}_{\overleftarrow{e}_l} = \delta_{n_l} \mathbb{E}\Big[\sigma^s_{\overleftarrow{e}_l}\left(z_{w_{\overrightarrow{e}_l}}, z_{w_{\overrightarrow{e}_l}}\rho^{-1}_{w_{\overrightarrow{e}_l}}\nu^s_{\overrightarrow{e}_l} + Z^s_{\overrightarrow{e}_l}, w_{\overrightarrow{e}_{l+1}}\hat{\nu}^s_{\overrightarrow{e}_{l+1}}Z^s_{\overleftarrow{e}_{l+1}}\right)$$
$$\sigma^t_{\overleftarrow{e}_l}\left(z_{w_{\overrightarrow{e}_l}}, z_{w_{\overrightarrow{e}_l}}\rho^{-1}_{w_{\overrightarrow{e}_l}}\nu^t_{\overrightarrow{e}_l} + Z^t_{\overrightarrow{e}_l}, w_{\overrightarrow{e}_{l+1}}\hat{\nu}^t_{\overrightarrow{e}_{l+1}}Z^t_{\overleftarrow{e}_{l+1}}\right)\Big] \tag{5.34}$$

- for l=L

$$\nu^0_{\overrightarrow{e}_L} = \delta_{n_{L-1}}\mathbb{E}\left[w_{\overrightarrow{e}_l}\sigma^0_{\overrightarrow{e}_L}(x^0_{\overleftarrow{e}_L})\right],\ \kappa^{1,1}_{\overrightarrow{e}_L} = \delta_{n_{L-1}}\mathbb{E}\left[\sigma^0_{\overrightarrow{e}_L}(x^0_{\overleftarrow{e}_L})\sigma^0_{\overrightarrow{e}_L}(x^0_{\overleftarrow{e}_L})\right] \tag{5.35}$$

$$\nu^{t+1}_{\overrightarrow{e}_L} = \delta_{n_{L-1}}\mathbb{E}\left[w_{\overrightarrow{e}_L}\sigma^t_{\overrightarrow{e}_L}\left(z_{w_{\overrightarrow{e}_{L-1}}}\rho^{-1}_{w_{\overrightarrow{e}_{L-1}}}\nu^t_{\overrightarrow{e}_{L-1}} + Z^t_{\overrightarrow{e}_{L-1}}, w_{\overrightarrow{e}_L}\hat{\nu}^t_{\overleftarrow{e}_L} + Z^t_{\overleftarrow{e}_L}\right)\right] \tag{5.36}$$

$$\kappa^{s+1,t+1}_{\overrightarrow{e}_L} = \kappa^{t+1,s+1}_{\overrightarrow{e}_L} = \tag{5.37}$$

$$\delta_{n_{L-1}}\mathbb{E}\Bigg[\left(\sigma^s_{\overrightarrow{e}_L}\left(z_{w_{\overrightarrow{e}_{L-1}}}\rho^{-1}_{w_{\overrightarrow{e}_{L-1}}}\nu^s_{\overrightarrow{e}_{L-1}} + Z^s_{\overrightarrow{e}_{L-1}}, w_{\overrightarrow{e}_L}\hat{\nu}^s_{\overleftarrow{e}_L} + Z^s_{\overleftarrow{e}_L}\right) - w_{\overrightarrow{e}_L}\rho^{-1}_{w_{\overrightarrow{e}_L}}\nu^{s+1}_{\overrightarrow{e}_L}\right)$$
$$\left(\sigma^t_{\overrightarrow{e}_L}\left(z_{w_{\overrightarrow{e}_{L-1}}}\rho^{-1}_{w_{\overrightarrow{e}_{L-1}}}\nu^t_{\overrightarrow{e}_{L-1}} + Z^t_{\overrightarrow{e}_{L-1}}, w_{\overrightarrow{e}_L}\hat{\nu}^t_{\overleftarrow{e}_L} + Z^t_{\overleftarrow{e}_L}\right) - w_{\overrightarrow{e}_L}\rho^{-1}_{w_{\overrightarrow{e}_L}}\nu^{t+1}_{\overrightarrow{e}_L}\right)\Bigg] \tag{5.38}$$

$$\hat{\nu}^0_{\overleftarrow{e}_L}, \kappa^{1,1}_{\overleftarrow{e}_L} = \delta_{n_L}\mathbb{E}\left[\sigma^0_{\overleftarrow{e}_L}\left(z_{w_{\overrightarrow{e}_L}}, x^0_{\overrightarrow{e}_L}\right)\sigma^0_{\overleftarrow{e}_L}\left(z_{w_{\overrightarrow{e}_L}}\right)\right] \tag{5.39}$$

$$\hat{\nu}^{t+1}_{\overleftarrow{e}_L} = \delta_{n_L}\mathbb{E}\left[\frac{\partial\sigma^t_{\overleftarrow{e}_L,i}}{\partial z_{w_{\overleftarrow{e}_L}},i,\varphi_{\overleftarrow{e}_L}}\left(z_{w_{\overrightarrow{e}_L}}, z_{w_{\overrightarrow{e}_L}}\rho^{-1}_{w_{\overrightarrow{e}_L}}\nu^t_{\overrightarrow{e}_L} + Z^t_{\overrightarrow{e}_L}\right)\right] \tag{5.40}$$

$$\kappa^{s+1,t+1}_{\overleftarrow{e}_L} = \delta_{n_L}\mathbb{E}\Big[\sigma^s_{\overleftarrow{e}_L}\left(z_{w_{\overrightarrow{e}_L}}, z_{w_{\overrightarrow{e}_L}}\rho^{-1}_{w_{\overrightarrow{e}_L}}\nu^s_{\overrightarrow{e}_L} + Z^s_{\overrightarrow{e}_L}\right)$$
$$\sigma^t_{\overleftarrow{e}_L}\left(z_{w_{\overrightarrow{e}_L}}, z_{w_{\overrightarrow{e}_L}}\rho^{-1}_{w_{\overrightarrow{e}_L}}\nu^t_{\overrightarrow{e}_L} + Z^t_{\overrightarrow{e}_L}\right)\Big] \tag{5.41}$$

Lemma 23. *Under the assumptions of section 5.1.2, define, as above, independently for each $\overrightarrow{e}_l$, $Z^0_{\overrightarrow{e}_l} = x^0_{\overrightarrow{e}}$ and $(Z^1_{\overrightarrow{e}_l}, \dots, Z^t_{\overrightarrow{e}_l})$ a centered Gaussian random vector of covariance $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}_l})_{r,s\leqslant t}$ (and similarly for left-oriented edges). Then for any $1 \leqslant l \leqslant L$, for any sequence of uniformly (in n) pseudo-Lipschitz function $\Phi_l : (\mathbb{R}^{n_{l-1}\times(t+1)})^2 \to \mathbb{R}$*

$$\Phi\left(\left(\mathbf{x}^s_{\overrightarrow{e}_l}\right)_{0\leqslant s\leqslant t}, \left(\mathbf{x}^s_{\overleftarrow{e}_l}\right)_{0\leqslant s\leqslant t, \overleftarrow{e}_{l-1}\in\overleftarrow{E}}\right) \overset{\mathrm{P}}{\simeq}$$
$$\mathbb{E}\left[\Phi\left(\left(z_{w_{\overrightarrow{e}_l}}\rho^{-1}_{w_{\overrightarrow{e}_l}}\nu^s_{\overrightarrow{e}_l} + Z^s_{\overrightarrow{e}_l}\right)_{0\leqslant s\leqslant t, \overleftarrow{e}_l\in\overleftarrow{E}}, \left(w_{\overrightarrow{e}_{l-1}}\hat{\nu}^s_{\overleftarrow{e}_{l-1}} + Z^s_{\overleftarrow{e}_{l-1}}\right)_{0\leqslant s\leqslant t}\right)\right]$$

Proof. Consider the following iteration, corresponding to the algorithm presented in the previous section Eq.(5.1) with $q = 1$ indexed on the same graph as above, but where the matrices $\mathbf{A}_{\overrightarrow{e}_l}$ are replaced with random convolutional ones, denoted $\hat{\mathbf{A}}_{\overrightarrow{e}_l}$ such that

$$\forall\ \overrightarrow{e} \in \overrightarrow{E}\ \hat{\mathbf{A}}_{\overrightarrow{e}_l} \sim \mathcal{M}(D_{\overrightarrow{e}_l}, P_{\overrightarrow{e}_l}, k_{\overrightarrow{e}_l}, q_{\overrightarrow{e}_l}) \tag{5.42}$$

where $\mathbf{A}_{\overrightarrow{e}_l} \in \mathbb{R}^{D_{\overrightarrow{e}_l}q_{\overrightarrow{e}_l}\times P_{\overrightarrow{e}_l}q_{\overrightarrow{e}_l}}$, and we remind that we chose variances of $1/N$. Since we assume that $q = 1$, thus the Onsager terms are scalars, which we denote with lowercase letters $b^t_{\overrightarrow{e}}$. The

corresponding iteration then reads:

$$
\begin{aligned}
\mathbf{x}_{\overrightarrow{e_1}}^{t+1} &= \hat{\mathbf{A}}_{\overrightarrow{e_1}}\mathbf{m}_{\overrightarrow{e_1}}^{t} - b_{\overrightarrow{e_1}}^{t}\mathbf{m}_{\overleftarrow{e_1}}^{t-1}\,,\\
\mathbf{m}_{\overrightarrow{e_1}}^{t} &= f_{\overrightarrow{e}_1}^{t}\left(\mathbf{x}_{\overleftarrow{e_1}}^{t}\right)\,,\\
\mathbf{x}_{\overleftarrow{e_1}}^{t+1} &= \hat{\mathbf{A}}_{\overrightarrow{e_1}}^{\top}\mathbf{m}_{\overleftarrow{e_1}}^{t} - b_{\overleftarrow{e_1}}^{t}\mathbf{m}_{\overrightarrow{e_1}}^{t-1}\,,\\
\mathbf{m}_{\overleftarrow{e_1}}^{t} &= f_{\overleftarrow{e_1}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_1}\mathbf{w}_{\overrightarrow{e}_1}, \mathbf{x}_{\overrightarrow{e_1}}^{t}, \mathbf{x}_{\overleftarrow{e_2}}^{t}\right)\,,\\
&\\
\mathbf{x}_{\overrightarrow{e_2}}^{t+1} &= \hat{\mathbf{A}}_{\overrightarrow{e_2}}\mathbf{m}_{\overrightarrow{e_2}}^{t} - b_{\overrightarrow{e_2}}^{t}\mathbf{m}_{\overleftarrow{e_2}}^{t-1}\,,\\
\mathbf{m}_{\overrightarrow{e_2}}^{t} &= f_{\overrightarrow{e}_2}^{t}\left(\mathbf{x}_{\overrightarrow{e_1}}^{t}, \mathbf{x}_{\overleftarrow{e_2}}^{t}\right)\,,\\
\mathbf{x}_{\overleftarrow{e_2}}^{t+1} &= \hat{\mathbf{A}}_{\overrightarrow{e_2}}^{\top}\mathbf{m}_{\overleftarrow{e_2}}^{t} - b_{\overleftarrow{e_2}}^{t}\mathbf{m}_{\overrightarrow{e_2}}^{t-1}\,,\\
\mathbf{m}_{\overleftarrow{e_2}}^{t} &= f_{\overleftarrow{e_2}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_2}\mathbf{w}_{\overrightarrow{e}_2}, \mathbf{x}_{\overrightarrow{e_2}}^{t}, \mathbf{x}_{\overleftarrow{e_3}}^{t}\right)\,,\\
&\\
&\vdots\\
&\\
\mathbf{x}_{\overrightarrow{e_L}}^{t+1} &= \hat{\mathbf{A}}_{\overrightarrow{e_L}}\mathbf{m}_{\overrightarrow{e_L}}^{t} - b_{\overrightarrow{e_L}}^{t}\mathbf{m}_{\overleftarrow{e_L}}^{t-1}\,,\\
\mathbf{m}_{\overrightarrow{e_L}}^{t} &= f_{\overrightarrow{e}_L}^{t}\left(\mathbf{x}_{\overrightarrow{e}_{L-1}}^{t}, \mathbf{x}_{\overleftarrow{e_L}}^{t}\right)\,,\\
\mathbf{x}_{\overleftarrow{e_L}}^{t+1} &= \hat{\mathbf{A}}_{\overrightarrow{e_L}}^{\top}\mathbf{m}_{\overleftarrow{e_L}}^{t} - b_{\overleftarrow{e_L}}^{t}\mathbf{m}_{\overrightarrow{e_L}}^{t-1}\,,\\
\mathbf{m}_{\overleftarrow{e_L}}^{t} &= f_{\overleftarrow{e_L}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_L}\mathbf{w}_{\overrightarrow{e}_L}, \mathbf{x}_{\overrightarrow{e_L}}^{t}\right)
\end{aligned}
\tag{5.43}
$$

Then, according to Lemma 22, for any $1 \leqslant l \leqslant L$, there exists a pair of orthogonal matrices $\mathbf{U}_{\overrightarrow{e}_l} \in \mathbb{R}^{D_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l} \times D_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}, \mathbf{V}_{\overrightarrow{e}_l} \in \mathbb{R}^{P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l} \times P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}$ such that $\hat{\mathbf{A}}_{\overrightarrow{e}_l} = \mathbf{U}_{\overrightarrow{e}_l}\tilde{\mathbf{A}}_{\overrightarrow{e}_l}\mathbf{V}_{\overrightarrow{e}_l}^{\top}$ and $\tilde{\mathbf{A}}_{\overrightarrow{e}_l} = \left[\left(\mathcal{P}_{P_{\overrightarrow{e}_l}, q_{\overrightarrow{e}_l}}\right)^{i-1}\mathbf{Q}_{\overrightarrow{e}_l}\right]_{i=1}^{q_{\overrightarrow{e}_l}}$, where $\mathbf{Q}_{\overrightarrow{e}_l} \in \mathbb{R}^{D_{\overrightarrow{e}_l} \times P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}$ is composed of $q_{\overrightarrow{e}_l}$ blocks of size $D_{\overrightarrow{e}_l} \times P_{\overrightarrow{e}_l}$, denoted $\mathbf{Q}_{\overrightarrow{e}_l}^{j}$, verifying

- for any $1 \leqslant j \leqslant k_{\overrightarrow{e}}$, $\mathbf{Q}_{\overrightarrow{e}}^{j}$ has i.i.d. $\mathcal{N}(0, \frac{1}{N})$ elements
- for any $k_{\overrightarrow{e}} < j \leqslant q_{\overrightarrow{e}}$, all elements of $\mathbf{Q}_{\overrightarrow{e}}^{j}$ are zero.

In the preceding definition of $\tilde{\mathbf{A}}_{\overrightarrow{e}_l}$, $\mathbf{Q}_{\overrightarrow{e}_l}$ is understood as a vector of size $\mathbb{R}^{P_{\overrightarrow{e}} q_{\overrightarrow{e}}}$ with elements in $\mathbb{R}^{D_{\overrightarrow{e}}}$, such that the permutation matrix $\mathcal{P}_{P_{\overrightarrow{e}}, q_{\overrightarrow{e}}}$ shifts blocks of size $D_{\overrightarrow{e}} \times P_{\overrightarrow{e}}$, yielding

$$
\tilde{\mathbf{A}}_{\overrightarrow{e}} = \begin{bmatrix}
\mathbf{Q}_{\overrightarrow{e}_l}^{(1)} & \mathbf{Q}_{\overrightarrow{e}_l}^{(2)} & \dots & \mathbf{Q}_{\overrightarrow{e}_l}^{(k_{\overrightarrow{e}})} & & & \\
 & \mathbf{Q}_{\overrightarrow{e}_l}^{(1)} & \mathbf{Q}_{\overrightarrow{e}_l}^{(2)} & \dots & \mathbf{Q}_{\overrightarrow{e}_l}^{(k_{\overrightarrow{e}})} & & \\
 & & \mathbf{Q}_{\overrightarrow{e}_l}^{(1)} & \mathbf{Q}_{\overrightarrow{e}_l}^{(2)} & \dots & \mathbf{Q}_{\overrightarrow{e}_l}^{(k_{\overrightarrow{e}})} & \vdots \\
\vdots & \vdots & \ddots & & & & \\
\mathbf{Q}_{\overrightarrow{e}_l}^{(2)} & \mathbf{Q}_{\overrightarrow{e}_l}^{(3)} & \dots & \mathbf{Q}_{\overrightarrow{e}_l}^{(k_{\overrightarrow{e}})} & & & \mathbf{Q}_{\overrightarrow{e}_l}^{(1)}
\end{bmatrix}
\tag{5.44}
$$

The iteration then reads

$$
\begin{aligned}
\mathbf{x}_{\overrightarrow{e_1}}^{t+1} &= \mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{A}}_{\overrightarrow{e}_1}\mathbf{V}_{\overrightarrow{e}_1}^{\top}\mathbf{m}_{\overrightarrow{e_1}}^{t} - b_{\overrightarrow{e_1}}^{t}\mathbf{m}_{\overleftarrow{e_1}}^{t-1}\,,\\
\mathbf{m}_{\overrightarrow{e_1}}^{t} &= f_{\overrightarrow{e}_1}^{t}\left(\mathbf{x}_{\overleftarrow{e_1}}^{t}\right)\,,\\
\mathbf{x}_{\overleftarrow{e_1}}^{t+1} &= \mathbf{V}_{\overrightarrow{e}_1}\tilde{\mathbf{A}}_{\overrightarrow{e}_1}^{\top}\mathbf{U}_{\overrightarrow{e}_1}^{\top}\mathbf{m}_{\overleftarrow{e_1}}^{t} - b_{\overleftarrow{e_1}}^{t}\mathbf{m}_{\overrightarrow{e_1}}^{t-1}\,,\\
\mathbf{m}_{\overleftarrow{e_1}}^{t} &= f_{\overleftarrow{e_1}}^{t}\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{A}}_{\overrightarrow{e}_1}\mathbf{V}_{\overrightarrow{e}_1}^{\top}\mathbf{w}_{\overrightarrow{e}_1}, \mathbf{x}_{\overrightarrow{e_1}}^{t}, \mathbf{x}_{\overleftarrow{e_2}}^{t}\right)\,,\\
\\
\mathbf{x}_{\overrightarrow{e_2}}^{t+1} &= \mathbf{U}_{\overrightarrow{e}_2}\tilde{\mathbf{A}}_{\overrightarrow{e}_2}\mathbf{V}_{\overrightarrow{e}_2}^{\top}\mathbf{m}_{\overrightarrow{e_2}}^{t} - b_{\overrightarrow{e_2}}^{t}\mathbf{m}_{\overleftarrow{e_2}}^{t-1}\,,\\
\mathbf{m}_{\overrightarrow{e_2}}^{t} &= f_{\overrightarrow{e}_2}^{t}\left(\mathbf{x}_{\overrightarrow{e_1}}^{t}, \mathbf{x}_{\overleftarrow{e_2}}^{t}\right)\,,\\
\mathbf{x}_{\overleftarrow{e_2}}^{t+1} &= \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{A}}_{\overrightarrow{e}}^{\top}\mathbf{U}_{\overrightarrow{e}_2}^{\top}\mathbf{m}_{\overleftarrow{e_2}}^{t} - b_{\overleftarrow{e_2}}^{t}\mathbf{m}_{\overrightarrow{e_2}}^{t-1}\,,\\
\mathbf{m}_{\overleftarrow{e_2}}^{t} &= f_{\overleftarrow{e_2}}^{t}\left(\mathbf{U}_{\overrightarrow{e}_2}\tilde{\mathbf{A}}_{\overrightarrow{e}_2}\mathbf{V}_{\overrightarrow{e}_2}^{\top}\mathbf{w}_{\overrightarrow{e}_2}, \mathbf{x}_{\overrightarrow{e_2}}^{t}, \mathbf{x}_{\overleftarrow{e_3}}^{t}\right)\,,
\end{aligned}
\tag{5.45}
$$

$$
\vdots
$$

$$
\begin{aligned}
\mathbf{x}_{\overrightarrow{e_L}}^{t+1} &= \mathbf{U}_{\overrightarrow{e}_L}\tilde{\mathbf{A}}_{\overrightarrow{e}_L}\mathbf{V}_{\overrightarrow{e}_L}^{\top}\mathbf{m}_{\overrightarrow{e_L}}^{t} - b_{\overrightarrow{e_L}}^{t}\mathbf{m}_{\overleftarrow{e_L}}^{t-1}\,,\\
\mathbf{m}_{\overrightarrow{e_L}}^{t} &= f_{\overrightarrow{e}_L}^{t}\left(\mathbf{x}_{\overrightarrow{e}_{L-1}}^{t}, \mathbf{x}_{\overleftarrow{e_L}}^{t}\right)\,,\\
\mathbf{x}_{\overleftarrow{e_L}}^{t+1} &= \mathbf{V}_{\overrightarrow{e}_L}\tilde{\mathbf{A}}_{\overrightarrow{e}_L}^{\top}\mathbf{U}_{\overrightarrow{e}_L}^{\top}\mathbf{m}_{\overleftarrow{e_L}}^{t} - b_{\overleftarrow{e_L}}^{t}\mathbf{m}_{\overrightarrow{e_L}}^{t-1}\,,\\
\mathbf{m}_{\overleftarrow{e_L}}^{t} &= f_{\overleftarrow{e_L}}^{t}\left(\mathbf{U}_{\overrightarrow{e}_L}\tilde{\mathbf{A}}_{\overrightarrow{e}_L}\mathbf{V}_{\overrightarrow{e}_L}^{\top}\mathbf{w}_{\overrightarrow{e}_L}, \mathbf{x}_{\overrightarrow{e_L}}^{t}\right)
\end{aligned}
$$

Since we will not be making any change of variable on the $\mathbf{w}_{\overrightarrow{e}_l}$, we will keep the $\hat{\mathbf{A}}_{\overrightarrow{e}_l}$ notation for the quantities related to the planted model. Define, for any $1 \leqslant l \leqslant L$ and any $t \in \mathbb{N}$:

$$
\begin{aligned}
&\tilde{\mathbf{x}}_{\overrightarrow{e}_l} = \mathbf{U}_{\overrightarrow{e}_l}^{\top}\mathbf{x}_{\overrightarrow{e}_l} \qquad \tilde{\mathbf{x}}_{\overleftarrow{e}_l} = \mathbf{V}_{\overrightarrow{e}_l}^{\top}\mathbf{x}_{\overleftarrow{e}_l}\\
&\tilde{\mathbf{m}}_{\overrightarrow{e}_l}^{t} = \mathbf{V}_{\overrightarrow{e}_l}^{\top}\mathbf{m}_{\overrightarrow{e}_l}^{t} \qquad \tilde{\mathbf{m}}_{\overleftarrow{e}_l}^{t} = \mathbf{U}_{\overrightarrow{e}_l}^{\top}\mathbf{m}_{\overleftarrow{e}_l}^{t}\\
&\tilde{f}_{\overrightarrow{e}_1}^{t}(\tilde{\mathbf{x}}_{\overleftarrow{e_1}}^{t}) = \mathbf{V}_{\overrightarrow{e}_1}^{\top}f_{\overrightarrow{e}_1}^{t}\left(\mathbf{V}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}_{\overleftarrow{e_1}}^{t}\right)\\
&\tilde{f}_{\overleftarrow{e_1}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_1}\mathbf{w}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}_{\overrightarrow{e_1}}^{t}, \tilde{\mathbf{x}}_{\overleftarrow{e_2}}^{t}\right) = \mathbf{U}_{\overrightarrow{e}_1}^{\top}f_{\overleftarrow{e_1}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_1}\mathbf{w}_{\overrightarrow{e}_1}, \mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}_{\overrightarrow{e_1}}^{t}, \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}_{\overleftarrow{e_2}}^{t}\right)\\
&\tilde{f}_{\overrightarrow{e}_2}^{t}\left(\tilde{\mathbf{x}}_{\overrightarrow{e_1}}^{t}, \tilde{\mathbf{x}}_{\overleftarrow{e_2}}^{t}\right) = \mathbf{V}_{\overrightarrow{e}_2}^{\top}f_{\overrightarrow{e}_2}^{t}\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}_{\overrightarrow{e_1}}^{t}, \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}_{\overleftarrow{e_2}}^{t}\right)\\
&\tilde{f}_{\overleftarrow{e_2}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_2}\mathbf{w}_{\overrightarrow{e}_2}, \tilde{\mathbf{x}}_{\overrightarrow{e_2}}^{t}, \tilde{\mathbf{x}}_{\overleftarrow{e_3}}^{t}\right) = \mathbf{U}_{\overrightarrow{e}_2}^{\top}f_{\overleftarrow{e_2}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_2}\mathbf{w}_{\overrightarrow{e}_2}, \mathbf{U}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}_{\overrightarrow{e_2}}^{t}, \mathbf{V}_{\overrightarrow{e}_3}\tilde{\mathbf{x}}_{\overleftarrow{e_3}}^{t}\right)
\end{aligned}
$$

$$
\vdots
$$

$$
\begin{aligned}
&\tilde{f}_{\overrightarrow{e}_L}^{t}\left(\tilde{\mathbf{x}}_{\overrightarrow{e}_{L-1}}^{t}, \tilde{\mathbf{x}}_{\overleftarrow{e_L}}^{t}\right) = \mathbf{V}_{\overrightarrow{e}_L}^{\top}f_{\overrightarrow{e}_L}^{t}\left(\mathbf{U}_{\overrightarrow{e}_{L-1}}\tilde{\mathbf{x}}_{\overrightarrow{e}_{L-1}}^{t}, \mathbf{V}_{\overrightarrow{e}_L}\tilde{\mathbf{x}}_{\overleftarrow{e_L}}^{t}\right)\\
&\tilde{f}_{\overleftarrow{e_L}}^{t}\left(\hat{\mathbf{A}}_{\overrightarrow{e}_L}\mathbf{w}_{\overrightarrow{e}_L}, \tilde{\mathbf{x}}_{\overrightarrow{e_L}}^{t}\right) = \mathbf{U}_{\overrightarrow{e}_L}^{\top}f_{\overleftarrow{e_L}}^{t}\left(\mathbf{U}_{\overrightarrow{e}_L}\tilde{\mathbf{A}}_{\overrightarrow{e}_L}\mathbf{V}_{\overrightarrow{e}_L}\mathbf{w}_{\overrightarrow{e}_L}, \mathbf{U}_{\overrightarrow{e}_L}\tilde{\mathbf{x}}_{\overrightarrow{e_L}}^{t}\right)
\end{aligned}
$$

Using the orthogonality of the permutation matrices $\mathbf{U}_{\overrightarrow{e}}, \mathbf{V}_{\overrightarrow{e}}$, the iteration may be rewritten

$$
\begin{aligned}
\tilde{\mathbf{x}}^{t+1}_{\overrightarrow{e_1}} &= \tilde{\mathbf{A}}_{\overrightarrow{e}_1} \tilde{\mathbf{m}}^{t}_{\overrightarrow{e_1}} - b^{t}_{\overrightarrow{e_1}} \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_1}} \,, \\
\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_1}} &= \tilde{f}^{t}_{\overrightarrow{e}_1}(\tilde{\mathbf{x}}^{t}_{\overleftarrow{e_1}}) \,, \\
\tilde{\mathbf{x}}^{t+1}_{\overleftarrow{e_1}} &= \tilde{\mathbf{A}}^{\top}_{\overrightarrow{e}_1} \tilde{\mathbf{m}}^{t}_{\overleftarrow{e_1}} - b^{t}_{\overleftarrow{e_1}} \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_1}} \,, \\
\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_1}} &= \tilde{f}^{t}_{\overleftarrow{e_1}} \left(\hat{\mathbf{A}}_{\overrightarrow{e}_1} \mathbf{w}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}^{t}_{\overrightarrow{e_1}}, \tilde{\mathbf{x}}^{t}_{\overleftarrow{e_2}} \right) \,, \\
\\
\tilde{\mathbf{x}}^{t+1}_{\overrightarrow{e_2}} &= \tilde{\mathbf{A}}_{\overrightarrow{e}_2} \tilde{\mathbf{m}}^{t}_{\overrightarrow{e_2}} - b^{t}_{\overrightarrow{e_2}} \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_2}} \,, \\
\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_2}} &= \tilde{f}^{t}_{\overrightarrow{e}_2} \left(\tilde{\mathbf{x}}^{t}_{\overrightarrow{e_1}}, \tilde{\mathbf{x}}^{t}_{\overleftarrow{e_2}} \right) \,, \\
\tilde{\mathbf{x}}^{t+1}_{\overleftarrow{e_2}} &= \tilde{\mathbf{A}}^{\top}_{\overrightarrow{e}} \tilde{\mathbf{m}}^{t}_{\overleftarrow{e_2}} - b^{t}_{\overleftarrow{e_2}} \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_2}} \,, \\
\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_2}} &= \tilde{f}^{t}_{\overleftarrow{e_2}} \left(\hat{\mathbf{A}}_{\overrightarrow{e}_2} \mathbf{w}_{\overrightarrow{e}_2}, \tilde{\mathbf{x}}^{t}_{\overrightarrow{e_2}}, \tilde{\mathbf{x}}^{t}_{\overleftarrow{e_3}} \right) \,, \\
\\
&\vdots \\
\\
\tilde{\mathbf{x}}^{t+1}_{\overrightarrow{e_L}} &= \tilde{\mathbf{A}}_{\overrightarrow{e}_L} \tilde{\mathbf{m}}^{t}_{\overrightarrow{e_L}} - b^{t}_{\overrightarrow{e_L}} \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_L}} \,, \\
\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_L}} &= \tilde{f}^{t}_{\overrightarrow{e}_L} \left(\tilde{\mathbf{x}}^{t}_{\overrightarrow{e}_{L-1}}, \tilde{\mathbf{x}}^{t}_{\overleftarrow{e_L}} \right) \,, \\
\tilde{\mathbf{x}}^{t+1}_{\overleftarrow{e_L}} &= \tilde{\mathbf{A}}^{\top}_{\overrightarrow{e}_L} \tilde{\mathbf{m}}^{t}_{\overleftarrow{e_L}} - b^{t}_{\overleftarrow{e_L}} \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_L}} \,, \\
\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_L}} &= \tilde{f}^{t}_{\overleftarrow{e_L}} \left(\hat{\mathbf{A}}_{\overrightarrow{e}_L} \mathbf{w}_{\overrightarrow{e}_L}, \tilde{\mathbf{x}}^{t}_{\overrightarrow{e_L}} \right)
\end{aligned}
\tag{5.46}
$$

Recall, for any $1 \leqslant l \leqslant L$, the dimensions $\tilde{\mathbf{A}}_{\overrightarrow{e}_l} \in \mathbb{R}^{D_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l} \times P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}$ and $\tilde{f}^{t}_{\overrightarrow{e}_l}(...) \in \mathbb{R}^{P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}$. Consider then

$$
\tilde{f}^{t}_{\overrightarrow{e}_l}(...) = \begin{bmatrix} \left(\tilde{f}^{t}_{\overrightarrow{e}_l} \right)^{(1)}(...) \\ \vdots \\ \left(\tilde{f}^{t}_{\overrightarrow{e}_l} \right)^{(q_{\overrightarrow{e}_l})}(...) \end{bmatrix}
\tag{5.47}
$$

where, for any $1 \leqslant k \leqslant q_{\overrightarrow{e}_l}$, $(\tilde{f}^{t}_{\overrightarrow{e}_l})^{(k)}(...) \in \mathbb{R}^{P_{\overrightarrow{e}}}$. The product $\tilde{\mathbf{A}}_{\overrightarrow{e}_l} \tilde{f}^{t}_{\overrightarrow{e}_l}(...) \in \mathbb{R}^{D_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}$ then reads,

using the circulant structure of $\tilde{\mathbf{A}}_{\overrightarrow{\ell}_l}$

$$\begin{bmatrix} \mathbf{Q}^{(1)}_{\overrightarrow{\ell}_l} & \mathbf{Q}^{(2)}_{\overrightarrow{\ell}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{\ell}})}_{\overrightarrow{\ell}_l} & & & \\ & \mathbf{Q}^{(1)}_{\overrightarrow{\ell}_l} & \mathbf{Q}^{(2)}_{\overrightarrow{\ell}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{\ell}})}_{\overrightarrow{\ell}_l} & & \\ & & \mathbf{Q}^{(1)}_{\overrightarrow{\ell}_l} & \mathbf{Q}^{(2)}_{\overrightarrow{\ell}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{\ell}})}_{\overrightarrow{\ell}_l} & \vdots \\ \vdots & \vdots & \ddots & & & & \\ \mathbf{Q}^{(2)}_{\overrightarrow{\ell}_l} & \mathbf{Q}^{(3)}_{\overrightarrow{\ell}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{\ell}})}_{\overrightarrow{\ell}_l} & & & \mathbf{Q}^{(1)}_{\overrightarrow{\ell}_l} \end{bmatrix} \begin{bmatrix} \left(\tilde{f}^t_{\overrightarrow{\ell}_l}\right)^{(1)}(...) \\ \vdots \\ \left(\tilde{f}^t_{\overrightarrow{\ell}_l}\right)^{(q_{\overrightarrow{\ell}_l})}(...) \end{bmatrix} \tag{5.48}$$

$$= \left[\left(\left(\mathcal{P}_{P_{\overrightarrow{\ell}_l}, q_{\overrightarrow{\ell}_l}} \right)^{i-1} \mathbf{Q}_{\overrightarrow{\ell}_l} \right) \tilde{f}^t_{\overrightarrow{\ell}_l}(...) \right]_{i=1}^{q_{\overrightarrow{\ell}_l}} \tag{5.49}$$

$$= \left[\sum_{j=1}^{k_{\overrightarrow{\ell}_l}} \mathbf{Q}^{(j)}_{\overrightarrow{\ell}_l} (\tilde{f}^t_{\overrightarrow{\ell}_l})^{(\lfloor j+n-2 \rfloor_{q_{\overrightarrow{\ell}_l}}+1)}(...) \right]_{n=1}^{q_{\overrightarrow{\ell}_l}} \tag{5.50}$$

where the notation $\lfloor . \rfloor_{q_{\overrightarrow{\ell}_l}}$ denotes the modulo $q_{\overrightarrow{\ell}_l}$, i.e. the remainder of the euclidian division by $q_{\overrightarrow{\ell}_l}$. Now define

$$\tilde{F}^t_{\overrightarrow{\ell}_l}(...) = \begin{bmatrix} \left[\left(\mathcal{P}_{P_{\overrightarrow{\ell}_l}, q_{\overrightarrow{\ell}_l}} \right)^{1-i} \left[(\tilde{f}^t_{\overrightarrow{\ell}_l})^{(1)} \dots (\tilde{f}^t_{\overrightarrow{\ell}_l})^{(q_{\overrightarrow{\ell}_l})} \right] \right]_{i=1}^{k_{\overrightarrow{\ell}_l}} \in \mathbb{R}^{P_{\overrightarrow{\ell}_l} k_{\overrightarrow{\ell}_l} \times q_{\overrightarrow{\ell}_l}} \\ \left[0_{P_{\overrightarrow{\ell}_l}} \dots 0_{P_{\overrightarrow{\ell}_l}} \right]_{j=1}^{q_{\overrightarrow{\ell}_l} - k_{\overrightarrow{\ell}_l}} \end{bmatrix} \in \mathbb{R}^{P_{\overrightarrow{\ell}_l} q_{\overrightarrow{\ell}_l} \times q_{\overrightarrow{\ell}_l}} \tag{5.51}$$

and the matrix $\tilde{\mathbf{Q}}_{\overrightarrow{\ell}_l} \in \mathbb{R}^{D_{\overrightarrow{\ell}_l} q_{\overrightarrow{\ell}_l} \times P_{\overrightarrow{\ell}_l} q_{\overrightarrow{\ell}_l}}$ is a dense Gaussian matrix with i.i.d. elements. Then

$$\tilde{\mathbf{Q}}_{\overrightarrow{\ell}_l} \tilde{F}^t_{\overrightarrow{\ell}_l}(...) = \begin{bmatrix} \sum_{j=1}^{k_{\overrightarrow{\ell}_l}} \mathbf{Q}^{(j)}_{\overrightarrow{\ell}_l} (\tilde{f}^t_{\overrightarrow{\ell}_l})^{\lfloor j-1 \rfloor_{q_{\overrightarrow{\ell}_l}}+1}(...) & \dots & \sum_{j=1}^{k_{\overrightarrow{\ell}_l}} (\mathbf{Q}^{(j)}_{\overrightarrow{\ell}_l}) (\tilde{f}^t_{\overrightarrow{\ell}_l})^{\lfloor j+q_{\overrightarrow{\ell}_l}-2 \rfloor_{q_{\overrightarrow{\ell}_l}}+1}(...) \\ & \dots & \\ & \dots & \end{bmatrix} \\ \in \mathbb{R}^{D_{\overrightarrow{\ell}_l} q_{\overrightarrow{\ell}_l} \times q_{\overrightarrow{\ell}_l}}$$

where each $\dots$ is an identical copy of the first $D_{\overrightarrow{\ell}_l} \times q_{\overrightarrow{\ell}_l}$ block, for a total of $k_{\overrightarrow{\ell}_l}$ blocks. This means the $D_{\overrightarrow{\ell}_l} q_{\overrightarrow{\ell}_l}$ output of the product $\tilde{\mathbf{A}}_{\overrightarrow{\ell}_l} f^t_{\overrightarrow{\ell}_l}(...)$ may be rewritten as a $D_{\overrightarrow{\ell}_l} \times q_{\overrightarrow{\ell}_l}$ matrix (copied $k_{\overrightarrow{\ell}_l}$ times) resulting from the product of a dense Gaussian matrix with i.i.d. elements and a matrix valued function $\tilde{F}^t_{\overrightarrow{\ell}_l}$ which verifies the same regularity conditions as $f^t_{\overrightarrow{\ell}_l}$. Note that, owing to the separability assumption, we may use any permutation of the $(\tilde{f}^t_{\overrightarrow{\ell}_l})^{(i)}, 1 \leqslant i \leqslant q_{\overrightarrow{\ell}_l}$ and will thus drop the permutations to write

$$\tilde{F}^t_{\overrightarrow{\ell}_l}(...) = \begin{bmatrix} \left[(\tilde{f}^t_{\overrightarrow{\ell}_l})^{(1)} \dots (\tilde{f}^t_{\overrightarrow{\ell}_l})^{(q_{\overrightarrow{\ell}_l})} \right]_{i=1}^{k_{\overrightarrow{\ell}_l}} \in \mathbb{R}^{P_{\overrightarrow{\ell}_l} k_{\overrightarrow{\ell}_l} \times q_{\overrightarrow{\ell}_l}} \\ \left[0_{P_{\overrightarrow{\ell}_l}} \dots 0_{P_{\overrightarrow{\ell}_l}} \right]_{j=1}^{q_{\overrightarrow{\ell}_l} - k_{\overrightarrow{\ell}_l}} \end{bmatrix} \in \mathbb{R}^{P_{\overrightarrow{\ell}_l} q_{\overrightarrow{\ell}_l} \times q_{\overrightarrow{\ell}_l}} \tag{5.52}$$

Similarly, for products of the form $\left(\tilde{\mathbf{A}}_{\overrightarrow{e}_l}\right)^\top \tilde{f}^t_{\overleftarrow{e}_l}(...) \in \mathbb{R}^{P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}$, we may write:

$$\begin{bmatrix} \mathbf{Q}^{(1)}_{\overrightarrow{e}_l} & \mathbf{Q}^{(2)}_{\overrightarrow{e}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{e}_l})}_{\overrightarrow{e}_l} & & & \\ & \mathbf{Q}^{(1)}_{\overrightarrow{e}_l} & \mathbf{Q}^{(2)}_{\overrightarrow{e}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{e}_l})}_{\overrightarrow{e}_l} & & \\ & & \mathbf{Q}^{(1)}_{\overrightarrow{e}_l} & \mathbf{Q}^{(2)}_{\overrightarrow{e}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{e}_l})}_{\overrightarrow{e}_l} & \vdots \\ \vdots & \vdots & \ddots & & & & \\ \mathbf{Q}^{(2)}_{\overrightarrow{e}_l} & \mathbf{Q}^{(3)}_{\overrightarrow{e}_l} & \dots & \mathbf{Q}^{(k_{\overrightarrow{e}_l})}_{\overrightarrow{e}_l} & & & \mathbf{Q}^{(1)}_{\overrightarrow{e}_l} \end{bmatrix}^\top \begin{bmatrix} \left(\tilde{f}^t_{\overleftarrow{e}_l}\right)^{(1)}(...) \\ \vdots \\ \left(\tilde{f}^t_{\overleftarrow{e}_l}\right)^{(q_{\overrightarrow{e}_l})}(...) \end{bmatrix} \tag{5.53}$$

$$= \left[\left(\left(\mathcal{P}_{P_{\overrightarrow{e}_l}, q_{\overrightarrow{e}_l}}\right)^{i-1} \left[(\mathbf{Q}^{(1)}_{\overrightarrow{e}_l})^\top (0 \dots 0)(\mathbf{Q}^{(k_{\overrightarrow{e}_l})}_{\overrightarrow{e}_l})^\top \dots (\mathbf{Q}^{(2)}_{\overrightarrow{e}_l})^\top\right]\right) \tilde{f}^t_{\overleftarrow{e}_l}(...)\right]_{i=1}^{q_{\overrightarrow{e}_l}} \tag{5.54}$$

Then, using once again the separability assumption, we may define:

$$\tilde{F}^t_{\overleftarrow{e}_l}(...) = \begin{bmatrix} \left[(\tilde{f}^t_{\overleftarrow{e}_l})^{(1)} \dots (\tilde{f}^t_{\overleftarrow{e}_l})^{(q_{\overleftarrow{e}_l})}\right]_{i=1}^{k_{\overrightarrow{e}_l}} \in \mathbb{R}^{D_{\overrightarrow{e}_l} k_{\overrightarrow{e}_l} \times q_{\overrightarrow{e}_l}} \\ \left[0_{D_{\overrightarrow{e}_l}} \dots 0_{D_{\overrightarrow{e}_l}}\right] \end{bmatrix} \in \mathbb{R}^{D_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l} \times q_{\overrightarrow{e}_l}} \tag{5.55}$$

such that the term $\tilde{\mathbf{Q}}^\top_{\overrightarrow{e}_l} \tilde{F}^t_{\overleftarrow{e}_l}(...)$ also contains $k_{\overrightarrow{e}_l}$ copies of a $P_{\overrightarrow{e}_l} \times q_{\overrightarrow{e}_l}$ block containing the $q_{\overrightarrow{e}_l}$ blocks of size $P_{\overrightarrow{e}_l}$ of the original $P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}$ vector $\tilde{\mathbf{A}}^\top_{\overrightarrow{e}_l} \tilde{f}_{\overleftarrow{e}_l}(...)$. The iterates of the sequences defined by Eq.(5.46) may then be rewritten as a subset of the lines of the following matrix valued iteration, i.e.:

$$\begin{aligned}
\tilde{\mathbf{X}}^{t+1}_{\overrightarrow{e_1}} &= \tilde{\mathbf{Q}}_{\overrightarrow{e}_1} \tilde{\mathbf{m}}^t_{\overrightarrow{e_1}} - b^t_{\overrightarrow{e_1}} \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_1}}, \\
\tilde{\mathbf{m}}^t_{\overrightarrow{e_1}} &= \tilde{F}^t_{\overrightarrow{e}_1}(\tilde{\mathbf{X}}^t_{\overleftarrow{e_1}}), \\
\tilde{\mathbf{X}}^{t+1}_{\overleftarrow{e_1}} &= \tilde{\mathbf{Q}}^\top_{\overrightarrow{e}_1} \tilde{\mathbf{m}}^t_{\overleftarrow{e_1}} - b^t_{\overleftarrow{e_1}} \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_1}}, \\
\tilde{\mathbf{m}}^t_{\overleftarrow{e_1}} &= \tilde{F}^t_{\overleftarrow{e_1}}\left(\tilde{\mathbf{Q}}_{\overrightarrow{e}_1} \mathbf{W}_{\overrightarrow{e}_1}, \tilde{\mathbf{X}}^t_{\overrightarrow{e_1}}, \tilde{\mathbf{X}}^t_{\overleftarrow{e_2}}\right), \\
\\
\tilde{\mathbf{X}}^{t+1}_{\overrightarrow{e_2}} &= \tilde{\mathbf{Q}}_{\overrightarrow{e}_2} \tilde{\mathbf{m}}^t_{\overrightarrow{e_2}} - b^t_{\overrightarrow{e_2}} \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_2}}, \\
\tilde{\mathbf{m}}^t_{\overrightarrow{e_2}} &= \tilde{F}^t_{\overrightarrow{e}_2}\left(\tilde{\mathbf{X}}^t_{\overrightarrow{e_1}}, \tilde{\mathbf{X}}^t_{\overleftarrow{e_2}}\right), \\
\tilde{\mathbf{X}}^{t+1}_{\overleftarrow{e_2}} &= \tilde{\mathbf{Q}}^\top_{\overrightarrow{e}} \tilde{\mathbf{m}}^t_{\overleftarrow{e_2}} - b^t_{\overleftarrow{e_2}} \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_2}}, \\
\tilde{\mathbf{m}}^t_{\overleftarrow{e_2}} &= \tilde{F}^t_{\overleftarrow{e_2}}\left(\tilde{\mathbf{Q}}_{\overrightarrow{e}_2} \mathbf{W}_{\overrightarrow{e}_2}, \tilde{\mathbf{X}}^t_{\overrightarrow{e_2}}, \tilde{\mathbf{X}}^t_{\overleftarrow{e_3}}\right), \\
&\vdots \\
\tilde{\mathbf{X}}^{t+1}_{\overrightarrow{e_L}} &= \tilde{\mathbf{Q}}_{\overrightarrow{e}_L} \tilde{\mathbf{m}}^t_{\overrightarrow{e_L}} - b^t_{\overrightarrow{e_L}} \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_L}}, \\
\tilde{\mathbf{m}}^t_{\overrightarrow{e_L}} &= \tilde{F}^t_{\overrightarrow{e}_L}\left(\tilde{\mathbf{X}}^t_{\overrightarrow{e}_{L-1}}, \tilde{\mathbf{X}}^t_{\overleftarrow{e_L}}\right), \\
\tilde{\mathbf{X}}^{t+1}_{\overleftarrow{e_L}} &= \tilde{\mathbf{Q}}^\top_{\overrightarrow{e}_L} \tilde{\mathbf{m}}^t_{\overleftarrow{e_L}} - b^t_{\overleftarrow{e_L}} \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_L}}, \\
\tilde{\mathbf{m}}^t_{\overleftarrow{e_L}} &= \tilde{F}^t_{\overleftarrow{e_L}}\left(\tilde{\mathbf{Q}}_{\overrightarrow{e}_L} \mathbf{W}_{\overrightarrow{e}_L}, \tilde{\mathbf{X}}^t_{\overrightarrow{e_L}}\right)
\end{aligned} \tag{5.56}$$

where each $\mathbf{W}_{\overrightarrow{e}_l}$ contains $k_{\overrightarrow{e}_l}$ copies of the initial $\mathbf{w}_{\overrightarrow{e}_l}$ reorganised into matrices as described above. The dimensions of the variables are Note that at this point we have almost reached an iteration verifying the structure of that appearing in Theorem 8, except the Onsager term isn't, a priori, the correct one. Consider the following iteration, where we replaced the original, scalar Onsager terms with the correct, matrix-valued ones:

$$\begin{aligned}
\tilde{\mathbf{X}}^{t+1}_{\overrightarrow{e_1}} &= \tilde{\mathbf{Q}}_{\overrightarrow{e}_1}\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_1}} - \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_1}}\left(\tilde{\mathbf{b}}^{t}_{\overrightarrow{e_1}}\right)^{\top}, \\
\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_1}} &= \tilde{F}^{t}_{\overrightarrow{e}_1}(\tilde{\mathbf{X}}^{t}_{\overleftarrow{e_1}}), \\
\tilde{\mathbf{X}}^{t+1}_{\overleftarrow{e_1}} &= \tilde{\mathbf{Q}}^{\top}_{\overrightarrow{e}_1}\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_1}} - \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_1}}\left(\tilde{\mathbf{b}}^{t}_{\overleftarrow{e_1}}\right)^{\top}, \\
\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_1}} &= \tilde{F}^{t}_{\overleftarrow{e_1}}\left(\tilde{\mathbf{Q}}_{\overrightarrow{e}_1}\mathbf{W}_{\overrightarrow{e}_1}, \tilde{\mathbf{X}}^{t}_{\overrightarrow{e_1}}, \tilde{\mathbf{X}}^{t}_{\overleftarrow{e_2}}\right), \\
\tilde{\mathbf{X}}^{t+1}_{\overrightarrow{e_2}} &= \tilde{\mathbf{Q}}_{\overrightarrow{e}_2}\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_2}} - \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_2}}\left(\tilde{\mathbf{b}}^{t}_{\overrightarrow{e_2}}\right)^{\top}, \\
\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_2}} &= \tilde{F}^{t}_{\overrightarrow{e}_2}\left(\tilde{\mathbf{X}}^{t}_{\overrightarrow{e_1}}, \tilde{\mathbf{X}}^{t}_{\overleftarrow{e_2}}\right), \\
\tilde{\mathbf{X}}^{t+1}_{\overleftarrow{e_2}} &= \tilde{\mathbf{Q}}^{\top}_{\overrightarrow{e}}\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_2}} - \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_2}}\left(\tilde{\mathbf{b}}^{t}_{\overleftarrow{e_2}}\right)^{\top}, \\
\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_2}} &= \tilde{F}^{t}_{\overleftarrow{e_2}}\left(\tilde{\mathbf{Q}}_{\overrightarrow{e}_2}\mathbf{W}_{\overrightarrow{e}_2}, \tilde{\mathbf{X}}^{t}_{\overrightarrow{e_2}}, \tilde{\mathbf{X}}^{t}_{\overleftarrow{e_3}}\right)
\end{aligned} \tag{5.57}$$

$$\vdots$$

$$\begin{aligned}
\tilde{\mathbf{X}}^{t+1}_{\overrightarrow{e_L}} &= \tilde{\mathbf{Q}}_{\overrightarrow{e}_L}\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_L}} - \tilde{\mathbf{m}}^{t-1}_{\overleftarrow{e_L}}\left(\tilde{\mathbf{b}}^{t}_{\overrightarrow{e_L}}\right)^{\top}, \\
\tilde{\mathbf{m}}^{t}_{\overrightarrow{e_L}} &= \tilde{F}^{t}_{\overrightarrow{e}_L}\left(\tilde{\mathbf{X}}^{t}_{\overrightarrow{e}_{L-1}}, \tilde{\mathbf{X}}^{t}_{\overleftarrow{e_L}}\right), \\
\tilde{\mathbf{X}}^{t+1}_{\overleftarrow{e_L}} &= \tilde{\mathbf{Q}}^{\top}_{\overrightarrow{e}_L}\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_L}} - \tilde{\mathbf{m}}^{t-1}_{\overrightarrow{e_L}}\left(\tilde{\mathbf{b}}^{t}_{\overleftarrow{e_L}}\right)^{\top}, \\
\tilde{\mathbf{m}}^{t}_{\overleftarrow{e_L}} &= \tilde{F}^{t}_{\overleftarrow{e_L}}\left(\tilde{\mathbf{Q}}_{\overrightarrow{e}_L}\mathbf{W}_{\overrightarrow{e}_L}, \tilde{\mathbf{X}}^{t}_{\overrightarrow{e_L}}\right)
\end{aligned} \tag{5.58}$$

where, for any $\overrightarrow{e} \in \overrightarrow{E}$ and any $t \in \mathbb{N}$ for the right oriented edges

$$\mathbf{b}^{t}_{\overrightarrow{e}_l} = \frac{1}{N}\sum_{i=1}^{n_{l-1}} \frac{\partial \tilde{F}^{t}_{\overrightarrow{e}_l,i}}{\partial \mathbf{X}_{\overleftarrow{e}_l,i}}\left(\left(\mathbf{X}^{t}_{\overrightarrow{e}'_l}\right)_{\overrightarrow{e}'_l:\overrightarrow{e}'_l\to\overrightarrow{e}_l}\right) \qquad \in \mathbb{R}^{q_{\overrightarrow{e}_l}\times q_{\overrightarrow{e}_l}}.$$

and left oriented edges

$$\mathbf{b}^{t}_{\overleftarrow{e}_l} = \frac{1}{N}\sum_{i=1}^{n_l} \frac{\partial \tilde{F}^{t}_{\overleftarrow{e}_l,i}}{\partial \mathbf{X}_{\overrightarrow{e}_l,i}}\left(\tilde{\mathbf{Q}}_{\overrightarrow{e}_l}\mathbf{W}_{\overrightarrow{e}_l}, \left(\mathbf{X}^{t}_{\overleftarrow{e}'_l}\right)_{\overleftarrow{e}'_l:\overleftarrow{e}'_l\to\overleftarrow{e}_l}\right) \qquad \in \mathbb{R}^{q_{\overleftarrow{e}_l}\times q_{\overleftarrow{e}_l}}.$$

Using the separability assumption, we can simplify this expression. To take a concrete example, consider $\tilde{F}^{t}_{\overrightarrow{e}_2}\left(\tilde{\mathbf{X}}^{t}_{\overrightarrow{e_1}}, \tilde{\mathbf{X}}^{t}_{\overleftarrow{e_2}}\right)$. Let's start with the dimensions. Recall

$$\tilde{f}^{t}_{\overrightarrow{e}_2}\left(\tilde{\mathbf{x}}^{t}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}^{t}_{\overleftarrow{e}_2}\right) \in \mathbb{R}^{P_{\overrightarrow{e}_2}q_{\overrightarrow{e}_2}} = \mathbf{V}^{\top}_{\overrightarrow{e}_2} f^{t}_{\overrightarrow{e}_2}\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}^{t}_{\overrightarrow{e_1}}, \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}^{t}_{\overleftarrow{e_2}}\right) \tag{5.59}$$

$$\text{where} \quad \tilde{\mathbf{x}}^{t}_{\overrightarrow{e}_1} \in \mathbb{R}^{D_{\overrightarrow{e}_1}q_{\overrightarrow{e}_1}} = \mathbb{R}^{P_{\overrightarrow{e}_2}q_{\overrightarrow{e}_2}} \text{ and } \tilde{\mathbf{x}}^{t}_{\overleftarrow{e}_2} \in \mathbb{R}^{P_{\overrightarrow{e}_2}q_{\overrightarrow{e}_2}} \tag{5.60}$$

using the separability assumption, we may write

$$\forall\, 1 \leqslant i \leqslant P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2} \tag{5.61}$$

$$\left(f^t_{\overrightarrow{e}_2}\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}^t_{\overrightarrow{e_1}}, \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}^t_{\overleftarrow{e_2}}\right)\right)_i = \sigma^t_{\overrightarrow{e}_2}\left(\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}^t_{\overrightarrow{e_1}}\right)_i, \left(\mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}^t_{\overleftarrow{e_2}}\right)_i\right) \tag{5.62}$$

And

$$\tilde{F}^t_{\overrightarrow{e}_2}\left(\tilde{\mathbf{X}}^t_{\overrightarrow{e}_1}, \tilde{\mathbf{X}}^t_{\overleftarrow{e}_2}\right) \in \mathbb{R}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2} \times q_{\overrightarrow{e}_2}} \tag{5.63}$$

$$\text{where} \quad \tilde{\mathbf{X}}^t_{\overrightarrow{e}_1} \mathbb{R}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2} \times q_{\overrightarrow{e}_2}} \text{ and } \tilde{\mathbf{X}}^t_{\overleftarrow{e}_2} \in \mathbb{R}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2} \times q_{\overrightarrow{e}_2}} \tag{5.64}$$

$$\tilde{F}^t_{\overrightarrow{e}_2}\left(\tilde{\mathbf{X}}^t_{\overrightarrow{e}_1}, \tilde{\mathbf{X}}^t_{\overleftarrow{e}_2}\right) = \begin{bmatrix}\left[(\tilde{f}^t_{\overrightarrow{e}_l})^{(1)}(\tilde{\mathbf{x}}^{t,(1)}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}^{t,(1)}_{\overleftarrow{e}_2}) \dots (\tilde{f}^t_{\overrightarrow{e}_l})^{(q_{\overrightarrow{e}_l})}(\tilde{\mathbf{x}}^{t,(q_{\overrightarrow{e}_l})}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}^{t,(q_{\overrightarrow{e}_l})}_{\overleftarrow{e}_2})\right]_{i=1}^{k_{\overrightarrow{e}_2}} \\ 0_{P_{\overrightarrow{e}_2}(q_{\overrightarrow{e}_2} - k_{\overrightarrow{e}_2}) \times q_{\overrightarrow{e}_2}}\end{bmatrix} \tag{5.65}$$

$$= \left[(\tilde{g}^t_{\overrightarrow{e}_l})^{(1)}(\tilde{\mathbf{x}}^{t,(1)}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}^{t,(1)}_{\overleftarrow{e}_2}) \dots (\tilde{g}^t_{\overrightarrow{e}_l})^{(q_{\overrightarrow{e}_l})}(\tilde{\mathbf{x}}^{t,(q_{\overrightarrow{e}_l})}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}^{t,(q_{\overrightarrow{e}_l})}_{\overleftarrow{e}_2})\right]_{i=1}^{q_{\overrightarrow{e}_2}} \tag{5.66}$$

where each $\tilde{\mathbf{x}}^{t,(i)}_{\overleftarrow{e}_2} \in \mathbb{R}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2}}$. Recall that, for any $1 \leqslant i \leqslant Pk$, $\tilde{F}^t_{\overrightarrow{e}_2, i} : \mathbb{R}^{q_{\overrightarrow{e}_2}} \to \mathbb{R}^{q_{\overrightarrow{e}_2}}$. Then, for any $1 \leqslant k, l \leqslant q_{\overrightarrow{e}_2}$

$$\left(\tilde{\mathbf{b}}^t_{\overrightarrow{e}_2}\right)_{k,l} = \frac{1}{N}\sum_{i=1}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2}} \frac{\partial \tilde{F}^t_{\overrightarrow{e}_2, i, k}}{\partial \mathbf{X}_{\overleftarrow{e}_2, i, l}}\left(\tilde{\mathbf{X}}^t_{\overrightarrow{e}_1}, \tilde{\mathbf{X}}^t_{\overleftarrow{e}_2}\right) \tag{5.67}$$

$$= \frac{1}{N}\sum_{i=1}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2}} \frac{\partial (\tilde{g}^t_{\overrightarrow{e}_2, i})^{(k)}}{\partial \tilde{\mathbf{x}}^{(l)}_{\overleftarrow{e}_2, i}}(\tilde{\mathbf{x}}^{t,(k)}_{\overrightarrow{e}_1}, \tilde{\mathbf{x}}^{t,(k)}_{\overleftarrow{e}_2}) \tag{5.68}$$

$$= \frac{1}{N}\sum_{i=1}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2}} \frac{\partial}{\partial \tilde{\mathbf{x}}^{t,(l)}_{\overleftarrow{e}_2}} \mathbf{V}^\top_{\overrightarrow{e}_2} (g^t_{\overrightarrow{e}_2})^{(k)}\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}^{t,(l)}_{\overrightarrow{e_1}}, \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}^{t,(l)}_{\overleftarrow{e_2}}\right) \tag{5.69}$$

$$= \frac{1}{N}\mathrm{Tr}\left(\mathbf{V}^\top_{\overrightarrow{e}_2} \mathcal{J}_{(g^t_{\overrightarrow{e}_2})^{(k)}}\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}^{t,(l)}_{\overrightarrow{e_1}}, \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}^{t,(l)}_{\overleftarrow{e_2}}\right)\mathbf{V}_{\overrightarrow{e}_2}\right)\delta_{k,l} \tag{5.70}$$

$$= \frac{1}{N}\mathrm{Tr}\left(\mathcal{J}_{(g^t_{\overrightarrow{e}_2})}\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}^t_{\overrightarrow{e_1}}, \mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}^t_{\overleftarrow{e_2}}\right)\right)\delta_{k,l} \tag{5.71}$$

$$= \frac{1}{N}\sum_{i=1}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2}} (\sigma^t)'_{\overrightarrow{e}_2}\left(\left(\mathbf{U}_{\overrightarrow{e}_1}\tilde{\mathbf{x}}^t_{\overrightarrow{e_1}}\right)_i, \left(\mathbf{V}_{\overrightarrow{e}_2}\tilde{\mathbf{x}}^t_{\overleftarrow{e_2}}\right)_i\right)\delta_{k,l} \tag{5.72}$$

where we wrote $\mathcal{J}_{(g^t_{\overrightarrow{e}_2})^{(k)}}$ the $N \times N$ Jacobian matrix of the function $(g^t_{\overrightarrow{e}_2})^{(k)} : \mathbb{R}^N \to \mathbb{R}^N$. Using [37] corollary 2, the Onsager term can be replaced by any estimator based on the asymptotically Gaussian iterates converging, in the high-dimensional limit, to the correct expectation. Using the permutation invariance of the Gaussian distribution, we can therefore replace each element of the matrix the Onsager term with

$$\frac{1}{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2}}\sum_{i=1}^{P_{\overrightarrow{e}_2} q_{\overrightarrow{e}_2}} (\sigma^t)'_{\overrightarrow{e}_2}\left(\left(\tilde{\mathbf{x}}^t_{\overrightarrow{e_1}}\right)_i, \left(\tilde{\mathbf{x}}^t_{\overleftarrow{e_2}}\right)_i\right)\delta_{k,l} \tag{5.73}$$

which amounts to

$$\tilde{\mathbf{b}}^t_{\overrightarrow{e}_2} = b^t_{\overrightarrow{e}_2}\mathbf{I}_{q_{\overrightarrow{e}_2} \times q_{\overrightarrow{e}_2}} \tag{5.74}$$

We therefore obtain an exact reformulation of the initial MLAMP iteration with convolutional matrices in terms of a subset (first line of size $P_{\overrightarrow{e}_l} \times q_{\overrightarrow{e}_l}$ for right oriented edges and $D_{\overrightarrow{e}_l} \times q_{\overrightarrow{e}_l}$ for left-oriented variables) of the variables of a matrix-valued iteration with dense Gaussian matrices verifying the SE equations. Isolating the aforementioned first lines, recalling that the SE equations prescribes i.i.d. lines in the asymptotically Gaussian fields, we recover that, for any $1 \leqslant l \leqslant L$, the variable $\mathbf{x}_{\overrightarrow{e}_l} \in \mathbb{R}^{P_{\overrightarrow{e}_l} q_{\overrightarrow{e}_l}}$ is composed of $q_{\overrightarrow{e}_l}$ copies of block of size $P_{\overrightarrow{e}_l}$ with i.i.d. Gaussian elements distributed according to the SE equations (5.1.2). The distribution of the variables associated to left-oriented edges is obtained similarly. Note that, from a finite size point of view, the effect of $D_{\overrightarrow{e}_l}, P_{\overrightarrow{e}_l}$ is different from that of $q_{\overrightarrow{e}_l}$: the former results in subGaussian concentration i.e. exponential in the dimension, while the latter only represents copies (and not i.i.d. samples), and thus only has an averaging effect. This is observed in simulations. □

5.1.3 Bayes-optimal MLAMP with random convolutional matrices

In this section, we specialize the equations obtained in the previous section to the Bayes-optimal MLAMP iteration of the main body of the paper. Several functions are reminded for convenience. Consider the MLAMP iteration outlined in section 4.1.2. The scalar updates described in Eq.(4.4) can be rewritten as vector-valued updates as follows, for any $t \in \mathbb{N}$, and any $0 \leqslant l \leqslant L$:

$$\boldsymbol{\omega}^{(l)}(t) = \mathbf{W}^{(l)} \hat{\mathbf{h}}^{(l)}(t) - V^{(l)}(t) \mathbf{g}^{(l)}(t-1) \tag{5.75}$$

$$\mathbf{B}^{(l)}(t) = \left(\mathbf{W}^{(l)}\right)^{\top} \mathbf{g}^{(l)}(t) - \hat{V}^{(l)}(t) \hat{\mathbf{h}}(t). \tag{5.76}$$

To define the update functions and terms $V^{(l)}, \hat{V}^{(l)}$, the following partition functions were introduced.

- for $l = 1$

$$\mathcal{Z}^{(1)}\left(y, V^{(1)}, \omega^{(1)}\right) = \frac{1}{\sqrt{2\pi V^{(1)}}} \int dz P_{out}^{(1)}(y|z) e^{-\frac{(z-\omega^{(1)})^2}{2V^{(1)}}} \tag{5.77}$$

- for any $2 \leqslant l \leqslant L-1$:

$$\mathcal{Z}^{(l)}\left(A^{(l-1)}, B^{(l-1)}, V^{(l)}, \omega^{(l)}\right) = \frac{1}{\sqrt{2\pi V^{(l)}}} \int dh dz P_{out}^{(l)}(h|z) e^{-\frac{1}{2}A^{(l-1)}h^2 + B^{(l-1)}h} e^{-\frac{(z-\omega^{(l)})^2}{2V^{(l)}}} \tag{5.78}$$

- for $l = L$

$$\mathcal{Z}^{(L)}(A^{(L)}, B^{(L)}) = \int dh P_X(h) e^{-\frac{1}{2}A^{(L)}h^2 + B^{(L)}h} \tag{5.79}$$

We then define the layer-dependent, time-dependent, scalar update functions $f^{(l),t}, \tilde{f}^{(l),t}$

$$\forall\, (B, \omega) \in \mathbb{R}^2$$

$$f^{(1),t}(\omega) = \partial_\omega \log \mathcal{Z}^{(1)}\left(y, V^{(1)}(t), \omega\right) \tag{5.80}$$

$$f^{(l),t}(B, \omega) = \partial_\omega \log \mathcal{Z}^{(l)}\left(A^{(l-1)}(t), B, V^{(l)}(t), \omega\right) \quad 2 \leqslant l \leqslant L \tag{5.81}$$

$$\tilde{f}^{(l),t}(B, \omega) = \partial_B \log \mathcal{Z}^{(l+1)}\left(A^{(l)}(t-1), B, V^{(l+1)}(t-1), \omega\right) \quad 1 \leqslant l \leqslant L-1 \tag{5.82}$$

$$\tilde{f}^{(L,t)}(B) = \partial_B \log \mathcal{Z}^{(L+1)}\left(A^{(L)}(t-1), B\right), \tag{5.83}$$

and their corresponding separable, vector valued counterparts $\mathbf{f}^{(l)}, \tilde{\mathbf{f}}^{(l)}$, which leads to the following iteration

$$\boldsymbol{\omega}^{(l)}(t) = \mathbf{W}^{(l)}\tilde{\mathbf{f}}^{(l),t}(\mathbf{B}^{(l),t-1}, \boldsymbol{\omega}^{(l+1),t-1}) - V^{(l)}(t)\mathbf{f}^{(l),t-1}(\mathbf{B}^{(l-1),t-1}, \boldsymbol{\omega}^{(l),t-1}) \tag{5.84}$$

$$\mathbf{B}^{(l)}(t) = \left(\mathbf{W}^{(l)}\right)^{\top}\mathbf{f}^{(l),t}(\mathbf{B}^{(l-1),t}, \boldsymbol{\omega}^{(l),t}) - \hat{V}^{(l)}(t)\tilde{\mathbf{f}}^{(l),t}(\mathbf{B}^{(l),t-1}, \boldsymbol{\omega}^{(l+1),t-1}), \tag{5.85}$$

where the Onsager terms $V^{(l),t}$ and $\hat{V}^{(l),t}$ reduce to, using the separability of the update functions,

$$V^{(l),t} = \frac{1}{n_l}\sum_{i=1}^{n_{l-1}} \partial_B \tilde{f}^{(l),t}(B_i^{(l),t-1}, \omega_i^{(l+1),t-1}) \tag{5.86}$$

$$\hat{V}^{(l),t} = \frac{1}{n_l}\sum_{j=1}^{n_l} \partial_\omega f^{(l),t}(B_j^{(l-1),t}, \omega_j^{(l),t}) = -A^{(l),t} \tag{5.87}$$

We now show that the update functions defined above are Lipschitz continuous and increasing, thus ensuring that the integrals are well defined through positivity of the parameters $V, \hat{V}$.

Lemma 24. *For any $1 \leqslant l \leqslant L$, and any $t \in \mathbb{N}$, the functions $f^{(l),t}, \tilde{f}^{(l),t}$ are Lipschitz continuous in B, ω. Furthermore, the functions $f^{(l),t}, \tilde{f}^{(l),t}$ are respectively decreasing in ω and increasing in B. As a consequence, the variance terms $A^{(l),t}$ and $V^{(l),t}$ are strictly positive.*

Proof. Recall the partition function, omitting the layer index since all regularity assumptions are the same for all layers and time indices,

$$\mathcal{Z}(A, B, V, \omega) := \frac{1}{\sqrt{2\pi V}}\int P(h \mid z)\exp\left(Bh - \frac{1}{2}Ah^2 - \frac{(z-\omega)^2}{2V}\right) dh\, dz \tag{5.88}$$

recalling $p(h|z) = \int p(\xi)\delta(h - f_\xi(z))d\xi$, integrating in h yields

$$\mathcal{Z}(A, B, V, \omega) := \frac{1}{\sqrt{2\pi V}}\int P(\xi)\exp\left(Bf_\xi(z) - \frac{1}{2}Af_\xi(z)^2 - \frac{(z-\omega)^2}{2V}\right) d\xi\, dz \tag{5.89}$$

Starting with $\tilde{f}$, we can straightforwardly verify the conditions to apply the dominated convergence theorem and differentiate under the integral to obtain

$$\begin{aligned}
&\partial_B \tilde{f}(B, \omega) = \partial_B^2 \log\left(\mathcal{Z}(A, B, V, \omega)\right) \\
&= \frac{1}{(\sqrt{2\pi V}\mathcal{Z}(A, B, V, \omega))^2}\Bigg(\int P(\xi) f_\xi^2(z)\exp\left(Bf_\xi(z) - \frac{1}{2}Af_\xi(z)^2 - \frac{(z-\omega)^2}{2V}\right) d\xi\, dz \times \\
&\int P(\xi)\exp\left(Bf_\xi(z) - \frac{1}{2}Af_\xi(z)^2 - \frac{(z-\omega)^2}{2V}\right) d\xi\, dz - \\
&\qquad\qquad \left(\int P(\xi) f_\xi(z)\exp\left(Bf_\xi(z) - \frac{1}{2}Af_\xi(z)^2 - \frac{(z-\omega)^2}{2V}\right) d\xi\, dz\right)^2\Bigg) \geqslant 0
\end{aligned} \tag{5.90}$$

where the positivity comes from the Cauchy-Schwarz inequality and positivity of the term $P(\xi)\exp\left(Bf_\xi(z) - \frac{1}{2}Af_\xi(z)^2 - \frac{(z-\omega)^2}{2V}\right)$. Turning to f, we complete the square in the variable h to obtain

$$\mathcal{Z}(A, B, V, \omega) := \frac{\exp\left(\frac{B^2}{2A}\right)}{\sqrt{2\pi V}}\int P(\xi)\exp\left(-\frac{A}{2}\left(f_\xi(z) - \frac{B}{A}\right)^2\right)\exp\left(-\frac{(z-\omega)^2}{2V}\right) d\xi\, dz \tag{5.91}$$

and differentiating under the integral yields

$$f(B,\omega) = \partial_\omega \log\left(\mathcal{Z}(A,B,V,\omega)\right) \tag{5.92}$$

$$= \frac{1}{V}\left(\frac{\int P(\xi) z \exp\left(-\frac{A}{2}\left(f_\xi(z)-\frac{B}{A}\right)^2\right)\exp\left(-\frac{(z-\omega)^2}{2V}\right)\,d\xi\,dz}{\left(\int P(\xi)\exp\left(-\frac{A}{2}\left(f_\xi(z)-\frac{B}{A}\right)^2\right)\exp\left(-\frac{(z-\omega)^2}{2V}\right)\,d\xi\,dz\right)} - \omega\right) \tag{5.93}$$

where the term $\frac{\int P(\xi) z \exp\left(-\frac{A}{2}\left(f_\xi(z)-\frac{B}{A}\right)^2\right)\exp\left(-\frac{(z-\omega)^2}{2V}\right)\,d\xi\,dz}{\left(\int P(\xi)\exp\left(-\frac{A}{2}\left(f_\xi(z)-\frac{B}{A}\right)^2\right)\exp\left(-\frac{(z-\omega)^2}{2V}\right)\,d\xi\,dz\right)}$ is the conditional mean of the distribution with density $\frac{\int P(\xi)\exp\left(-\frac{A}{2}\left(f_\xi(z)-\frac{B}{A}\right)^2\right)\exp\left(-\frac{(z-\omega)^2}{2V}\right)d\xi}{\left(\int P(\xi)\exp\left(-\frac{A}{2}\left(f_\xi(z)-\frac{B}{A}\right)^2\right)\exp\left(-\frac{(z-\omega)^2}{2V}\right)\,d\xi\,dz\right)}$. The Lipschitz property is straightforward to verify using the polynomial bound assumption on the activation functions and the inverse exponential factors. □

In the Bayes-optimal MLAMP, see [188], the planted vectors $\mathbf{w}_{\overrightarrow{e}_l}$ are chosen as independently distributed as the asymptotic SE representation of the output of the previous layer, and are therefore Lipschitz transforms of subGaussian random variables, and thus are also subgaussian. Using the permuation invariance of the Gaussian distribution, the quantities $\mathbf{z}_{\overrightarrow{e}_l} = \hat{\mathbf{A}}_{\overrightarrow{e}_l}$ remain Gaussian. We can therefore apply the result of Lemma 23 to this iteration and obtain that iterates of Eq.(4.4) verify the SE equations from Lemma 23 with the corresponding update functions. Furthermore, in the Bayes optimal case, the Nishimori conditions, see e.g. [154], allow to only keep the parameters $\nu_{\overrightarrow{e}_l}, \hat{\nu}_{\overleftarrow{e}_l}$ to describe the distribution of of the iterates, recovering the equations of Theorem 7. Finally, the rescaling of the variances to go from the factors δ_l to the β_l of the main can be done by rescaling each non-linearity $f^t_{\overrightarrow{e}_l}$ by $\sqrt{N/n_{l-1}}$ (and similary for the $f^t_{\overleftarrow{e}_l}$ with $\sqrt{N/n_l}$) as done in [135, 37].

Chapter 6

Asymptotics of stochastic gradient descent

The results presented in this chapter are unpublished and part of a work currently in preparation.

We prove closed-form equations for the exact high-dimensional asymptotics of a family of first order gradient-based methods, learning an estimator (e.g. M-estimator, shallow neural network, ...) from observations on Gaussian data with empirical risk minimization. This includes widely used algorithms such as stochastic gradient descent (SGD) or Nesterov acceleration. We show that the obtained equations match those resulting from the discretization of dynamical mean-field theory (DMFT) equations from statistical physics when applied to gradient flow. Our proof method has the benefit of being quite streamlined, notably with respect to previous literature which often involves a rather high level of technicality. Notably, we give an explicit description of how memory kernels build up in the effective dynamics, and include non-separable update functions, allowing datasets with non-identity covariance matrices. Finally, we provide numerical implementations of the equations for SGD with varying batch-sizes and learning rates.

6.1 Introduction

Stochastic gradient descent methods are one of the cornerstones of optimization and thus, modern machine-learning. Notably, stochastic gradient descent and its variants have become the method of choice for the optimization of large deep learning architectures, see e.g. [157, 144, 249]. Gradient based dynamics are, however, not restricted to the field of machine learning and computational mathematics, as they are also at the center of out-of-equilibrium statistical mechanics through the notion of Langevin dynamics, see e.g. [196]. Obtaining an exact understanding of these procedures has been a long-standing problem, notably for spin glasses where a significant set of results has been obtained, first using heuristic, theoretical physics [268, 269, 67, 68] methods and then rigorous probability theory [11, 35, 56, 168]. In theoretical physics, the effective dynamics describing the high-dimensional behavior of gradient flow is called dynamical mean-field theory (DMFT), in reference to the reduction of a system of strongly correlated degrees of freedom to low-dimensional order parameters whose evolution can be tracked analytically by a set of self-consistent equations. In the continuous time limit, those equations take the form of a stochastic integro-differential system involving memory kernels and additive Gaussian processes, whose parameters are all related to the form of the gradient, temperature (of the thermal noise), or other characteristics of the original

system. In recent years, DMFT equations have been used by physicists to study a wide variety of high-dimensional disordered dynamical systems (see, e.g., [184, 276, 186, 248]), including constraint satisfaction and learning problems [5, 198, 200, 187, 260, 199].

While the recent work of [56] provides game-changing progress into the rigorous establishment of the DMFT, it does not account for stochasticity of the gradient descent algorithms and their proof is limited to the data matrix to be random, with i.i.d. centered subgaussian entries. In the present work we remove these two limitations and establish the DMFT equations for a broad class of stochastic algorithms (including SGD, various momentum methods or Langevin algorithms), and for a broader class of data (including Gaussian with a rather generic covariance).

Theoretical physics works on DMFT aim to describe the continuous time dynamics, because the physical dynamics simply is continuous. When gradient based methods are used as algorithms they are always run in discrete time and thus for algorithmic purposes analysis of the discrete dynamics is of larger interest. In previous theoretical physics works the DMFT is always presented for the continuous (flow) limit of the dynamics. In this paper we prove that the discrete DMFT equations provide exact asymptotic analysis for the discrete gradient descent methods as well. This has been noticed empirically in [198]. While a larger part of [56] is devoted to proving the continuous-time equations, they also establish the discrete time DMFT. In the present paper we will only consider the discrete version because (a) our main motivation is analysis of actual algorithms, (b) the exactness of the discrete DMFT is not discussed in the literature and we thus want to rectify that.

Our proof of dynamical mean-field theory equations applies to a wide range of supervised learning problems, where an estimator is learned using stochastic gradient descent on a cost function defined by empirical risk minimization. In this regard, consider the following optimization problem

$$\hat{\mathbf{w}} \in \inf_{\mathbf{w}\in\mathbb{R}^{d\times q}} \mathcal{L}(\mathbf{X}\mathbf{w}, \mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{6.1}$$

$$\text{where } \mathbf{y} = \Phi_0\left(\mathbf{X}\mathbf{w}^*\right), \tag{6.2}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the design matrix, the observed labels $\mathbf{y} \in \mathbb{R}^n$ are generated according to a ground truth parametrized by a continuous, separable function $\Phi_0 : \mathbb{R}^{n\times q} \to \mathbb{R}^n$ and ground-truth vector $\mathbf{w}^* \in \mathbb{R}^{d\times q}$, and the loss and regularization $\mathcal{L}, \mathbf{F}$ are differentiable functions. The number of samples n and dimension of the inputs d will be taken to infinity (the high-dimensional limit), while the number of weight vectors q will remain finite. We will consider a generic family of discrete-time dynamics in Theorem 9, which includes stochastic gradient descent methods widely used in practice: a candidate $\hat{\mathbf{w}}$ is estimated using gradient descent by producing the following sequence of iterates

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \gamma^t \left(\mathbf{X}^\top \nabla \mathcal{L}^t(\mathbf{X}\mathbf{w}^t, \mathbf{y}) + \nabla \mathbf{F}(\mathbf{w}^t)\right) \tag{6.3}$$

where γ^t is the scalar learning rate, and the time-dependent gradient represents potential modifications of the gradient descent, for instance mini-batch sampling with batch-size being a finite fraction of d in the high-dimensional limit.

Our main result is an asymptotically (i.e. in the high-dimensional limit) exact characterization of the distribution of the iterates $\mathbf{w}^t$ and preactivations $\mathbf{X}\mathbf{w}^t$ at each time step, in the weak sense. In particular, our results encompass the following special cases:

1. an exact asymptotic characterization of discrete-time (multi-pass) stochastic gradient descent with mini-batch sizes proportional to the data dimension;

2. a data matrix $\mathbf{X}$ with any positive definite covariance $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ with bounded spectral norm;
3. a finite number q of learners;
4. time dependent update functions which may include stochastic effects such as mini-batch sampling, learning rate schedules and thermal noise (i.e., *Langevin equation*), and any differentiable regularization;
5. momentum methods such as Polyak's heavy ball and Nesterov accelerated gradient.

6.2 Related works

Rigorous proofs of dynamical mean-field theory equations first appeared in the context of spin glasses in the works [11, 35], who applied large deviation theory to the paths generated by the Langevin dynamics corresponding to the Hamiltonians of the Sherrington-Kirkpatrick and spherical p-spin models.

More recently, [56] proposed a different proof for the DMFT of the high-dimensional asymptotics of first order flows for the empirical risk minimization problem (6.2). This new approach was based on an approximate message passing (AMP) iteration with memory, building upon an implicit mapping between the AMP iterates and the discretized gradient flow, and using the high-dimensional concentration properties of AMP iterations, the state evolution (SE) equations. Our proof instead is based on iterative Gaussian conditioning, and as a consequence is simpler and more direct. Iterative Gaussian conditioning is a technique introduced in the study of SE equations for AMP iterations [28, 135, 42, 37, 110]. In AMP iterations, the so-called Onsager correction applied at each time step drastically simplifies the high-dimensional effective dynamics, leading to a Markovian Gaussian process. Since gradient descent has no Onsager correction, one key aspect of the proof is to show how the dynamics may be decomposed and reformulated into asymptotically tractable memory terms and additive Gaussian processes. As a result, our proof is completely explicit and we provide intuition on how the different terms appear in subsections 6.4.1 before moving to the general case in Appendix 6.6.

Our proof technique based on the iterative conditioning has important benefits as it becomes straightforward to account for additional stochastic effects that are independent on the design matrix, notably mini-batch sampling or thermal noise, as well as potential momentum terms. Additionally, we allow non-separable, time-dependent update functions, which enables to handle design matrices with arbitrary well-conditioned covariance and bounded spectral norm. We do not study the continuous time limit, provided in [56] for gradient flow on separable cost functions. Notably, they prove the existence and uniqueness of the solution to the stochastic integro-differential system describing the high-dimensional gradient flow dynamics under suitable conditions. They also benefit from the universality results for AMP iterations, [27, 62], allowing design matrices with independent sub-Gaussian entries and identity covariance.

Finally, it is interesting to note that, although methods from theoretical physics are often not rigorous, a direct parallel can be drawn between our proof and derivation of the dynamical cavity method as formulated in [172], [196] and references therein for earlier appearances. Indeed, the dynamical cavity method relies on a orthogonal decomposition of the samples and iterates along a chosen direction, resulting in approximately independent Gaussian terms with different scalings. As a low dimensional projection, the term aligned with the chosen direction is of finite order, while the orthogonal component contains a number of directions proportional to the dimension and thus

remains of extensive order. A Taylor expansion then allows to simplify the dynamics and obtain the DMFT equations with some algebra. In the present rigorous proof, we also perform orthogonal decompositions, but in the direction of previous iterates. For a finite number of iterations and width q of the iterates, the component resulting from this projection is also of low-order, while the orthogonal component remains extensive. The proof, done by induction, then boils down to a precise control of the correlations of the different terms and concentration of various inner products appearing due to the projections using the induction hypothesis.

6.3 Main result

Our main result characterizes the high-dimensional dynamics of a family of iterations that includes gradient descent iteration Eq. (6.3), and takes the generic form

$$\mathbf{v}^{t+1} = \mathbf{h}^t\left(\left\{\mathbf{v}^k\right\}_{k=0}^t\right) + \mathbf{X}^\top \mathbf{g}^t(\mathbf{r}^t) \tag{6.4}$$

$$\mathbf{r}^t = \mathbf{X}\sum_{k=0}^t \mathbf{v}^k \tag{6.5}$$

The update functions $\mathbf{g}^t, \mathbf{h}^t$ will belong to the regularity class of pseudo-Lipschitz functions, which will also be used to characterize the (weak) convergence of random matrices (of finite width) in the rest of the paper. This family of functions is commonly used in the AMP literature, see e.g. [37], and is reminded in Appendix 6.5. Note that, when considering a planted model as in Eq. (6.2) and the corresponding gradient based dynamics will involve a sequence of functions $\mathbf{g}^t$ implicitly depending on the data matrix $\mathbf{X}$ through the observed labels $\mathbf{y}$. Following [56], this additional dependence can be dealt with by considering an augmented variable $[\mathbf{w}|\mathbf{w}_*]$ and a corresponding update function involving the gradient step on $\mathbf{w}_0$, which is made possible by the validity of the result for matrix-valued variables of finite width. It can also be dealt with using an orthogonal decomposition in the direction of $\mathbf{w}_*$, see e.g. [110], however we will use the former formulation to avoid redundant derivations.

6.3.1 Examples of algorithms belonging to the considered family

Stochastic gradient-descent Consider the following stochastic gradient-descent dynamics with constant step-size γ

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \gamma\left(\frac{1}{b}\mathbf{X}^\top \mathbf{s}^t \odot \nabla\mathcal{L}(\mathbf{X}\mathbf{w}^t) + \nabla\mathbf{F}(\mathbf{w}^t)\right). \tag{6.6}$$

where $\mathbf{s}^t \in \mathbb{R}^n$ is a random vector with i.i.d. elements sampled at each time step according to a Bernoulli distribution with parameter b, and $\odot$ is the Hadamard product. Now define the increment variable $\mathbf{v}^t = \mathbf{w}^t - \mathbf{w}^{t-1}$ such that, for any $t \in \mathbb{N}$, $\mathbf{w}^t = \sum_{k=0}^t \mathbf{v}^t$ with the convention $\mathbf{v}^{-1} = 0$; the preactivation term $\mathbf{r}^t = \mathbf{X}\mathbf{w}^t \in \mathbb{R}^{n\times q}$, such that the stochastic gradient-descent iteration may be rewritten

$$\mathbf{v}^{t+1} = -\gamma\nabla\mathbf{F}\left(\sum_{k=0}^t \mathbf{v}^t\right) - \gamma\mathbf{X}^\top\mathbf{s}^t \odot \nabla\mathcal{L}(\mathbf{r}^t) \tag{6.7}$$

$$\mathbf{r}^t = \mathbf{X}\sum_{k=0}^t \mathbf{v}^t \tag{6.8}$$

which fits the form of Eq. (6.4-6.5) by choosing $\mathbf{g}^t(\mathbf{r}^t) = -\gamma \mathbf{s}^t \odot \nabla \mathcal{L}(\mathbf{r}^t)$, $\mathbf{h}^t(\mathbf{w}^t) = -\gamma \nabla \mathbf{F}(\mathbf{w}^t)$. Notice that our characterization requires that the size of the training mini batch is a finite fraction of the full dataset.

Langevin algorithm The discretized Langevin algorithm amounts to adding independent Gaussian noise to the gradient descent, leading to the following iteration

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \gamma \left(\mathbf{X}^\top \mathbf{s}^t \odot \nabla \mathcal{L}(\mathbf{X}\mathbf{w}^t) + \nabla \mathbf{F}(\mathbf{w}^t) \right) + \gamma \sqrt{T} \mathbf{z}^t \tag{6.9}$$

where $\mathbf{z}^t \in \mathbb{R}^d$ has i.i.d. standard normal elements and is independent from all other problem parameters and $\mathbf{z}^{t'}$ for all $t' \neq t$. It is then straightforward to redefine the function $\mathbf{h}^t(\mathbf{w}^t) = -\gamma \nabla \mathbf{F}(\mathbf{w}^t) + \sqrt{T}\mathbf{z}^t$, which will simply lead to an additive noise with variance T at each time step in the Gaussian process u^t of the field ν^{t+1} in Corollary 2. This modification is also observed when discretizing the DMFT equations obtained from theoretical physics methods [198].

Polyak momentum Polyak momentum [236] (or heavy-ball method) reads

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \gamma \left(\mathbf{X}^\top \nabla \mathcal{L}(\mathbf{X}\mathbf{w}^t) + \nabla \mathbf{F}(\mathbf{w}^t) \right) + \beta \left(\mathbf{w}^t - \mathbf{w}^{t-1} \right) \tag{6.10}$$

with gradient step size α and momentum parameter β. Using the same intermediate variables as those introduced for the reformulation of the stochastic gradient-descent iteration Eq.(6.6) into dynamics of the form of Eq. (6.4-6.5), we obtain

$$\mathbf{v}^{t+1} = -\gamma \nabla \mathbf{F}(\sum_{k=0}^{t} \mathbf{v}^t) - \gamma \mathbf{X}^\top \nabla \mathcal{L}(\mathbf{r}^t) + \beta \mathbf{v}^t \tag{6.11}$$

$$\mathbf{r}^t = \mathbf{X} \sum_{k=0}^{t} \mathbf{v}^t \tag{6.12}$$

which fits the form of Eq. (6.4-6.5) by choosing $\mathbf{g}^t(\mathbf{r}^t) = -\gamma \nabla \mathcal{L}(\mathbf{r}^t)$, and $\mathbf{h}^t(\left\{\mathbf{v}^k\right\}_{k=0}^t) = -\gamma \nabla \mathbf{F}(\sum_{k=0}^t \mathbf{v}^k) + \beta \mathbf{v}^t$.

Nesterov accelerated gradient Nesterov accelerated gradient [213] is defined as an iteration of three sequences parametrized by stepsizes $\tau^t, \gamma^t, \nu^t, \alpha^t$ and initialized with $\mathbf{w}^0, \mathbf{z}^0$, taking the form

$$\mathbf{y}^t = \mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t) \tag{6.13}$$

$$\mathbf{w}^{t+1} = \mathbf{y}^t - \gamma^t \left(\mathbf{X}^\top \nabla \mathcal{L}(\mathbf{X}\mathbf{y}^t) + \nabla \mathbf{F}(\mathbf{y}^t) \right) \tag{6.14}$$

$$\mathbf{z}^{t+1} = \mathbf{z}^t + \mu^t \left(\mathbf{y}^t - \mathbf{z}^t \right) - \alpha^t \left(\mathbf{X}^\top \nabla \mathcal{L}(\mathbf{X}\mathbf{y}^t) + \nabla \mathbf{F}(\mathbf{y}^t) \right) \tag{6.15}$$

Defining the variables $\mathbf{u}^{t+1} = \mathbf{w}^{t+1} - \mathbf{w}^t \in \mathbb{R}^d, \tilde{\mathbf{u}}^{t+1} = \mathbf{z}^{t+1} - \mathbf{z}^t \in \mathbb{R}^d, \mathbf{v}^t = [\mathbf{u}^t | \tilde{\mathbf{u}}^t] \in \mathbb{R}^{d \times 2}, \mathbf{x}^t = [\mathbf{w}^t | \mathbf{z}^t] = \sum_{k=0}^t \mathbf{v}^k \in \mathbb{R}^{d \times 2}$, $\mathbf{r}^t = \mathbf{X} \sum_{k=0}^t \mathbf{v}^k$, we may fit these equations to the form of Eq. (6.4-6.5)

by defining

$$\mathbf{h}^t : \mathbb{R}^{d\times 2(t+1)} \to \mathbb{R}^{d\times 2} \tag{6.16}$$

$$\left\{\mathbf{v}^k\right\}_{k=0}^{t} \to \left[\sum_{k=0}^{t} \mathbf{v}^k \begin{bmatrix} -\tau^t \\ \tau^t \end{bmatrix} | \sum_{k=0}^{t} \mathbf{v}^k \begin{bmatrix} \mu^t(1-\tau^t) \\ \mu^t(\tau^t - 1) \end{bmatrix}\right] \tag{6.17}$$

$$+ \left[-\gamma^t \nabla F\left(\sum_{k=0}^{t} \mathbf{v}^k \begin{bmatrix} 1-\tau^t \\ \tau^t \end{bmatrix}\right) | - \alpha^t \nabla F\left(\sum_{k=0}^{t} \mathbf{v}^k \begin{bmatrix} 1-\tau^t \\ \tau^t \end{bmatrix}\right)\right] \tag{6.18}$$

$$\mathbf{g}^t : \mathbb{R}^{n\times 2} \to \mathbb{R}^{n\times 2} \tag{6.19}$$

$$\mathbf{r}^t \to \left[-\gamma^t \nabla \mathcal{L}\left(\mathbf{r}^t \begin{bmatrix} 1-\tau^t \\ \tau^t \end{bmatrix}\right) | - \alpha^t \nabla \mathcal{L}\left(\mathbf{r}^t \begin{bmatrix} 1-\tau^t \\ \tau^t \end{bmatrix}\right)\right] \tag{6.20}$$

The details of this mapping are given in Appendix 6.7.

6.3.2 Statement of the main theorem

We now state the required assumptions for our main result to hold.

Assumptions

(A1) the dimensions of the problem n, d go to infinity with finite ratio $n/d = \alpha$;

(A2) the matrix $\mathbf{X}$ has i.i.d. $\mathcal{N}(0, \frac{1}{d})$ elements;

(A3) for any $t \in \mathbb{N}$, the functions $\mathbf{g}^t : \mathbb{R}^{n\times q} \to \mathbb{R}^{n\times q}, \mathbf{h}^t : \mathbb{R}^{d\times q} \to \mathbb{R}^{d\times q}$ are pseudo-Lipschitz continuous of order k (in their arguments), and may involve random effects (accounted for by random variables, not considered as arguments) independent of the matrix $\mathbf{X}$, initialization $\mathbf{w}^0$ and ground truth $\mathbf{w}^*$. If these functions contain said additional random effects, the pseudo-Lipschitz property is assumed to be verified with high probability as the dimensions go to infinity;

(A4) the columns of the initalization $\mathbf{w}^0$ and planted model $\mathbf{w}^*$ are drawn from distributions in $\mathbb{R}^d$ verifying dimension-free log-Sobolev inequalities and are independent of other random parameters of the dynamics;

(A5) for any time $t \in \mathbb{N}$, for any arguments verifying a dimension-free log-Sobolev inequality, the inner products of the expectations of the functions $\mathbf{g}^s, \mathbf{g}^t$ and $\mathbf{h}^s, \mathbf{h}^t$, for any $t \in \mathbb{N}$, for any $0 \leqslant s \leqslant t$, converge with high probability to finite constants.

The last condition is a short reformulation of the stability conditions (A5-A7) of [37, 110]. The log-Sobolev assumption may be replaced with slower decaying distributions (e.g. subGaussian) if more regular, for instance separable and/or Lipschitz, are used. We keep the log-Sobolev assumption for simplicity and clarity of presentation in the non-separable case. Our main result is presented in the following theorem:

Theorem 9. *(High-dimensional dynamics of gradient-based methods) Consider the following discrete time stochastic process*

$$\nu^{t+1} = \theta^t \Gamma^t + \mathbf{h}^t \left(\left\{ \nu^k \right\}_{k=0}^t \right) + \sum_{k=0}^{t-1} \theta^k R_g(t,k) + \mathbf{u}^t \in \mathbb{R}^{d \times q} \tag{6.21}$$

$$\theta^t = \sum_{k=0}^{t} \nu^k \in \mathbb{R}^{d \times q} \tag{6.22}$$

$$\eta^t = \sum_{k=0}^{t-1} \mathbf{g}^k(\eta^k) R_\theta(t,k) + \omega^t \in \mathbb{R}^{n \times q} \tag{6.23}$$

$$R_\theta(t,s) = \lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{d} \mathbb{E} \left[\frac{\partial \theta_i^t}{\partial u_i^s} \right] \in \mathbb{R}^{q \times q} \tag{6.24}$$

$$R_g(t,s) = \lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{n} \mathbb{E} \left[\frac{\partial g_i^t}{\partial \omega_i^s}(\eta^t) \right] \in \mathbb{R}^{q \times q} \tag{6.25}$$

$$\Gamma^t = \lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{n} \mathbb{E} \left[\frac{\mathrm{d} g_i^t}{\mathrm{d} \eta_i^t}(\eta^t) \right] \in \mathbb{R}^{q \times q} \tag{6.26}$$

$$C_\theta(t,s) = \lim_{d \to \infty} \frac{1}{d} \mathbb{E} \left[\left(\theta^t \right)^\top \theta^s \right] \in \mathbb{R}^{q \times q} \tag{6.27}$$

$$C_g(t,s) = \lim_{d \to \infty} \frac{1}{d} \mathbb{E} \left[\mathbf{g}^s(\eta^s)^\top \mathbf{g}^t(\eta^t) \right] \in \mathbb{R}^{q \times q} \tag{6.28}$$

initialized with $\nu^0 = \mathbf{v}^0$, where $\mathbf{u}^t, \omega^t$ have i.i.d. lines in $\mathbb{R}^q$ which are Gaussian processes with covariances $C_g^{s,t}, C_\theta^{s,t}$. Consider the iteration Eq. (6.4-6.5). Then, under assumptions (A1)-(A5), for any $t \in \mathbb{N}$, and any pseudo-Lipschitz functions $\Psi : \mathbb{R}^{d \times q(t+1)} \to \mathbb{R}$ and $\Phi : \mathbb{R}^{n \times qt} \to \mathbb{R}$:

$$\begin{aligned} &\Psi(\mathbf{w}^0, ..., \mathbf{w}^t) \xrightarrow[n,d \to \infty]{\text{w.h.p.}} \mathbb{E} \left[\Psi(\theta^0, ..., \theta^t) \right] ; \text{ and} \\ &\Phi(\mathbf{r}^0, ..., \mathbf{r}^{t-1}) \xrightarrow[n,d \to \infty]{\text{w.h.p.}} \mathbb{E} \left[\Phi(\eta^0, ..., \eta^{t-1}) \right] . \end{aligned} \tag{6.29}$$

The following corollary gives the high-dimensional dynamics for the SGD iteration described at Eq.(6.6) with separable functions. Assume that the loss function $\mathcal{L}$ and regularization $\mathbf{F}$ are separable with the respective component-wise scalar functions l, f, and that $\mathcal{L}$ is twice differentiable. The non-linearities $\mathbf{g}^t, \mathbf{h}^t$ are then also separable with component-wise functions $g^t(r^t) = -\gamma s^t l'(r^t)$ and $h^t(w^t) = -\gamma f'(w^t)$. Since the variables $\boldsymbol{\nu}^t, \boldsymbol{\eta}^t$, respectively in $\mathbb{R}^{d \times q}$ and $\mathbb{R}^{n \times q}$.

Corollary 2. *Consider the SGD iteration of Eq.(6.6) and assume that the loss function $\mathcal{L}$ is twice*

differentiable. Consider the following discrete-time stochastic process

$$\nu^{t+1} = \Gamma^t \theta^t - \gamma f'(\theta^t) + \sum_{k=0}^{t-1} R_g(t,k)\theta^k + u^t \in \mathbb{R}^q \tag{6.30}$$

$$\theta^t = \sum_{k=0}^{t} \nu^t \in \mathbb{R}^q \tag{6.31}$$

$$\eta^t = -\gamma \sum_{k=0}^{t-1} R_\theta(t,k) s^k l'(\eta^k) + \omega^t \in \mathbb{R}^q \tag{6.32}$$

$$R_\theta(t,s) = \mathbb{E}\left[\frac{\partial \theta^t}{\partial u^s}\right] \in \mathbb{R}^{q\times q} \tag{6.33}$$

$$R_g(t,s) = -\alpha\gamma \mathbb{E}\left[s^t \frac{\partial l'}{\partial \omega^s}(\eta^t)\right] \in \mathbb{R}^{q\times q} \tag{6.34}$$

$$\Gamma^t = -\alpha\gamma \mathbb{E}\left[s^t l''(\eta^t)\right] \in \mathbb{R}^{q\times q} \tag{6.35}$$

$$C_\theta(t,s) = \mathbb{E}\left[\theta^s \left(\theta^t\right)^\top\right] \in \mathbb{R}^{q\times q} \tag{6.36}$$

$$C_g(t,s) = \alpha\gamma^2 \mathbb{E}\left[s^s s^t l'(\eta^s) l'(\eta^t)^\top\right] \in \mathbb{R}^{q\times q} \tag{6.37}$$

initialized with $\boldsymbol{\nu}^0 = \mathbf{v}^0$*, where* u^t, ω^t *are Gaussian processes in* $\mathbb{R}^q$ *with covariances* $C_g(s,t), C_\theta(s,t)$*. Then, under assumptions (A1)-(A5), for any* $t \in \mathbb{N}$*, and any pseudo-Lipschitz functions* $\psi : \mathbb{R}^{q(t+1)} \to \mathbb{R}$ *and* $\phi : \mathbb{R}^{qt} \to \mathbb{R}$*:*

$$\frac{1}{d}\sum_{i=1}^{d} \psi((\mathbf{w}^0, ..., \mathbf{w}^t)_i) \xrightarrow[n,d\to\infty]{\text{w.h.p.}} \mathbb{E}\left[\psi(\theta^0, ..., \theta^t)\right], \tag{6.38}$$

$$\frac{1}{n}\sum_{j=1}^{n} \phi((\mathbf{r}^0, ..., \mathbf{r}^{t-1})_i) \xrightarrow[n,d\to\infty]{\text{w.h.p.}} \mathbb{E}\left[\phi(\eta^0, ..., \eta^{t-1})\right] \tag{6.39}$$

We remind that, to obtain the correlation with a planted vector $\mathbf{w}^*$ as in problem 6.2, we may use the same mapping from section 4.1 from [56].

6.4 Proof

In the next two subsections, we provide intuition on our proof method. Subsection 6.4.1 gives the exact asymptotic characterization of a gradient descent iteration with no regularization and a sample splitting assumption, where a fresh data matrix is sampled at each time step. This drastically simplifies the analysis and gives a simple result. We then move to the generic case, proving Theorem 9 using an induction on the variables $\mathbf{r}^t, \mathbf{u}^{t+1}$. The full induction step for $\mathbf{r}^t$ is given in the main text, while the induction step on $\mathbf{u}^{t+1}$, similar in spirit, is deferred to Appendix 6.6. Notations and useful lemmas are gathered in Appendix 6.5. We note that gradient-descent with sample-splitting was recently studied in [58] using Gaussian comparison inequalities.

6.4.1 A first example: gradient descent with sample splitting

Under the sample splitting assumption, the gradient descent iteration reads (for $q = 1$):

$$\forall t \in \mathbb{N}^* \quad \mathbf{w}^{t+1} = \mathbf{w}^t - \gamma^t (\mathbf{A}^t)^\top \nabla \mathbf{f}(\mathbf{A}^t \mathbf{w}^t) \tag{6.40}$$

where, for any $t \in \mathbb{N}$, $\mathbf{A}^t \in \mathbb{R}^{n\times d}$ is a matrix with i.i.d. Gaussian elements and variance $1/d$ independent on all other $\{\mathbf{A}^i\}_{i\neq t}$, $\gamma^t \in \mathbb{R}$ is a scalar step-size and $\mathbf{f}$ is a twice differentiable, deterministic function with pseudo-Lipschitz gradient $\nabla\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. We also assume that $\mathbf{f}$ is separable, with an elementwise operation f. The iteration is initialized with $\mathbf{w}^0 \in \mathbb{R}^d$, a random vector independent on $\mathbf{A}$ with i.i.d. subGaussian elements. Starting at $t = 0$, we condition equation (6.40) on (the sigma algebra generated by) $\mathbf{w}^0, \mathbf{A}^0\mathbf{w}^0$, and obtain, using lemma 2:

$$\mathbf{w}^1|_{\mathbf{w}^0,\mathbf{A}^0\mathbf{w}^0} = \mathbf{w}^0 - \gamma^0 \left(\mathbf{A}^0\mathbf{P}_{\mathbf{w}^0} + \tilde{\mathbf{A}}^0\mathbf{P}^{\perp}_{\mathbf{w}^0}\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \tag{6.41}$$

$$= \mathbf{w}^0 - \gamma^0\mathbf{w}^0\frac{1}{\|\mathbf{w}^0\|_2^2}\left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) - \gamma^0\mathbf{P}^{\perp}_{\mathbf{w}^0}\tilde{\mathbf{A}}^\top\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \tag{6.42}$$

Owing to the sample splitting assumption, the vector $\mathbf{A}^0\mathbf{w}^0$ has i.i.d. entries distributed according to $\mathcal{N}(0, \frac{1}{d}\|\mathbf{w}^0\|_2^2)$. We can then write

$$\frac{1}{\|\mathbf{w}^0\|_2^2}\left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) = \frac{1}{\frac{1}{d}\|\mathbf{w}^0\|_2^2}\frac{1}{d}\left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \tag{6.43}$$

The term $\frac{1}{d}\left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)$ is a scalar valued, pseudo-Lipschitz function of $\mathbf{A}^0\mathbf{w}^0$, and the subgaussian assumption on $\mathbf{w}^0$ ensures that the quantity $\frac{1}{d}\|\mathbf{w}^0\|_2^2$ converges almost surely to a finite, deterministic quantity. We can thus use lemma 1, the continuous mapping theorem (in the form of Slutsky's lemma), and Stein's lemma to show that

$$\frac{1}{\|\mathbf{w}^0\|_2^2}\left(\mathbf{A}^0\mathbf{w}^0\right)^\top \nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) \overset{\mathrm{P}}{\simeq} \alpha\mathbb{E}\left[f''(z^0)\right] \tag{6.44}$$

where $z^0 \sim \mathcal{N}(0, \rho^0)$ and we introduced $\rho^0 = \lim_{d\to\infty}\frac{1}{d}\|\mathbf{w}^0\|_2^2$. Turning to the part orthogonal to $\mathbf{w}^0$ and using the fact that the projector $\mathbf{P}_{\mathbf{w}^0}$ is of rank 1, the elements of $\tilde{\mathbf{A}}$ have variance $\frac{1}{d}$ and $\|\mathbf{w}^0\|_2^2$ is of order d, lemma 21 shows that

$$\frac{1}{\sqrt{d}}\left\|\mathbf{P}^{\perp}_{\mathbf{w}^0}\tilde{\mathbf{A}}^\top\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0) - (\tilde{\mathbf{A}}^0)^\top\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\right\|_2 \overset{\mathrm{P}}{\simeq} 0 \tag{6.45}$$

where $(\tilde{\mathbf{A}}^0)^\top\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)$ is a vector with i.i.d elements distributed as $\mathcal{N}(0, \frac{1}{d}\|\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\|_2^2)$. Once again, the function $\frac{1}{d}\|\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\|_2^2$ is scalar valued and pseudo-Lipschitz, thus lemma 1 and the continuous mapping theorem show that, for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ of order 2,

$$\frac{1}{d}\sum_{i=1}^{d}\psi(\left(\mathbf{P}^{\perp}_{\mathbf{w}^0}\tilde{\mathbf{A}}^\top\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\right)_i) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\psi(u^0)\right] \tag{6.46}$$

where $u^0 \sim \mathcal{N}(0, \tau_0)$ and we have introduced $\tau_0 = \lim_{n,d\to\infty}\frac{1}{d}\|\nabla\mathbf{f}(\mathbf{A}^0\mathbf{w}^0)\|_2^2 = \alpha\mathbb{E}\left[(f'(z^0))^2\right]$. Using these results, we may now lift the conditioning and use the definition of pseudo-Lipschitz function to recover the scalar equation describing the high-dimensional behaviour of $\mathbf{w}^1$. A straightforward induction shows that, for any $t \in \mathbb{N}$, the quantity $\frac{1}{d}\|\mathbf{w}^t\|_2^2$ is almost surely bounded, and the same conditioning argument can be applied along the sample splitting assumption to reach the following theorem

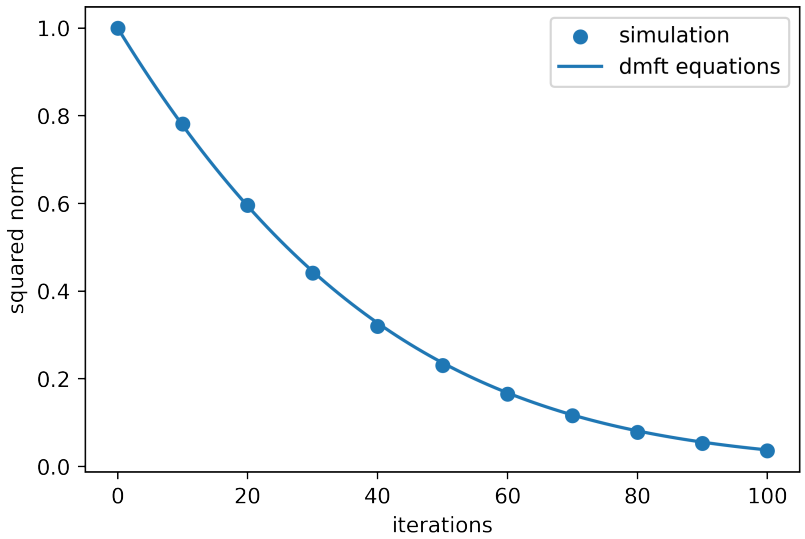


Figure 6.1: Gradient descent with sample splitting where $f'(z) = \tanh(z)$ Due to the regularity of the update function and sample splitting assumption, the concentration is very fast and . almost perfect matching is obtained between the theoretical and empirical curves with low dimensions (n=50,d=100) and no averaging.

Theorem 10. *(High-dimensional dynamics of gradient descent with sample splitting) Consider the iteration Eq. (6.40) with its set of assumptions described above. Define the following discrete-time one-dimensional stochastic process, initialized with a subgaussian random variable ω^0 with variance ρ^0:*

$$\omega^{t+1} = \left(1 - \gamma^t \alpha \mathbb{E}\left[f''(z^t)\right]\right)\omega^t + \gamma^t u^t \tag{6.47}$$

where $\rho^t = \mathbb{E}\left[(\omega^t)^2\right]$, $\tau^t = \alpha \mathbb{E}\left[(f'(z^t))^2\right]$. z^t, u^t are independent normal random variables with zero mean and respective variances ρ^t, τ^t. Then, for any $t \in \mathbb{N}$ and any pseudo-Lipschitz function of order 2 $\psi : \mathbb{R} \to \mathbb{R}$, the following holds

$$\lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi(w_i^t) \overset{\text{w.h.p.}}{=} \mathbb{E}\left[\psi(\omega^t)\right] \tag{6.48}$$

We have obtained a full description of the asymptotic distribution of $\mathbf{w}^t$ in terms of a scalar equation. The sample splitting assumption however, is unrealistic. Let us move to the generic case that corresponds to the usual gradient descent.

6.4.2 The general case

Without the sample splitting assumption, the iterates $\mathbf{x}^t$ and the design matrix $\mathbf{X}$ are correlated at each time step and thus there is no simple concentration towards a markovian model. We need to account for the correlation beyond the previous time step, leading to the appearance of memory kernels. Recall the dynamics (6.4-6.5), where we introduce an additional intermediate variable

$\mathbf{m}^t = \mathbf{g}(\mathbf{r}^t)$:

$$\mathbf{v}^{t+1} = \mathbf{h}^t(\left\{\mathbf{v}^k\right\}_{k=0}^t) + \mathbf{X}^\top \mathbf{m}^t \tag{6.49}$$

$$\mathbf{m}^t = \mathbf{g}^t(\mathbf{r}^t) \tag{6.50}$$

$$\mathbf{r}^t = \mathbf{X}\sum_{k=0}^t \mathbf{v}^t \tag{6.51}$$

The proof is done by induction on t.

Initialization At initialization, we have

$$\mathbf{v}^0 = \mathbf{w}^0 \sim P_{\mathbf{v}_0} \quad \text{by definition} \quad \mathbf{v}^0 = \boldsymbol{\nu}^0 \tag{6.52}$$

$$\mathbf{r}^0 = \mathbf{X}\mathbf{v}^0 \xrightarrow[n,d\to\infty]{Plk} \boldsymbol{\eta}^0 \sim \mathcal{N}(0, C_\theta(0,0)\otimes \mathbf{I}_n) \quad \text{where} \quad C_\theta(0,0) = \lim_{d\to\infty}\frac{1}{d}\mathbb{E}\left[(\mathbf{v}^0)^\top \mathbf{v}^0\right] \tag{6.53}$$

Where the second line is a direct consequence of the independence of the initialization with the data matrix and the continuous mapping theorem. Note that

$$\boldsymbol{\eta}^0 = \boldsymbol{\omega}^0 \sim \mathcal{N}(0, C_\theta(0,0)\otimes \mathbf{I}_n) \tag{6.54}$$

Let's do the step for $\mathbf{v}^1$.

$$\mathbf{v}^1 = \mathbf{h}^0\left(\mathbf{v}^0\right) + \mathbf{X}^\top \mathbf{m}^0 \tag{6.55}$$

conditioning on the $\sigma-$algebra $\mathfrak{S}^0 = \sigma\left(\mathbf{v}^0, \mathbf{r}^0\right)$ and using Lemma 2, we obtain

$$\mathbf{v}^1|_{\mathfrak{S}^0} = \mathbf{h}^0(\mathbf{v}^0) + \left(\mathbf{X}|_{\mathfrak{S}^0}\right)^\top \mathbf{m}^0 \tag{6.56}$$

$$= \mathbf{h}^0(\mathbf{v}^0) + \left(\mathbf{P}_{\mathbf{v}^0}\mathbf{X}^\top + \mathbf{P}^\perp_{\mathbf{v}^0}\tilde{\mathbf{X}}^\top\right)\mathbf{m}^0 \tag{6.57}$$

$$= \mathbf{h}^0(\mathbf{v}^0) + \mathbf{v}^0\left((\mathbf{v}^0)^\top\mathbf{v}^0\right)^{-1}(\mathbf{v}^0\mathbf{X})^\top g^0(\mathbf{r}^0) + \mathbf{P}^\perp_{\mathbf{v}^0}\tilde{\mathbf{X}}^\top\mathbf{m}^0 \tag{6.58}$$

$$\xrightarrow[n,d\to\infty]{Plk} \mathbf{h}^0(\mathbf{v}^0) + \mathbf{v}^0\left(\frac{1}{d}(\mathbf{v}^0)^\top\mathbf{v}^0\right)^{-1}\frac{1}{d}(\boldsymbol{\eta}^0)^\top\mathbf{g}^0(\boldsymbol{\eta}^0) + \mathbf{u}^0 \tag{6.59}$$

$$\xrightarrow[n,d\to\infty]{Plk} \mathbf{h}^0(\mathbf{v}^0) + \mathbf{v}^0\left(C_\theta(0,0)\right)^{-1}C_\theta(0,0)\frac{1}{d}\sum_{i=1}^n \mathbb{E}\left[\frac{\partial g_i^0}{\partial \eta_i^0}(\boldsymbol{\eta}^0)\right] + \mathbf{u}^0 \tag{6.60}$$

$$\xrightarrow[n,d\to\infty]{Plk} \mathbf{h}^0(\mathbf{v}^0) + \mathbf{v}^0\Gamma^0 + \mathbf{u}^0 \tag{6.61}$$

where $C_g(0,0) = \lim_{d\to\infty}\frac{1}{d}\mathbb{E}\left[\mathbf{g}^0(\boldsymbol{\eta}^0)^\top\mathbf{g}^0(\boldsymbol{\eta}^0)\right]$, $\mathbf{u}^0 \sim \mathcal{N}(0, C_g(0,0)\otimes\mathbf{I}_d)$ and we remind $\Gamma^0 = \lim_{d\to\infty}\frac{1}{d}\sum_{i=1}^n \mathbb{E}\left[\frac{\partial g_i^0}{\partial \eta_i^0}(\eta^0)\right]$. The convergence of the term $\mathbf{P}^\perp_{\mathbf{v}^0}\tilde{\mathbf{X}}^\top\mathbf{m}^0$ to u^0 comes from lemma 21 while the appearance of Γ^0 is due to Stein's lemma 17. This concludes the initialization.

Induction Assume that Theorem 9 is verified up to time t, i.e. for all iterates up to $\mathbf{r}^{t-1}, \mathbf{v}^t$. We prove the property for $\mathbf{r}^t, \mathbf{v}^{t+1}$.

We shall condition on the σ-algebra generated by $\mathbf{v}^0, ..., \mathbf{v}^t, \mathbf{r}^0, ..., \mathbf{r}^{t-1}$, denoted $\mathfrak{S}^t$. A short induction and application of the Doob-Dynkin Lemma show that this σ-algebra is the same as that generated by $\mathbf{v}^0, \mathbf{X}^\top \mathbf{m}^0, ..., \mathbf{X}^\top \mathbf{m}^{t-1}, \mathbf{X}\mathbf{w}^0, ..., \mathbf{X}\mathbf{w}^{t-1}$, where we remind that $\mathbf{w}^s = \sum_{k=0}^{s} \mathbf{v}^0$ with $\mathbf{w}^0 = \mathbf{v}^0$. We define the matrices

$$\mathbf{M}_{t-1} = \left[\mathbf{m}^0 | \mathbf{m}^1 | ... | \mathbf{m}^{t-1}\right], \mathbf{W}_{t-1} = \left[\mathbf{w}^0 | \mathbf{w}^1 | ... | \mathbf{w}^{t-1}\right] \tag{6.62}$$

Starting with $\mathbf{r}^t$, we may write

$$\begin{aligned}
\mathbf{r}^t|_{\mathfrak{S}^t} &= \left(\mathbf{X}\sum_{k=0}^{t} \mathbf{v}^k\right)|_{\mathfrak{S}^t} && (6.63)\\
&= \mathbf{r}_{t-1} + \mathbf{X}|_{\mathfrak{S}^t}\mathbf{v}^t && (6.64)\\
&= \mathbf{r}^{t-1} + \left(\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{X} + \mathbf{X}\mathbf{P}_{\mathbf{W}_{t-1}} - \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{X}\mathbf{P}_{\mathbf{W}_{t-1}} + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{X}}\mathbf{P}^{\perp}_{\mathbf{W}_{t-1}}\right)\mathbf{v}^t && (6.65)\\
&= \mathbf{r}^{t-1} + \left(\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{X}\mathbf{P}^{\perp}_{\mathbf{W}_{t-1}} + \mathbf{X}\mathbf{P}_{\mathbf{W}_{t-1}} + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{X}}\mathbf{P}^{\perp}_{\mathbf{W}_{t-1}}\right)\mathbf{v}^t && (6.66)\\
& && (6.67)
\end{aligned}$$

where $\tilde{\mathbf{X}}$ is a copy of $\mathbf{X}$ independent on $\mathfrak{S}^t$.
At this point, we introduce an assumption guaranteeing the projectors are well-defined, in similar fashion to [37, 110]. It will be relaxed at the end of the proof, in Appendix 6.6.1.

Non-degeneracy assumption We say that the iteration SGD satisfies the non-degeneracy assumption if :

- almost surely, for all t and all $N \geqslant t$, $\mathbf{M}_{t-1}, \mathbf{W}_{t-1}$ have full column rank.
- for all t, there exists some constant $c_{M,t}, c_{W,t} > 0$—independent of n—such that almost surely, there exists n_0 (random) such that, for $n \geqslant n_0$, $\sigma_{\min}(\mathbf{M}_{t-1})/\sqrt{N} \geqslant c_{M,t} > 0$ and $\sigma_{\min}(\mathbf{W}_{t-1})/\sqrt{N} \geqslant c_{W,t} > 0$.

Let's look at each term separately, starting with

$$\begin{aligned}
\mathbf{X}\mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{v}^t &= \mathbf{X}\mathbf{W}_{t-1}\left(\mathbf{W}_{t-1}^\top\mathbf{W}_{t-1}\right)^{-1}\mathbf{W}_{t-1}^\top\mathbf{v}^t && (6.68)\\
&= \left[\mathbf{r}^0 | \mathbf{r}^1 | ... | \mathbf{r}^{t-1}\right]\boldsymbol{\alpha}^t && (6.69)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\alpha}^t &= \left(\mathbf{W}_{t-1}^\top\mathbf{W}_{t-1}\right)^{-1}\mathbf{W}_{t-1}^\top\mathbf{v}^t \in \mathbb{R}^{tq\times q}\\
&= \left(\frac{1}{d}\mathbf{W}_{t-1}^\top\mathbf{W}_{t-1}\right)^{-1}\frac{1}{d}\mathbf{W}_{t-1}^\top\mathbf{v}^t && (6.70)
\end{aligned}$$

which is a low-dimensional $(tq \times q)$ pseudo-Lipschitz function of $\mathbf{v}^0, ..., \mathbf{v}^t$. Thus, owing to the induction hypothesis, non-degeneracy assumption and lemma, $\boldsymbol{\alpha}^t$ converges to a determinstic limit $\boldsymbol{\alpha}^{t,*} \in \mathbb{R}^{tq\times q}$ representing the coefficients of the projection of the columns of $\mathbf{v}^t$ onto the subspace

spanned by the columns of $\mathbf{W}_{t-1}$. Using the induction hypothesis and non-degeneracy assumption, we also have

$$\boldsymbol{\alpha}^{t,*} = \lim_{n\to\infty} \left(\frac{1}{d}\boldsymbol{\Theta}_{t-1}^\top \boldsymbol{\Theta}_{t-1}\right)^{-1} \frac{1}{d}\boldsymbol{\Theta}_{t-1}^\top \left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right) \tag{6.71}$$

$$\overset{\mathrm{P}}{\simeq} \lim_{n\to\infty} \mathbb{E}\left[\left(\frac{1}{d}\boldsymbol{\Theta}_{t-1}^\top \boldsymbol{\Theta}_{t-1}\right)^{-1} \frac{1}{d}\boldsymbol{\Theta}_{t-1}^\top \left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right)\right] \tag{6.72}$$

where we defined the matrix $\boldsymbol{\Theta}_{t-1} = \left[\boldsymbol{\theta}^0|\boldsymbol{\theta}^1...|\boldsymbol{\theta}^{t-1}\right]$. We may then write

$$\mathbf{X}\mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{v}^t \xrightarrow[n,d\to\infty]{Plk} \sum_{k=0}^{t-1} \boldsymbol{\eta}^k \alpha_k^{t,*} \tag{6.73}$$

where each $\alpha_k^{t,*} \in \mathbb{R}^{q\times q}$ and $\boldsymbol{\eta}^k \in \mathbb{R}^{n\times q}$ are defined in Theorem 9.

Moving to the next term,

$$\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{X}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t = \mathbf{M}_{t-1}\left(\mathbf{M}_{t-1}^\top\mathbf{M}_{t-1}\right)^{-1}\mathbf{M}_{t-1}^\top\mathbf{X}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t \tag{6.74}$$

$$= \mathbf{M}_{t-1}\left(\frac{1}{d}\mathbf{M}_{t-1}^\top\mathbf{M}_{t-1}\right)^{-1}\frac{1}{d}\mathbf{M}_{t-1}^\top\mathbf{X}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t \tag{6.75}$$

where, using the definition of iteration Eq. (6.4-6.5)

$$\frac{1}{d}\mathbf{M}_{t-1}^\top\mathbf{X}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t = \frac{1}{d}\left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{v}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\left\{\mathbf{v}^k\right\}_{k=0}^{t-1})\right]^\top \mathbf{v}^t$$
$$- \frac{1}{d}\left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{v}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\left\{\mathbf{v}^k\right\}_{k=0}^{t-1})\right]^\top \mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{v}^t \tag{6.76}$$

and

$$\frac{1}{d}\left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{v}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\left\{\mathbf{v}^k\right\}_{k=0}^{t-1})\right]^\top \mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{v}^t = \tag{6.77}$$

$$= \frac{1}{d}\left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{v}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\left\{\mathbf{v}^k\right\}_{k=0}^{t-1})\right]^\top \mathbf{W}_{t-1}\left(\frac{1}{d}\mathbf{W}_{t-1}^\top\mathbf{W}_{t-1}\right)^{-1}\frac{1}{d}\mathbf{W}_{t-1}^\top\mathbf{v}^t \tag{6.78}$$

Using the induction hypothesis and pseudo-Lipschitz convergence lemma 1,

$$\frac{1}{d}\left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{w}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\left\{\mathbf{v}^k\right\}_{k=0}^{t-1})\right]^\top \mathbf{W}_{t-1} \overset{\mathrm{P}}{\simeq}$$
$$\frac{1}{d}\left[\Gamma^0\boldsymbol{\theta}^0 + \mathbf{u}^0|...|\Gamma^{t-1}\boldsymbol{\theta}^{t-1} + \sum_{k=0}^{t-2}\boldsymbol{\theta}^k R_l(t-1,k) + \mathbf{u}^{t-1}\right]^\top \boldsymbol{\Theta}_{t-1} \tag{6.79}$$

$$= \underbrace{\frac{1}{d}\left[\Gamma^0\boldsymbol{\theta}^0|...|\Gamma^{t-1}\boldsymbol{\theta}^{t-1} + \sum_{k=0}^{t-2}\boldsymbol{\theta}^k R_l(t-1,k)\right]^\top}_{\in \mathrm{span}(\boldsymbol{\Theta}_{t-1})}\boldsymbol{\Theta}_{t-1} + \frac{1}{d}\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right]^\top\boldsymbol{\Theta}_{t-1} \tag{6.80}$$

and

$$\frac{1}{d}\mathbf{W}_{t-1}^\top\mathbf{v}^t \overset{\mathrm{P}}{\simeq} \frac{1}{d}\boldsymbol{\Theta}_{t-1}^\top\left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right) \tag{6.81}$$

where we also have

$$\frac{1}{d}\mathbf{W}_{t-1}^\top \mathbf{W}_{t-1} \overset{\mathrm{P}}{\simeq} \frac{1}{d}\mathbf{\Theta}_{t-1}^\top \mathbf{\Theta}_{t-1} \succ 0_{tq\times tq} \quad \text{w.h.p.} \tag{6.82}$$

We thus reach

$$\frac{1}{d}\left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{w}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\mathbf{w}^{t-1})\right]^\top \mathbf{v}^t \overset{\mathrm{P}}{\simeq}$$

$$\frac{1}{d}\left[\mathbf{\Gamma}^0\boldsymbol{\theta}^0 + \mathbf{u}^0|...|\mathbf{\Gamma}^{t-1}\boldsymbol{\theta}^{t-1} + \sum_{k=0}^{t-2}\boldsymbol{\theta}^k R_l(t-1,k) + \mathbf{u}^{t-1}\right]^\top \left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right) \tag{6.83}$$

and $$\frac{1}{d}\left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{w}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\mathbf{w}^{t-1})\right]^\top \mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{v}^t \overset{\mathrm{P}}{\simeq}$$

$$\frac{1}{d}\left[\mathbf{\Gamma}^0\boldsymbol{\theta}^0 + \mathbf{u}^0|...|\mathbf{\Gamma}^{t-1}\boldsymbol{\theta}^{t-1} + \sum_{k=0}^{t-2}\theta^k R_l(t-1,k) + \mathbf{u}^{t-1}\right]^\top \tag{6.84}$$

$$\mathbf{\Theta}_{t-1}\left(\frac{1}{d}\mathbf{\Theta}_{t-1}^\top\mathbf{\Theta}_{t-1}\right)^{-1}\frac{1}{d}\mathbf{\Theta}_{t-1}^\top\left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right) \tag{6.85}$$

which, when combined, leads to

$$\frac{1}{d}\mathbf{M}_{t-1}^\top \mathbf{X}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t \overset{\mathrm{P}}{\simeq} \frac{1}{d}\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right]^\top\left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right) - \frac{1}{d}\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right]^\top \mathbf{P}_{\mathbf{\Theta}_{t-1}}\left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right) \tag{6.86}$$

$$\overset{\mathrm{P}}{\simeq} \frac{1}{d}\mathbb{E}\left[\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right]^\top\left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right)\right] - \frac{1}{d}\mathbb{E}\left[\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right]^\top\mathbf{\Theta}_{t-1}\right]\boldsymbol{\alpha}^{t,*} \tag{6.87}$$

Now, remembering the equation defining $\boldsymbol{\theta}^s$ for any $0 \leqslant s \leqslant t$, we may use Stein's lemma 17 to obtain

$$\forall\, 0 \leqslant r,s \leqslant t \quad \frac{1}{d}(\mathbf{u}^r)^\top\boldsymbol{\theta}^s(\mathbf{u}^0,\mathbf{u}^1,...,\mathbf{u}^{s-1}) \overset{\mathrm{P}}{\simeq} \frac{1}{d}\sum_{i=0}^{s-1}C_g(i,r)\sum_{j=1}^{d}\mathbb{E}\left[\frac{\partial\theta_j^s}{\partial u_j^i}\right]$$

$$\overset{\mathrm{P}}{\simeq} \sum_{i=0}^{s-1}C_g(i,r)R_\theta(s,i) \tag{6.88}$$

Letting $\mathbf{C}_{g,t}$ be the $tq\times tq$ covariance matrix of the lines of $\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right] \in \mathbb{R}^{d\times tq}$ for any t, we can write

$$\frac{1}{d}\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right]^\top\boldsymbol{\theta}^t \overset{\mathrm{P}}{\simeq} \mathbf{C}_{g,t}\begin{bmatrix}\frac{1}{d}\sum_{j=1}^d\mathbb{E}\left[\frac{\partial\theta_j^t}{\partial u_j^0}\right]\\ ...\\ \frac{1}{d}\sum_{j=1}^d\mathbb{E}\left[\frac{\partial\theta_j^t}{\partial u_j^{t-1}}\right]\end{bmatrix} \tag{6.89}$$

$$= \mathbf{C}_{g,t}\begin{bmatrix}R_\theta(t,0)\\ ...\\ R_\theta(t,t-1)\end{bmatrix} = \mathbf{C}_{g,t}\mathbf{R}_{\theta,t} \tag{6.90}$$

where we defined the $tq \times q$ matrix $\mathbf{R}_{\theta,t} = \begin{bmatrix} R_\theta(t,0) \\ ... \\ R_\theta(t,t-1) \end{bmatrix}$.

Similarly, for any $0 \leqslant s \leqslant t$

$$\frac{1}{d}\left[\mathbf{u}^0|...|\mathbf{u}^{t-1}\right]^\top \boldsymbol{\theta}^s \overset{\mathrm{P}}{\simeq} \mathbf{C}_{g,t} \begin{bmatrix} \frac{1}{d}\sum_{j=1}^d \mathbb{E}\left[\frac{\partial \theta_j^s}{\partial u_j^0}\right] \\ ... \\ \frac{1}{d}\sum_{j=1}^d \mathbb{E}\left[\frac{\partial \theta_j^s}{\partial u_j^{s-1}}\right] \\ 0 \\ ... \\ 0 \end{bmatrix} = \mathbf{C}_{g,t}\mathbf{R}_{\theta,s} \tag{6.91}$$

where the zeroes come from the fact that θ^s is not an algebraic function of the u^l for $l \geqslant s$, which is coherent with the causality from the physics approach, even though the Gaussian process u^l is correlated across all $0 \leqslant l \leqslant t-1$. Also, due to the induction hypothesis

$$\frac{1}{d}\mathbf{M}_{t-1}^\top \mathbf{M}_{t-1} \overset{\mathrm{P}}{\simeq} \mathbf{C}_{g,t} \tag{6.92}$$

We then have, using the non-degeneracy assumption

$$\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{X}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t \xrightarrow[n,d\to\infty]{Plk} \tag{6.93}$$

$$\mathbf{M}_{t-1}\left(\frac{1}{d}\mathbf{M}_{t-1}^\top\mathbf{M}_{t-1}\right)^{-1}\mathbf{C}_{g,t}\left(\mathbf{R}_{\theta,t} - \mathbf{R}_{\theta,t-1} - [\mathbf{R}_{\theta,0}|\mathbf{R}_{\theta,1}|...|\mathbf{R}_{\theta,t-1}]\,\boldsymbol{\alpha}^{t,*}\right) \tag{6.94}$$

$$\xrightarrow[n,d\to\infty]{Plk} \mathbf{M}_{t-1}\left(\mathbf{R}_{\theta,t} - \mathbf{R}_{\theta,t-1} - [\mathbf{R}_{\theta,0}|\mathbf{R}_{\theta,1}|...|\mathbf{R}_{\theta,t-1}]\,\boldsymbol{\alpha}^{t,*}\right) \tag{6.95}$$

Combining this with the induction hypothesis and lemma 1 and 21 , we may write

$$\mathbf{r}^t|_{\mathfrak{S}^t} \xrightarrow[n,d\to\infty]{Plk} \mathbf{r}^{t-1} + \sum_{k=0}^{t-1}\mathbf{r}^k\alpha_k^{t,*} + \mathbf{M}_{t-1}\left(\mathbf{R}_{\theta,t} - \mathbf{R}_{\theta,t-1} - [\mathbf{R}_{\theta,0}|\mathbf{R}_{\theta,1}|...|\mathbf{R}_{\theta,t-1}]\,\boldsymbol{\alpha}^{t,*}\right) \tag{6.96}$$

$$+ \tilde{\mathbf{X}}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t$$

$$\xrightarrow[n,d\to\infty]{Plk} \sum_{l=0}^{t-2}\mathbf{g}^l(\boldsymbol{\eta}^l)R_\theta(t-1,l) + \boldsymbol{\omega}^{t-1} + \sum_{k=0}^{t-1}\left(\sum_{l'=0}^{k}\mathbf{g}^{l'}(\boldsymbol{\eta}^{l'})R_\theta(k,l') + \boldsymbol{\omega}^k\right)\alpha_k^{t,*}$$

$$+ \mathbf{M}_{t-1}\left(\mathbf{R}_{\theta,t} - \mathbf{R}_{\theta,t-1} - [\mathbf{R}_{\theta,0}|\mathbf{R}_{\theta,1}|...|\mathbf{R}_{\theta,t-1}]\,\boldsymbol{\alpha}^{t,*}\right) + \tilde{\mathbf{X}}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t \tag{6.97}$$

where we used lemma 21 to remove the projector $\mathbf{P}^\perp_{\mathbf{M}_{t-1}}$ in the term $\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\tilde{\mathbf{X}}\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\mathbf{v}^t$. Recalling the definition of $\mathbf{m}^s = \mathbf{g}^s(\mathbf{r}^s)$, the induction hypothesis gives, for any $0 \leqslant s \leqslant t$,

$$\mathbf{M}_{t-1}\mathbf{R}_{\theta,s} \xrightarrow[n,d\to\infty]{Plk} \sum_{l=0}^{s-1}\mathbf{g}^l(\boldsymbol{\eta}^l)R_\theta(s,l) \tag{6.98}$$

All memory terms associated to $\mathbf{R}_{\theta,s}$ for $s \leqslant t-1$ thus simplify in Eq.(6.97), leading to

$$\mathbf{r}^t|_{\mathfrak{S}^t} \xrightarrow[n,d\to\infty]{Plk} \sum_{k=0}^{t-1}\mathbf{g}^k(\boldsymbol{\eta}^k)R_\theta(t,k) + \sum_{k=0}^{t-1}\boldsymbol{\omega}^k\alpha_k^{t,*} + \boldsymbol{\omega}^{t-1} + \tilde{\boldsymbol{\omega}}^t \tag{6.99}$$

where $\tilde{\boldsymbol{\omega}}^t \sim \mathcal{N}(0, \mathbf{C}^{\perp}_{\mathbf{v},t} \otimes \mathbf{I}_n)$, and $\mathbf{C}^{\perp}_{\mathbf{v},t} = \lim_{d\to\infty} \frac{1}{d}\left(\mathbf{P}^{\perp}_{\mathbf{W}_{t-1}}\mathbf{v}^t\right)^\top \left(\mathbf{P}^{\perp}_{\mathbf{W}_{t-1}}\mathbf{v}^t\right)$. We thus recover the correct memory term. We are left with checking that the Gaussian process term has the right covariance. Define

$$\boldsymbol{\omega}^t = \sum_{k=0}^{t-1} \boldsymbol{\omega}^k \alpha_k^{t,*} + \boldsymbol{\omega}^{t-1} + \tilde{\boldsymbol{\omega}}^t. \tag{6.100}$$

Which is indeed a Gaussian random vector (with elements in $\mathbb{R}^q$). To check that this is the correct covariance, we start by noticing that, for any $s < t$ Theorem 9 states that:

$$\frac{1}{d}(\mathbf{w}^s)^\top \mathbf{w}^t = \frac{1}{d}(\mathbf{w}^s)^\top \mathbf{w}^{t-1} + \frac{1}{d}(\mathbf{w}^s)^\top \mathbf{v}^t \tag{6.101}$$

$$\xrightarrow[n,d\to\infty]{Plk} C_\theta(s, t-1) + \frac{1}{d}(\mathbf{w}^s)^\top \mathbf{v}^t \tag{6.102}$$

Then, using the induction hypothesis and the fact that $\tilde{\boldsymbol{\omega}}^t$ is independent from any $\boldsymbol{\omega}^s, \forall s < t$:

$$\frac{1}{d}\mathbb{E}\left[(\boldsymbol{\omega}^s)^\top \boldsymbol{\omega}^t\right] = \frac{1}{d}\sum_{k=0}^{t-1} \mathbb{E}\left[(\boldsymbol{\omega}^s)^\top \boldsymbol{\omega}^s\right] \alpha_k^{t,*} + \frac{1}{d}\mathbb{E}\left[(\boldsymbol{\omega}^s)^\top \boldsymbol{\omega}^{t-1}\right] \tag{6.103}$$

$$= \sum_{k=0}^{t-1} C_\theta(s,k)\alpha_k^{t,*} + C_\theta(s, t-1) \tag{6.104}$$

$$\xrightarrow[n,d\to\infty]{Plk} \frac{1}{d}(\mathbf{w}^s)^\top \mathbf{W}_{t-1}\left(\mathbf{W}_{t-1}^\top \mathbf{W}_{t-1}\right)^{-1} \mathbf{W}_{t-1}^\top \mathbf{v}^t + C_\theta(s, t-1) \tag{6.105}$$

$$= \frac{1}{d}\left(\mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{w}^s\right)^\top \mathbf{v}^t + C_\theta(s, t-1) \tag{6.106}$$

$$\xrightarrow[n,d\to\infty]{Plk} \frac{1}{d}(\mathbf{w}^s)^\top \mathbf{v}^t + C_\theta(s, t-1) \tag{6.107}$$

We then check for $s = t$, noticing that

$$\frac{1}{d}(\mathbf{w}^t)^\top \mathbf{w}^t = \frac{1}{d}\left(\mathbf{w}^{t-1} + \mathbf{v}^t\right)^\top \left(\mathbf{w}^{t-1} + \mathbf{v}^t\right) \tag{6.108}$$

$$\xrightarrow[n,d\to\infty]{Plk} C_\theta(t-1, t-1) + \frac{1}{d}(\mathbf{v}^t)^\top \left(\mathbf{w}^{t-1} + \mathbf{v}^t\right) \tag{6.109}$$

$$\frac{1}{d}\mathbb{E}\left[(\boldsymbol{\omega}^t)^\top \boldsymbol{\omega}^t\right] = \frac{1}{d}\mathbb{E}\left[\left(\sum_{k=0}^{t-1} \boldsymbol{\omega}^k \alpha_k^{t,*} + \boldsymbol{\omega}^{t-1} + \tilde{\boldsymbol{\omega}}^t\right)^\top \left(\sum_{k=0}^{t-1} \boldsymbol{\omega}^k \alpha_k^{t,*} + \boldsymbol{\omega}^{t-1} + \tilde{\boldsymbol{\omega}}^t\right)\right] \tag{6.110}$$

$$= C_\theta(t-1, t-1) + \sum_{k,k'=0}^{t-1} (\alpha_{k'}^{t,*})^\top C_\theta(k,k')\alpha_k^{t,*} + 2\sum_{k=0}^{t-1} C_\theta(t-1,k)\alpha_k^{t,*} + C_{v,t}^{\perp} \tag{6.111}$$

$$\xrightarrow[n,d\to\infty]{Plk} \frac{1}{d}(\mathbf{w}^{t-1})^\top \mathbf{w}^{t-1} + \frac{1}{d}\left(\mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{v}^t\right)^\top \left(\mathbf{P}_{\mathbf{W}_{t-1}}\mathbf{v}^t\right) + \frac{1}{d}\left(\mathbf{P}^{\perp}_{\mathbf{W}_{t-1}}\mathbf{v}^t\right)^\top \left(\mathbf{P}^{\perp}_{\mathbf{W}_{t-1}}\mathbf{v}^t\right) \tag{6.112}$$

$$+ 2\frac{1}{d}\mathbf{w}_{t-1}^\top \mathbf{v}^t \tag{6.113}$$

$$\xrightarrow[n,d\to\infty]{Plk} \frac{1}{d}\left(\mathbf{w}^{t-1} + \mathbf{v}^t\right)^\top \left(\mathbf{w}^{t-1} + \mathbf{v}^t\right) \tag{6.114}$$

We thus recover the correct covariance and the statement is proven for $\mathbf{r}^t$. The rest of the proof consists in completing the induction on $\mathbf{u}^{t+1}$, in similar fashion to what has been presented for $\mathbf{r}^t$, and relaxing the non-degeneracy assumption using an existing argument from [37, 110]. The detail is given in appendix 6.6.

6.5 Useful definitions and probability results

Here we reproduce some definitions and useful intermediate lemmas from [28, 110] without proof.

Notations We adopt the same notations as in [110]. We introduce the folowing notion of convergence to lighten notations.

Definition 12 (pseudo-Lipschitz convergence)**.** *We say that the sequence of random matrices* $\mathbf{X}_n \in \mathbb{R}^{d\times q}$ *converges in the pseudo-Lipschitz sense of order* k *to* $\mathbf{Z} \in \mathbb{R}^{n\times q}$ *and denote* $\mathbf{X}_n \xrightarrow[n,d\to\infty]{Plk} \mathbf{Z}$ *if, for any sequence of uniformly pseudo-Lipschitz functions* $\phi_n : \mathbb{R}^{d\times q} \to \mathbb{R}$ *of order* k*, the following holds*

$$\lim_{n\to\infty} |\phi_n(\mathbf{X}_n) - \phi_n(\mathbf{Z})| \overset{\text{w.h.p.}}{=} 0 \tag{6.115}$$

where both $n, d \to \infty$ *with fixed ratio* α*, and* q *remains finite.*

Definition 1 shows that, if for all k $\left(\frac{\|\mathbf{X}_n\|_F}{\sqrt{d}}\right)^k, \left(\frac{\|\mathbf{Z}\|_F}{\sqrt{d}}\right)^k$ are bounded and the following holds $\frac{1}{\sqrt{N}}\|\mathbf{X}_n - \mathbf{Z}\|_F \xrightarrow[n,d\to\infty]{\text{w.h.p.}} 0$, we have pseudo-Lipschitz convergence of order k of $\mathbf{X}_n$ towards $\mathbf{Z}$. It is also straightforward to show that pseudo-Lipschitz convergence is stable under addition and multiplication by deterministic matrices. Note that, when separable test functions ϕ_n are used, pseudo-Lipschitz convergence is equivalent to convergence in the Wasserstein space of order k [289]. We now state the necessary assumptions for our main result to hold.

6.6 Proof of Theorem 9

This appendix provides the details for the second part of the induction proving Theorem 9, the first part of which we presented in section 6.4.2. At this point we completed the induction step for the variable $\mathbf{r}^t$. Moving to $\mathbf{v}^{t+1}$, we now need to condition on $\mathfrak{S}^t$ but also on $\mathbf{r}^t$ for which we just proved the statement, which amounts to conditioning on the values of $\mathbf{v}^0, \mathbf{X}^\top\mathbf{m}^0, ..., \mathbf{X}^\top\mathbf{m}^{t-1}$
$\mathbf{X}\mathbf{w}^0, ..., \mathbf{W}^t$. We will then perform orthogonal decomposition on the subspaces spanned by the matrices

$$\mathbf{M}_{t-1} = \left[\mathbf{m}^0|\mathbf{m}^1|...|\mathbf{m}^{t-1}\right], \mathbf{W}_t = \left[\mathbf{w}^0|\mathbf{w}^1|...|\mathbf{w}^{t-1}|\mathbf{w}^t\right] \tag{6.116}$$

where $\mathbf{M}_{t-1} \in \mathbb{R}^{n\times tq}$ and $\mathbf{W}_t \in \mathbb{R}^{d\times tq}$ We obtain

$$\mathbf{v}^{t+1}|_{\mathfrak{S}^t,\mathbf{r}^t} = \mathbf{h}^t(\left\{\mathbf{v}^k\right\}_{k=0}^t) + \mathbf{X}|^\top_{\mathfrak{S}^t,\mathbf{r}^t}\mathbf{m}^t \tag{6.117}$$

$$= \mathbf{h}^t(\left\{\mathbf{v}^k\right\}_{k=0}^t) + \left(\mathbf{X}^\top\mathbf{P}_{\mathbf{M}_{t-1}} + \mathbf{P}_{\mathbf{W}_t}\mathbf{X}^\top - \mathbf{P}_{\mathbf{W}_t}\mathbf{X}^\top\mathbf{P}_{\mathbf{M}_{t-1}} + \mathbf{P}^\perp_{\tilde{\mathbf{W}}}\tilde{\mathbf{X}}^\top\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\right)\mathbf{m}^t \tag{6.118}$$

$$= \mathbf{h}^t(\left\{\mathbf{v}^k\right\}_{k=0}^t) + \mathbf{X}^\top\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}_{\mathbf{W}_t}\mathbf{X}^\top\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}^\perp_{\tilde{\mathbf{W}}_t}\tilde{\mathbf{X}}^\top\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \tag{6.119}$$

As before, we treat each term separately, starting with

$$\mathbf{X}^\top\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{m}^t = \mathbf{X}^\top\mathbf{M}_{t-1}\left(\mathbf{M}_{t-1}^\top\mathbf{M}_{t-1}\right)^{-1}\mathbf{M}_{t-1}^\top\mathbf{m}^t \tag{6.120}$$

$$= \left[\mathbf{v}^1 - \mathbf{h}^0(\mathbf{w}^0)|...|\mathbf{v}^t - \mathbf{h}^{t-1}(\left\{\mathbf{v}^k\right\}_{k=0}^{t-1}))\right]^\top \boldsymbol{\beta}^t \tag{6.121}$$

$$= \sum_{k=0}^{t-1}\left(\mathbf{v}^{k+1} - \mathbf{h}^t(\left\{\mathbf{v}^l\right\}_{l=0}^k))\right)\boldsymbol{\eta}_k^t \tag{6.122}$$

where

$$\boldsymbol{\beta}^t = \left(\mathbf{M}_{t-1}^\top \mathbf{M}_{t-1}\right)^{-1} \mathbf{M}_{t-1}^\top \mathbf{m}^t \tag{6.123}$$

$$= \left(\frac{1}{n}\mathbf{M}_{t-1}^\top \mathbf{M}_{t-1}\right)^{-1} \frac{1}{n}\mathbf{M}_{t-1}^\top \mathbf{m}^t \tag{6.124}$$

$$\overset{\mathrm{P}}{\simeq} \boldsymbol{\beta}^{t,*} \in \mathbb{R}^{tq\times q} \tag{6.125}$$

with deterministic $\boldsymbol{\beta}^{t,*}$, where we used the non-degeneracy assumption and the induction hypothesis, in similar fashion to the claim for $\boldsymbol{\alpha}^{t,*}$. And

$$\mathbf{P}_{\mathbf{W}_t}\mathbf{X}^\top \mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t = \mathbf{W}_{t-1}\left(\mathbf{W}_t^\top \mathbf{W}_t\right)^{-1}\mathbf{W}_t^\top \mathbf{X}^\top \mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \tag{6.126}$$

$$= \mathbf{W}_t\left(\mathbf{W}_t^\top \mathbf{W}_t\right)^{-1}\left[\mathbf{r}^0|...|\mathbf{r}^t\right]^\top \mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \tag{6.127}$$

Using a similar argument as in the proof for $\mathbf{r}^t$, we may use the induction hypothesis and non-degeneracy assumption to write the limiting behaviour of the projectors to obtain

$$\frac{1}{n}\left[\mathbf{r}^0|...|\mathbf{r}^{t-1}\right]^\top \mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \overset{\mathrm{P}}{\simeq} \frac{1}{d}\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^t\right]^\top \mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \tag{6.128}$$

$$= \frac{1}{n}\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^t\right]^\top \mathbf{m}^t - \frac{1}{d}\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^t\right]^\top \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{m}^t \tag{6.129}$$

$$\overset{\mathrm{P}}{\simeq} \frac{1}{n}\mathbb{E}\left[\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^t\right]^\top \mathbf{m}^t\right] - \frac{1}{n}\mathbb{E}\left[\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^t\right]^\top \mathbf{M}_{t-1}\right]\boldsymbol{\beta}^{t,*} \tag{6.130}$$

where, for any $0 \leqslant s \leqslant t$, Stein's lemma gives

$$\frac{1}{n}\mathbb{E}\left[\left(\boldsymbol{\omega}^s\right)^\top \mathbf{m}^t\right] = \frac{1}{n}\mathbb{E}\left[\left(\boldsymbol{\omega}^s\right)^\top \mathbf{g}^t\left(\boldsymbol{\eta}^t\left(\boldsymbol{\omega}^0,...,\boldsymbol{\omega}^{t-1},\boldsymbol{\omega}^t\right)\right)\right] = \frac{1}{n}\sum_{i=0}^{t} C_\theta(s,i)\sum_{j=1}^{n}\mathbb{E}\left[\frac{\partial g_j^t}{\partial \omega_j^i}(\eta^t)\right] \tag{6.131}$$

From the definition of $\boldsymbol{\eta}^t$ in Theorem 9, the dependence on $\boldsymbol{\omega}^t$ in $\boldsymbol{\eta}^t$ is the identity. We may then write

$$\frac{1}{n}\mathbb{E}\left[\frac{\partial g_j^t}{\partial \omega_j^t}(\eta^t)\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\frac{dg_j^t}{d\eta_j^t}(\eta^t)\right] = \Gamma^t \tag{6.132}$$

We now define $\mathbf{C}_{\theta,t}$ the $(t+1)q\times(t+1)q$ covariance matrix of the lines of $\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^{t-1}|\boldsymbol{\omega}^t\right] \in \mathbb{R}^{n\times(t+1)q}$, and

$$\mathbf{R}_{g,t} = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\frac{\partial g_j^t}{\partial \boldsymbol{\omega}_j^0}(\eta^t)\right] \\ ... \\ \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\frac{\partial g_j^t}{\partial \boldsymbol{\omega}_j^{t-1}}(\boldsymbol{\eta}^t)\right] \\ \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\frac{dg_j^t}{d\boldsymbol{\eta}_j^t}(\boldsymbol{\eta}^t)\right] \end{bmatrix} \in \mathbb{R}^{(t+1)q\times q} \tag{6.133}$$

and write

$$\frac{1}{n}\mathbb{E}\left[\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^{t-1}\right]^\top \mathbf{m}^t\right] = \mathbf{C}_{\theta,t}\mathbf{R}_{g,t} \tag{6.134}$$

and, for any $0 \leqslant s < t$

$$\frac{1}{n}\mathbb{E}\left[\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^{t-1}\right]^\top \mathbf{m}^s\right] = \mathbf{C}_{\theta,t}\begin{bmatrix} \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\frac{\partial g_j^s}{\partial \omega_j^0}(\eta^s)\right] \\ ... \\ \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\frac{\partial g_j^s}{\partial \omega_j^{s-1}}(\eta^s)\right] \\ \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\frac{dg_j^s}{d\eta_j^s}(\eta^s)\right] \\ 0 \\ ... \\ 0 \end{bmatrix} = \mathbf{C}_{\theta,t}\mathbf{R}_{g,s} \tag{6.135}$$

where the zeroes come from the fact that η^s is not an algebraic function of the ω^l for $l > s$ which is, again, coherent with notions of causality. We thus reach the following equality

$$\frac{1}{n}\mathbb{E}\left[\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^t\right]^\top \mathbf{m}^t\right] - \frac{1}{n}\mathbb{E}\left[\left[\boldsymbol{\omega}^0|...|\boldsymbol{\omega}^t\right]^\top \mathbf{M}_{t-1}\right]\boldsymbol{\beta}^{t,*} \tag{6.136}$$

$$= \mathbf{C}_{\theta,t}\left(\mathbf{R}_{g,t} - \left[\mathbf{R}_{g,0}|\mathbf{R}_{g,1}|...|\mathbf{R}_{g,t-1}\right]\boldsymbol{\beta}^{t,*}\right) \tag{6.137}$$

Also, due to the induction hypothesis

$$\frac{1}{n}\mathbf{W}_t^T\mathbf{W}_t \overset{\mathrm{P}}{\simeq} \mathbf{C}_{\theta,t} \tag{6.138}$$

which leads to

$$\mathbf{P}_{\mathbf{W}_t}\mathbf{X}^\top\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \xrightarrow[n,d\to\infty]{Plk} \mathbf{W}_t\left(\mathbf{R}_{g,t} - \left[\mathbf{R}_{g,0}|\mathbf{R}_{g,1}|...|\mathbf{R}_{g,t-1}\right]\boldsymbol{\beta}^{t,*}\right) \tag{6.139}$$

Combining these results leads to

$$\mathbf{v}^{t+1}|_{\mathfrak{S}^t,\mathbf{r}^t} \xrightarrow[n,d\to\infty]{Plk} h^t(\mathbf{w}^t) + \sum_{k=0}^{t-1}\left(\mathbf{v}^{k+1} - \mathbf{h}^k(\left\{\mathbf{v}^l\right\}_{l=0}^k)\right)\beta_k^{*,t} \tag{6.140}$$

$$+ \mathbf{W}_t\left(\mathbf{R}_{g,t} - \left[\mathbf{R}_{g,0}|\mathbf{R}_{g,1}|...|\mathbf{R}_{g,t-1}\right]\boldsymbol{\beta}^{t,*}\right) + \tilde{\mathbf{X}}^\top\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \tag{6.141}$$

where we used Lemma 21 to remove the projector $\mathbf{P}^\perp_{\mathbf{W}_{t-1}}$ in the term $\mathbf{P}^\perp_{\mathbf{W}_{t-1}}\tilde{\mathbf{X}}^\top\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t$. We now use the induction hypothesis to write

$$\sum_{k=0}^{t-1}\left(\mathbf{v}^{k+1} - \mathbf{h}^k(\mathbf{w}^k)\right)\beta_k^{*,t} \xrightarrow[n,d\to\infty]{Plk} \sum_{k=0}^{t-1}\left(\boldsymbol{\theta}^k\Gamma^k + \sum_{l=0}^{k-1}\theta^l R_g(k,l) + \mathbf{u}^k\right)\beta_k^{*,t} \tag{6.142}$$

and to write

$$\mathbf{W}_t\left[\mathbf{R}_{g,0}|\mathbf{R}_{g,1}|...|\mathbf{R}_{g,t-1}\right]\boldsymbol{\beta}^{t,*} \xrightarrow[n,d\to\infty]{Plk} \sum_{k=0}^{t-1}\left(\boldsymbol{\theta}^k\Gamma^k + \sum_{l=0}^{k-1}\boldsymbol{\theta}^l R_g(k,l)\right)\beta_k^{*,t} \tag{6.143}$$

where we remind that, for any $s < t$, the elements of the last $q \times q$ block of $\mathbf{R}_{g,s}$ are all zeroes, and thus $\mathbf{w}^t$ does not appear in this sum. We reach

$$\mathbf{v}^{t+1}|_{\mathfrak{S}^t,\mathbf{r}^t} \xrightarrow[n,d\to\infty]{Plk} \mathbf{h}^t(\boldsymbol{\omega}^t) + \boldsymbol{\theta}^t\boldsymbol{\Gamma}^t + \sum_{k=0}^{t-1}\boldsymbol{\theta}^k R_g(t,k) + \sum_{k=0}^{t-1}\mathbf{u}^k\beta_k^{*,t} + \tilde{\mathbf{X}}^\top\mathbf{P}^\perp_{\mathbf{M}_{t-1}}\mathbf{m}^t \tag{6.144}$$

Owing to lemma 21

$$\tilde{\mathbf{X}}^\top \mathbf{P}^\perp_{\mathbf{M}_{t-1}} \mathbf{m}^t \xrightarrow[n,d\to\infty]{Plk} \tilde{\mathbf{u}}^t \tag{6.145}$$

where $\tilde{\mathbf{u}}^t \in \mathbb{R}^{d\times q}$ has i.i.d. Gaussian lines with covariance

$$C^\perp_{\mathbf{m},t} = \lim_{n,d\to\infty} \frac{1}{n} \left(\mathbf{P}^\perp_{\mathbf{M}_{t-1}} \mathbf{m}^t\right)^\top \mathbf{P}^\perp_{\mathbf{M}_{t-1}} \mathbf{m}^t \tag{6.146}$$

and is independent from all other random parameters of the problem. We recover a additive Gaussian process term

$$\mathbf{u}^t = \sum_{k=0}^{t-1} \mathbf{u}^k \beta_k^{*,t} + \tilde{\mathbf{u}}^t \tag{6.147}$$

To check it has the correct covariance profile, we evaluate, for any $s < t$

$$\begin{aligned}
\frac{1}{d}\mathbb{E}\left[(\mathbf{u}^s)^\top \mathbf{u}^t\right] &= \sum_{k=0}^{t-1} \mathbb{E}\left[(\mathbf{u}^s)^\top \mathbf{u}^k\right] \beta_k^{*,t} && (6.148)\\
&= \mathbf{C}_{g,t} \boldsymbol{\beta}^{*,t} && (6.149)\\
&\overset{\mathrm{P}}{\simeq} \frac{1}{d}(\mathbf{m}^s)^\top \mathbf{M}_{t-1} \left(\mathbf{M}_{t-1}^\top \mathbf{M}_{t-1}\right)^{-1} \mathbf{M}_{t-1}^\top \mathbf{m}^t && (6.150)\\
&= \frac{1}{d}(\mathbf{m}^s)^\top \mathbf{m}^t && (6.151)\\
&\overset{\mathrm{P}}{\simeq} \frac{1}{d}\mathbb{E}\left[\mathbf{g}^s(\boldsymbol{\eta}^s)^\top \mathbf{g}^t(\boldsymbol{\eta}^t)\right] && (6.152)
\end{aligned}$$

and for $s = t$

$$\begin{aligned}
\frac{1}{d}\mathbb{E}\left[(\mathbf{u}^t)^\top \mathbf{u}^t\right] &= \sum_{k=0}^{t-1}\sum_{k'=0}^{t-1} (\beta_k^{*,t})^\top \frac{1}{d}\mathbb{E}\left[(\mathbf{u}_k)^\top \mathbf{u}_{k'}\right] \beta_{k'}^{*,t} + \frac{1}{d}\mathbb{E}\left[(\tilde{\mathbf{u}}^t)^\top \tilde{\mathbf{u}}^t\right] && (6.153)\\
&\overset{\mathrm{P}}{\simeq} \frac{1}{d}(\mathbf{m}^t)^\top \mathbf{M}_{t-1} \left(\mathbf{M}_{t-1}^\top \mathbf{M}_{t-1}\right)^{-1} \mathbf{M}_{t-1}^\top \mathbf{m}^t + \frac{1}{d}\mathbf{m}^t \mathbf{P}^\perp_{\mathbf{M}_{t-1}} \mathbf{m}^t && (6.154)\\
&= \frac{1}{d}(\mathbf{m}^t)^\top \mathbf{m}^t && (6.155)\\
&\overset{\mathrm{P}}{\simeq} \frac{1}{d}\mathbb{E}\left[\mathbf{g}^t(\boldsymbol{\eta}^t)^\top \mathbf{g}^t(\boldsymbol{\eta}^t)\right] && (6.156)
\end{aligned}$$

which concludes the induction.

6.6.1 Relaxing the non-degeneracy assumption

The non-degeneracy assumption is relaxed using the same method as in [37, 110]. We can define an auxiliary, randomly perturbed iteration with

$$\hat{\mathbf{v}}^{t+1} = \hat{\mathbf{h}}^t \left(\left\{\hat{\mathbf{v}}^k\right\}_{k=0}^t\right) + \mathbf{X}^\top \hat{\mathbf{g}}^t(\hat{\mathbf{r}}^t) \tag{6.157}$$

$$\hat{\mathbf{r}}^t = \mathbf{X} \sum_{k=0}^t \hat{\mathbf{v}}^k \tag{6.158}$$

initialized with the same $\mathbf{v}_0$ as the original dynamics Eq.(6.4)-(6.5), and where the update functions are defined as

$$\hat{\mathbf{h}}^t\left(\left\{\hat{\mathbf{v}}^k\right\}_{k=0}^t\right) = \mathbf{h}^t\left(\left\{\hat{\mathbf{v}}^k\right\}_{k=0}^t\right) + \epsilon \mathbf{Y}_h^t \tag{6.159}$$

$$\hat{\mathbf{g}}^t(\hat{\mathbf{r}}^t) = \mathbf{g}^t(\hat{\mathbf{r}}^t) + \epsilon \mathbf{Y}_r^t \tag{6.160}$$

where, at each time step, $\mathbf{Y}_h^t \in \mathbb{R}^{d\times q}$ and $\mathbf{Y}_r^t \in \mathbb{R}^{n\times q}$ have i.i.d. standard normal elements and are independent from one another and from all other parameters from the problems. Since n, d are much larger than tq by assumption, standard results on Gaussian matrices [288] show that the Gram matrices being inverted in the projectors are almost surely full rank with smallest eigenvalue bounded away from 0 when n, d go to infinity. We thus have the rigorous system of equations for the perturbed iteration. Using inductions, one can then show that the iterates of the perturbed iterations uniformly converge to the original ones when taking ϵ to zero. Similarly, uniform convergence of the asymptotic Gaussian model of the perturbed iteration towards the one of the original iteration can be shown. Taking the limits on both sides concludes the proof. Since the procedure and technical steps are almost identical to those presented in [37, 110], we do not reproduce them here.

6.7 Detailed mapping for Nesterov acceleration

Recall the equations for Nesterov accelerated gradient

$$\mathbf{y}^t = \mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t) \tag{6.161}$$

$$\mathbf{w}^{t+1} = \mathbf{y}^t - \gamma^t\left(\mathbf{X}^\top \nabla\mathcal{L}(\mathbf{X}\mathbf{y}^t) + \nabla\mathbf{F}(\mathbf{y}^t)\right) \tag{6.162}$$

$$\mathbf{z}^{t+1} = \mathbf{z}^t + \mu^t\left(\mathbf{y}^t - \mathbf{z}^t\right) - \alpha^t\left(\mathbf{X}^\top \nabla\mathcal{L}(\mathbf{X}\mathbf{y}^t) + \nabla\mathbf{F}(\mathbf{y}^t)\right) \tag{6.163}$$

Replacing $\mathbf{y}^t$ using its definition leads to

$$\begin{aligned}
\mathbf{w}^{t+1} &= \mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t) - \gamma^t\left(\mathbf{X}^\top \nabla\mathcal{L}(\mathbf{X}\left(\mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t)\right)) + \nabla\mathbf{F}(\mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t))\right)\\
\mathbf{z}^{t+1} &= \mathbf{z}^t + \mu^t\left(\mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t) - \mathbf{z}^t\right)\\
&\qquad - \alpha^t\left(\mathbf{X}^\top \nabla\mathcal{L}\left(\mathbf{X}\left(\mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t)\right)\right) + \nabla\mathbf{F}\left(\mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t)\right)\right)
\end{aligned}$$

Define the variables $\mathbf{u}^{t+1} = \mathbf{w}^{t+1} - \mathbf{w}^t \in \mathbb{R}^d, \tilde{\mathbf{u}}^{t+1} = \mathbf{z}^{t+1} - \mathbf{z}^t \in \mathbb{R}^d, \mathbf{v}^t = [\mathbf{u}^t|\tilde{\mathbf{u}}^t] \in \mathbb{R}^{d\times 2}, \mathbf{x}^t = [\mathbf{w}^t|\mathbf{z}^t] = \sum_{k=0}^t \mathbf{v}^k \in \mathbb{R}^{d\times 2}$. Using these variables, we may write

$$\tau^t(\mathbf{z}^t - \mathbf{w}^t) = \sum_{k=0}^t \mathbf{v}^k \begin{bmatrix} -\tau^t \\ \tau^t \end{bmatrix}$$

$$\mathbf{X}\left(\mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t)\right) = \left(\mathbf{X}\sum_{k=0}^t \mathbf{v}^k\right)\begin{bmatrix} 1-\tau^t \\ \tau^t \end{bmatrix}$$

$$\mu^t(\mathbf{w}^t + \tau^t(\mathbf{z}^t - \mathbf{w}^t) - \mathbf{z}^t) = \sum_{k=0}^t \mathbf{v}^k \begin{bmatrix} \mu^t(1-\tau^t) \\ \mu^t(\tau^t - 1) \end{bmatrix}$$

Defining $\mathbf{r}^t = \mathbf{X}\sum_{k=0}^{t}\mathbf{v}^k$, we obtain

$$\mathbf{v}^{t+1} = \left[\sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}-\tau^t\\ \tau^t\end{bmatrix} \mid \sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}\mu^t(1-\tau^t)\\ \mu^t(\tau^t-1)\end{bmatrix}\right] \tag{6.164}$$

$$+\left[-\gamma^t\nabla F\left(\sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right) \mid -\alpha^t\nabla F\left(\sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right)\right] \tag{6.165}$$

$$+\mathbf{X}^\top\left[-\gamma^t\nabla\mathcal{L}\left(\mathbf{r}^t\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right) \mid -\alpha^t\nabla\mathcal{L}\left(\mathbf{r}^t\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right)\right] \tag{6.166}$$

$$\mathbf{r}^t = \mathbf{X}\sum_{k=0}^{t}\mathbf{v}^k \tag{6.167}$$

which fits the form of Eq. (6.4-6.5) by defining

$$\mathbf{h}^t : \mathbb{R}^{d\times 2(t+1)} \to \mathbb{R}^{d\times 2} \tag{6.168}$$

$$\left\{\mathbf{v}^k\right\}_{k=0}^{t} \to \left[\sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}-\tau^t\\ \tau^t\end{bmatrix} \mid \sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}\mu^t(1-\tau^t)\\ \mu^t(\tau^t-1)\end{bmatrix}\right] \tag{6.169}$$

$$+\left[-\gamma^t\nabla F\left(\sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right) \mid -\alpha^t\nabla F\left(\sum_{k=0}^{t}\mathbf{v}^k\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right)\right] \tag{6.170}$$

$$\mathbf{g}^t : \mathbb{R}^{n\times 2} \to \mathbb{R}^{n\times 2} \tag{6.171}$$

$$\mathbf{r}^t \to \left[-\gamma^t\nabla\mathcal{L}\left(\mathbf{r}^t\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right) \mid -\alpha^t\nabla\mathcal{L}\left(\mathbf{r}^t\begin{bmatrix}1-\tau^t\\ \tau^t\end{bmatrix}\right)\right] \tag{6.172}$$

Part II

Exact asymptotics for convex models : feature maps, ensembling and multiclass problems

Chapter 7

Learning curves of generic features maps for realistic datasets with a Gaussian covariate model

The results in this chapter are based on the paper [176].

Teacher-student models provide a framework in which the typical-case performance of high-dimensional supervised learning can be described in closed form. The assumptions of Gaussian i.i.d. input data underlying the canonical teacher-student model may, however, be perceived as too restrictive to capture the behaviour of realistic data sets. In this paper, we introduce a Gaussian covariate generalisation of the model where the teacher and student can act on different spaces, generated with fixed, but generic feature maps. While still solvable in a closed form, this generalization is able to capture the learning curves for a broad range of realistic data sets, thus redeeming the potential of the teacher-student framework. Our contribution is then two-fold: First, we prove a rigorous formula for the asymptotic training loss and generalisation error. Second, we present a number of situations where the learning curve of the model captures the one of a *realistic data set* learned with kernel regression and classification, with out-of-the-box feature maps such as random projections or scattering transforms, or with pre-learned ones - such as the features learned by training multi-layer neural networks. We discuss both the power and the limitations of the framework.

7.1 Introduction

Teacher-student models are a popular framework to study the high-dimensional asymptotic performance of learning problems with synthetic data, and have been the subject of intense investigations spanning three decades [263, 294, 94, 83, 90, 300, 82]. In the wake of understanding the limitations of classical statistical learning approaches [301, 30, 33], this direction is witnessing a renewal of interest [190, 125, 33, 53, 12, 253]. However, this framework is often assuming the input data to be Gaussian i.i.d., which is arguably too simplistic to be able to capture properties of realistic data. In this paper, we redeem this line of work by defining a Gaussian covariate model where the teacher and student act on different Gaussian correlated spaces with arbitrary covariance. We derive a rigorous asymptotic solution of this model generalizing the formulas found in the above mentioned classical works.

We then put forward a theory, supported by universality arguments and numerical experiments, that this model captures learning curves, i.e. the dependence of the training and test errors on the number of samples, for a generic class of feature maps applied to realistic datasets. These maps can be deterministic, random, or even learnt from the data. This analysis thus gives a unified framework to describe the learning curves of, for example, kernel regression and classification, the analysis of feature maps – random projections [239], neural tangent kernels [130], scattering transforms [9] – as well as the analysis of transfer learning performance on data generated by generative adversarial networks [120]. We also discuss limits of applicability of our results, by showing concrete situations where the learning curves of the Gaussian covariate model differ from the actual ones.

Model definition — The Gaussian covariate teacher-student model is defined via two vectors $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^d$, with correlation matrices $\Psi \in \mathbb{R}^{p\times p}, \Omega \in \mathbb{R}^{d\times d}$ and $\Phi \in \mathbb{R}^{p\times d}$, from which we draw n independent samples:

$$\begin{bmatrix} \boldsymbol{u}^\mu \\ \boldsymbol{v}^\mu \end{bmatrix} \in \mathbb{R}^{p+d} \underset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right), \qquad \mu = 1, \cdots, n. \tag{7.1}$$

The *labels* y^μ are generated by a **teacher** function that is only using the vectors $\boldsymbol{u}^\mu$:

$$y^\mu = f_0\left(\frac{1}{\sqrt{p}}\boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu\right), \tag{7.2}$$

where $f_0 : \mathbb{R} \to \mathbb{R}$ is a function that may include randomness such as, for instance, an additive Gaussian noise, and $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is a vector of teacher-weights with finite norm which can be either random or deterministic. Learning is performed by the **student** with weights $\boldsymbol{w}$ via empirical risk minimization that has access only to the features $\boldsymbol{v}^\mu$:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\arg\min}\left[\sum_{\mu=1}^{n} g\left(\frac{\boldsymbol{w}^\top \boldsymbol{v}^\mu}{\sqrt{d}}, y^\mu\right) + r(\boldsymbol{w})\right], \tag{7.3}$$

where r and g are proper, convex, lower-semicontinuous functions of $\boldsymbol{w} \in \mathbb{R}^d$ (e.g. g can be a logistic or a square loss and r a ℓ_p ($p=1,2$) regularization). The key quantities we want to compute in this model are the *averaged training and generalisation errors* for the estimator $\hat{\boldsymbol{w}}$,

$$\mathcal{E}_{\text{train.}}(\boldsymbol{w}) \equiv \frac{1}{n}\sum_{\mu=1}^{n} g\left(\frac{\boldsymbol{w}^\top \boldsymbol{v}^\mu}{\sqrt{d}}, y^\mu\right) \quad \text{and} \quad \mathcal{E}_{\text{gen.}}(\boldsymbol{w}) \equiv \mathbb{E}\left[\hat{g}\left(\hat{f}\left(\frac{\boldsymbol{v}_{\text{new}}^\top \boldsymbol{w}}{\sqrt{d}}\right), f_0\left(\frac{\boldsymbol{u}_{\text{new}}^\top \boldsymbol{\theta}_0}{\sqrt{p}}\right)\right)\right]. \tag{7.4}$$

where g is the loss function in eq. (7.3), $\hat{f}$ is a prediction function (e.g. $\hat{f} = \text{sign}$ for a classification task), $\hat{g}$ is a performance measure (e.g. $\hat{g}(\hat{y}, y) = (\hat{y} - y)^2$ for regression or $\hat{g}(\hat{y}, y) = \mathbb{P}(\hat{y} \neq y)$ for classification) and $(\boldsymbol{u}_{\text{new}}, \boldsymbol{v}_{\text{new}})$ is a fresh sample from the joint distribution of $\boldsymbol{u}$ and $\boldsymbol{v}$.

Our two **main technical contributions** are:

(C1) In Theorems 11 & 12, we give a rigorous closed-form characterisation of the properties of the estimator $\hat{\boldsymbol{w}}$ for the Gaussian covariate model (7.1), and the corresponding training and generalisation errors in the high-dimensional limit. We prove our result using Gaussian comparison inequalities [121];

(C2) We show how the same expression can be obtained using the replica method from statistical physics [196]. This is of additional interest given the wide range of applications of the replica approach in machine learning and computer science [195]. In particular, this allows to put on a rigorous basis many results previously derived with the replica method.

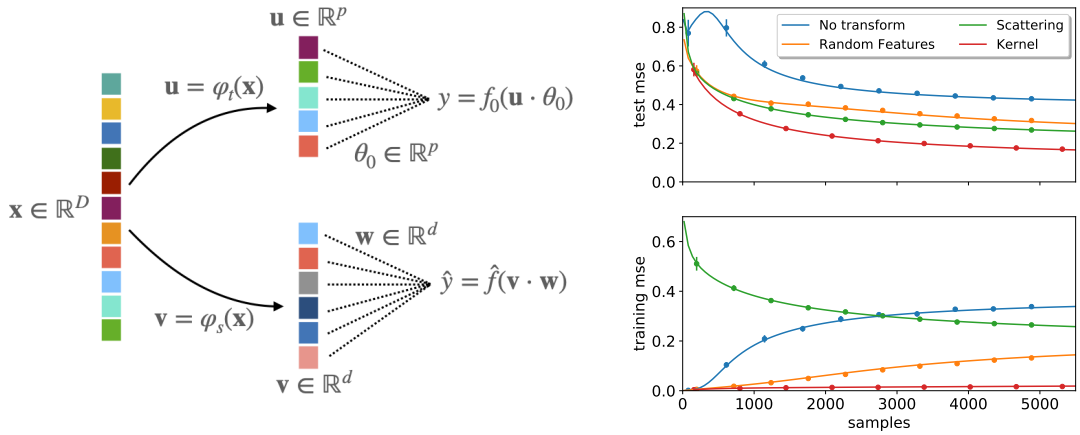


Figure 7.1: **Left:** Given a data set $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$, teacher $\boldsymbol{u} = \boldsymbol{\varphi}_t(\boldsymbol{x})$ and student maps $\boldsymbol{v} = \boldsymbol{\varphi}_t(\boldsymbol{x})$, we assume $[\boldsymbol{u}, \boldsymbol{v}]$ to be jointly Gaussian random variables and apply the results of the Gaussian covariate model (7.1). **Right:** Illustration on real data, here ridge regression on even vs odd MNIST digits, with regularisation $\lambda = 10^{-2}$. Full line is theory, points are simulations. We show the performance with no feature map (blue), random feature map with $\sigma = \text{erf}$ & Gaussian projection (orange), the scattering transform with parameters $J = 3, L = 8$ [9] (green), and of the limiting kernel of the random map [296] (red). The covariance Ω is empirically estimated from the full data set, while the other quantities appearing in the Theorem 11 are expressed directly as a function of the labels, see Section 7.2.4. Simulations are averaged over 10 independent runs.

Towards realistic data — In the second part of our paper, we argue that the above Gaussian covariate model (7.1) is generic enough to capture the learning behaviour of a broad range of realistic data. Let $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$ denote a data set with n independent samples on $\mathcal{X} \subset \mathbb{R}^D$. Based on this input, the **features** $\boldsymbol{u}, \boldsymbol{v}$ are given by (potentially) elaborated transformations of $\boldsymbol{x}$, i.e.

$$\boldsymbol{u} = \boldsymbol{\varphi}_t(\boldsymbol{x}) \in \mathbb{R}^p \quad \text{and} \quad \boldsymbol{v} = \boldsymbol{\varphi}_s(\boldsymbol{x}) \in \mathbb{R}^d \tag{7.5}$$

for given centred feature maps $\boldsymbol{\varphi}_t : \mathcal{X} \to \mathbb{R}^p$ and $\boldsymbol{\varphi}_s : \mathcal{X} \to \mathbb{R}^d$, see Fig. 7.1. Uncentered features can be taken into account by shifting the covariances, but we focus on the centred case to lighten notation.

The Gaussian covariate model (7.1) is exact in the case where $\boldsymbol{x}$ are Gaussian variables and the feature maps $(\boldsymbol{\varphi}_t, \boldsymbol{\varphi}_s)$ preserve the Gaussianity, for example linear features. In particular, this is the case for $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{x}$, which is the widely-studied vanilla teacher-student model [103]. The interest of the model (7.1) is that it also captures a range of cases in which the feature maps $\boldsymbol{\varphi}_t$ and $\boldsymbol{\varphi}_s$ are deterministic, or even learnt from the data. The covariance matrices Ψ, Φ, and Ω then represent different aspects of the data-generative process and learning model. The student (7.3) then corresponds to the last layer of the learning model. These observation can be distilled into the following conjecture:

Conjecture 1. (Gaussian equivalent model) For a wide class of data distributions $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$, and features maps $\boldsymbol{u} = \boldsymbol{\varphi}_t(\boldsymbol{x}), \boldsymbol{v} = \boldsymbol{\varphi}_s(\boldsymbol{x})$, the generalisation and training errors of estimator (7.3) are asymptotically captured by the equivalent Gaussian model (7.1), where $[\boldsymbol{u}, \boldsymbol{v}]$ are jointly Gaussian variables, and thus by the closed-form expressions of Theorem 11.

The second part of our **main contributions** are:
(C3) In Sec. 7.2.3 we show that the theoretical predictions from (C1) captures the learning curves in non-trivial cases, e.g. when input data are generated using a trained generative adversarial network, while extracting both the feature maps from a neural network trained on real data.

(C4) In Sec. 7.2.4, we show empirically that for ridge regression the asymptotic formula of Theorem 11 can be applied *directly* to real data sets, even though the Gaussian hypothesis is not satisfied. This universality-like property is a consequence of Theorem 13 and is illustrated in Fig. 7.1 (right) where the real learning curve of several features maps learning the odd-versus-even digit task on MNIST is compared to the theoretical prediction.

Related work — Rigorous results for teacher-student models: The Gaussian covariate model (7.1) contains the vanilla teacher-student model as a special case where one takes $\boldsymbol{u}$ and $\boldsymbol{v}$ *identical*, with unique covariance matrix Ω. This special case has been extensively studied in the statistical physics community using the heuristic replica method [103, 217, 263, 294, 94]. Many recent rigorous results for such models can be rederived as a special case of our formula, e.g. refs. [190, 125, 113, 33, 53, 281, 208, 12, 253, 57]. Numerous of these results are based on the same proof technique as we employed here: the Gordon's Gaussian min-max inequalities [121, 273, 220]. The asymptotic analysis of kernel ridge regression [45], of margin-based classification [129] also follow from our theorem. Other examples include models of the double descent phenomenon [205]. Closer to our work is the recent work of [76] on the random feature model. For ridge regression, there are also precise predictions thanks to random matrix theory [78, 125, 297, 170, 173, 21, 133]. A related set of results was obtained in [108] for orthogonal random matrix models. The main technical novelty of our proof is the handling of a generic loss and regularisation, not only ridge, representing convex empirical risk minimization, for both classification and regression, with the generic correlation structure of the model (7.1).

Gaussian equivalence: A similar Gaussian conjecture has been discussed in a series of recent works, and some authors proved partial results in this direction [125, 190, 208, 106, 117, 116, 76, 128]. Ref. [116] analyses a special case of the Gaussian model (corresponding to $\boldsymbol{\varphi}_t = \mathrm{id}$ here), and proves a Gaussian equivalence theorem (GET) for feature maps $\boldsymbol{\varphi}_s$ given by single-layer neural networks with fixed weights. They also show that for Gaussian data $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_D)$, feature maps of the form $\boldsymbol{v} = \sigma(\mathrm{W}\boldsymbol{x})$ (with some technical restriction on the weights) led to the jointly-Gaussian property for the two scalars $(\boldsymbol{v}\cdot\boldsymbol{w}, \boldsymbol{u}\cdot\boldsymbol{\theta}_0)$ for *almost* any vector $\boldsymbol{w}$. However, their stringent assumptions on random teacher weights limited the scope of applications to unrealistic label models. A related line of work discussed similar universality through the lens of random matrix theory [92, 230, 174]. In particular, Seddik et al. [261] showed that, in our notations, vectors $[\boldsymbol{u}, \boldsymbol{v}]$ obtained from Gaussian inputs $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_D)$ with Lipschitz feature maps satisfy a concentration property. In this case, again, one can expect the two scalars $(\boldsymbol{v}\cdot\boldsymbol{w}, \boldsymbol{u}\cdot\boldsymbol{\theta}_0)$ to be jointly Gaussian with high-probability on $\boldsymbol{w}$. Remarkably, in the case of random feature maps, [128] could go beyond this central-limit-like behavior and established the universality of the Gaussian covariate model (7.1) for the actual learned weights $\hat{\boldsymbol{w}}$.

7.2 Main technical result

Our main technical result is a closed-form expression for the asymptotic training and generalisation errors (7.4) of the Gaussian covariate model introduced above. We start by presenting our result in the most relevant setting for the applications of interest in Section 11.3, which is the case of the ℓ_2 regularization. Next, we briefly present our result in larger generality, which includes non-asymptotic results for non-separable losses and regularizations.

We start by defining key quantities that we will use to characterize the estimator $\hat{\boldsymbol{w}}$. Let

$\Omega = \mathrm{S}^\top \mathrm{diag}(\omega_i)\mathrm{S}$ be the spectral decomposition of Ω. Let:

$$\rho \equiv \frac{1}{d}\boldsymbol{\theta}_0^\top \Psi \boldsymbol{\theta}_0 \in \mathbb{R}, \qquad \bar{\boldsymbol{\theta}} \equiv \frac{\mathrm{S}\Phi^\top \boldsymbol{\theta}_0}{\sqrt{\rho}} \in \mathbb{R}^d \tag{7.6}$$

and define the joint empirical density $\hat{\mu}_d$ between $(\omega_i, \bar{\theta}_i)$:

$$\hat{\mu}_d(\omega, \bar{\theta}) \equiv \frac{1}{d}\sum_{i=1}^{d} \delta(\omega - \omega_i)\delta(\bar{\theta} - \bar{\theta}_i). \tag{7.7}$$

Note that $\Phi^\top \boldsymbol{\theta}_0$ is the projection of the teacher weights on the student space, and therefore $\bar{\boldsymbol{\theta}}$ is the rotated projection on the basis of the student covariance, rescaled by the teacher variance. Together with the student eigenvalues ω_i, these are relevant statistics of the model, encoded here in the joint distribution $\hat{\mu}_d$.

Assumptions — Consider the *high-dimensional* limit in which the number of samples n and the dimensions p, d go to infinity with fixed ratios:

$$\alpha \equiv \frac{n}{d}\,, \quad \text{and} \quad \gamma \equiv \frac{p}{d}\,. \tag{7.8}$$

Assume that the covariance matrices Ψ, Ω are positive-definite and that the Schur complement of the block covariance in equation (7.1) is positive semi-definite. Additionally, the spectral distributions of the matrices Φ, Ψ and Ω converge to distributions such that the limiting joint distribution μ is well-defined, and their maximum singular values are bounded with high probability as $n, p, d \to \infty$. Finally, regularity assumptions are made on the loss and regularization functions mainly to ensure feasibility of the minimization problem. We assume that the cost function $\mathbf{F} + g$ is coercive, i.e. $\lim_{\|\boldsymbol{w}\|_2 \to +\infty} (\mathbf{F}+g)(\boldsymbol{w}) = +\infty$ and that the following scaling condition holds : for all $n, d \in \mathbb{N}$, $\boldsymbol{z} \in \mathbb{R}^n$ and any constant $c > 0$, there exist a finite, positive constant C, such that, for any standard normal random vectors $\boldsymbol{h} \in \mathbb{R}^d$ and $\boldsymbol{g} \in \mathbb{R}^n$:

$$\|\boldsymbol{z}\|_2 \leqslant c\sqrt{n} \implies \sup_{\boldsymbol{x} \in \partial g(\boldsymbol{z})} \|\boldsymbol{x}\|_2 \leqslant C\sqrt{n}, \qquad \frac{1}{d}\mathbb{E}\left[\mathbf{F}(\boldsymbol{h})\right] < +\infty, \qquad \frac{1}{n}\mathbb{E}\left[g(\boldsymbol{g})\right] < +\infty \tag{7.9}$$

The relevance of these assumptions in a supervised machine learning context is discussed in Appendix 8.1. We are now in a position to state our result.

Theorem 11. *(Closed-form asymptotics for ℓ_2 regularization) In the asymptotic limit defined above, the training and generalisation errors (7.4) of the estimator $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ solving the empirical risk minimisation problem in eq. (7.3) with ℓ_2 regularization $r(\boldsymbol{w}) = \frac{\lambda}{2}||\boldsymbol{w}||_2^2$ verify:*

$$\begin{aligned}
&\mathcal{E}_{\text{train.}}(\hat{\boldsymbol{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[g\left(prox_{V^\star g(.,f_0(\sqrt{\rho}s))}\left(\frac{m^\star}{\sqrt{\rho}}s + \sqrt{q^\star - \frac{m^{\star 2}}{\rho}}h\right), f_0(\sqrt{\rho}s)\right)\right] \\
&\mathcal{E}_{\text{gen.}}(\hat{\boldsymbol{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{(\nu,\lambda)}\left[\hat{g}\left(\hat{f}(\lambda), f_0(\nu)\right)\right]
\end{aligned} \tag{7.10}$$

where prox stands for the proximal operator defined as

$$prox_{Vg(.,y)}(x) = \arg\min_z \{g(z,y) + \frac{1}{2V}(x-z)^2\} \tag{7.11}$$

and where (ν, λ) *are jointly Gaussian scalar variables:*

$$(\nu, \lambda) \sim \mathcal{N}\left(0, \begin{bmatrix} \rho & m^\star \\ m^\star & q^\star \end{bmatrix}\right), \tag{7.12}$$

and the overlap parameters $(V^\star, q^\star, m^\star)$ *are prescribed by the unique fixed point of the following set of self-consistent equations:*

$$\begin{cases} V = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\omega}{\lambda+\hat{V}\omega}\right] \\ m = \frac{\hat{m}}{\sqrt{\gamma}}\mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\bar{\theta}^2}{\lambda+\hat{V}\omega}\right] \\ q = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\hat{m}^2\bar{\theta}^2\omega+\hat{q}\omega^2}{\left(\lambda+\hat{V}\omega\right)^2}\right] \end{cases}, \quad \begin{cases} \hat{V} = \frac{\alpha}{V}(1 - \mathbb{E}_{s,h\sim\mathcal{N}(0,1)}[f_g'(V, m, q)]) \\ \hat{m} = \frac{1}{\sqrt{\rho\gamma}}\frac{\alpha}{V}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[s f_g(V, m, q) - \frac{m}{\sqrt{\rho}} f_g'(V, m, q)\right] \\ \hat{q} = \frac{\alpha}{V^2}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}h - f_g(V, m, q)\right)^2\right] \end{cases} \tag{7.13}$$

where we defined the scalar random functions $f_g(V, m, q) = prox_{Vg(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms + \sqrt{q - \rho^{-1}m^2}h)$ *and* $f_g'(V, m, h) = prox'_{Vg(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms + \sqrt{q - \rho^{-1}m^2}h)$ *as the first derivative of the proximal operator.*

Proof: This result is a consequence of Theorem 12, whose proof can be found in the chapter 8.

The parameters of the model $(\boldsymbol{\theta}_0, \Omega, \Phi, \Psi)$ only appear trough ρ, eq. (7.6), and the asymptotic limit μ of the joint distribution eq. (7.7) and $(f_0, \hat{f}, g, \lambda)$. One can easily iterate the above equations to find their fixed point, and extract (q^*, m^*) which appear in the expressions for the training and generalisation errors $(\mathcal{E}^\star_{\text{train}}, \mathcal{E}^\star_{\text{gen}})$, see eq. (7.4). Note that $(q^\star, m^\star)$ have an intuitive interpretation in terms of the estimator $\hat{\boldsymbol{w}} \in \mathbb{R}^d$:

$$q^\star \equiv \frac{1}{d}\hat{\boldsymbol{w}}^\top\Omega\hat{\boldsymbol{w}}, \qquad m^\star \equiv \frac{1}{\sqrt{dp}}\boldsymbol{\theta}_0^\top\Phi\hat{\boldsymbol{w}} \tag{7.14}$$

Or in words: $m^\star$ is the correlation between the estimator projected in the teacher space, while $q^\star$ is the reweighted norm of the estimator by the covariance Ω. The parameter V^* also has a concrete interpretation : it parametrizes the deformation that must be applied to a Gaussian field specified by the solution of the fixed point equations to obtain the asymptotic behaviour of $\hat{\mathbf{z}}$. It prescribes the degree of non-linearity given to the linear output by the chosen loss function. This is coherent with the robust regression viewpoint, where one introduces non-square losses to deal with the potential non-linearity of the generative model. $\hat{V}^*$ plays a similar role for the estimator $\hat{\mathbf{w}}$ through the proximal operator of the regularisation, see Theorem 14 and 15 in the Appendix. Two cases are of particular relevance for the experiments that follow. The first is the case of *ridge regression*, in which $f_0(x) = \hat{f}(x)$ and both the loss g and the performance measure $\hat{g}$ are taken to be the *mean-squared error* $\text{mse}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$, and the asymptotic errors are given by the simple closed-form expression:

$$\mathcal{E}^\star_{\text{gen}} = \rho + q^\star - 2m^\star, \qquad \mathcal{E}^\star_{\text{train}} = \frac{\mathcal{E}^\star_{\text{gen}}}{(1 + V^\star)^2}, \tag{7.15}$$

The second case of interest is the one of a binary classification task, for which $f_0(x) = \hat{f}(x) = \text{sign}(x)$, and we choose the performance measure to be the *classification error* $\hat{g}(y, \hat{y}) = \mathbb{P}(y \neq \hat{y})$. In the same notation as before, the asymptotic generalisation error in this case reads:

$$\mathcal{E}^\star_{\text{gen}} = \frac{1}{\pi}\cos^{-1}\left(\frac{m^\star}{\sqrt{\rho q^\star}}\right), \tag{7.16}$$

while the training error $\mathcal{E}^{\star}_{\text{train}}$ depends on the choice of g - which we will take to be the logistic loss $g(y,x) = \log\left(1+e^{-xy}\right)$ in all of the binary classification experiments.

As mentioned above, this paper includes stronger technical results including finite size corrections and precise characterization of the distribution of the estimator $\hat{\boldsymbol{w}}$, for generic, non-separable loss and regularization g and r. This type of distributional statement is encountered for special cases of the model in related works such as [204, 57, 208]. Define $\mathcal{V} \in \mathbb{R}^{n\times d}$ as the matrix of concatenated samples used by the student. Informally, in high-dimension, the estimator $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{z}} = \frac{1}{\sqrt{d}}\mathcal{V}\hat{\boldsymbol{w}}$ roughly behave as non-linear transforms of Gaussian random variables centered around the teacher vector $\boldsymbol{\theta}_0$ (or its projection on the covariance spaces) as follows:

$$\boldsymbol{w}^* = \Omega^{-1/2} \underset{\frac{1}{\hat{V}^*}\mathbf{F}(\Omega^{-1/2}.)}{\text{prox}}\left(\frac{1}{\hat{V}^*}(\hat{m}^*\boldsymbol{t} + \sqrt{\hat{q}^*}\boldsymbol{g})\right), \; \boldsymbol{z}^* = \underset{V^* g(.,\boldsymbol{z})}{\text{prox}}\left(\frac{m^*}{\sqrt{\rho}}\boldsymbol{s} + \sqrt{q^* - \frac{(m^*)^2}{\rho}}\boldsymbol{h}\right).$$

where $\boldsymbol{s}, \boldsymbol{h} \sim \mathcal{N}(0, \mathrm{I}_n)$ and $\boldsymbol{g} \sim \mathcal{N}(0, \mathrm{I}_d)$ are random vectors independent of the other quantities, $\boldsymbol{t} = \Omega^{-1/2}\Phi^\top\boldsymbol{\theta}_0$, $\boldsymbol{y} = \boldsymbol{f}_0\left(\sqrt{\rho}\boldsymbol{s}\right)$, and $(V^*, \hat{V}^*, q^*, \hat{q}^*, m^*, \hat{m}^*)$ is the unique solution to the fixed point equations presented in Lemma 36 of Chapter 8. Those fixed point equations are the generalization of (7.13) to generic, non-separable loss function and regularization. The formal concentration of measure result can then be stated in the following way:

Theorem 12. *(Non-asymptotic version, generic loss and regularization) Under Assumption (8.1), consider any optimal solution $\hat{\boldsymbol{w}}$ to 7.3. Then, there exist constants $C, c, c' > 0$ such that, for any Lipschitz function $\phi_1 : \mathbb{R}^d \to \mathbb{R}$, and separable, pseudo-Lipschitz function $\phi_2 : \mathbb{R}^n \to \mathbb{R}$ and any $0 < \epsilon < c'$:*

$$\mathbb{P}\left(\left|\phi_1\left(\frac{\hat{\boldsymbol{w}}}{\sqrt{d}}\right) - \mathbb{E}\left[\phi_1\left(\frac{\boldsymbol{w}^*}{\sqrt{d}}\right)\right]\right| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon^2}e^{-cn\epsilon^4}, \mathbb{P}\left(\left|\phi_2\left(\frac{\hat{\boldsymbol{z}}}{\sqrt{n}}\right) - \mathbb{E}\left[\phi_2\left(\frac{\boldsymbol{z}^*}{\sqrt{n}}\right)\right]\right| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon^2}e^{-cn\epsilon^4}.$$

Note that in this form, the dimensions n, p, d still appear explicitly, as we are characterizing the convergence of the estimator's distribution for large but finite dimension. The clearer, one-dimensional statements are recovered by taking the $n, p, d \to \infty$ limit with separable functions and an ℓ_2 regularization. Other simplified formulas can also be obtained from our general result in the case of an ℓ_1 penalty, but since this breaks rotational invariance, they do look more involved than the ℓ_2 case. From Theorem 12, one can deduce the expressions of a number of observables, represented by the test functions ϕ_1, ϕ_2, characterizing the performance of $\hat{\boldsymbol{w}}$, for instance the training and generalization error. A more detailed statement, along with the proof, is given in Chapter 8.

We now discuss how the theorems above are applied to characterise the learning curves for a range of concrete cases. We present a number of cases – some rather surprising – for which Conjecture 1 seems valid, and point out some where it is not. An out-of-the-box iterator for all the cases studied hereafter is provided in the GitHub repository for this manuscript at `https://github.com/IdePHICS/GCMProject`.

7.2.1 Random kitchen sink with Gaussian data

If we choose random feature maps $\boldsymbol{\varphi}_s(\boldsymbol{x}) = \sigma\left(\mathrm{F}\boldsymbol{x}\right)$ for a random matrix F and a chosen scalar function σ acting component-wise, we obtain the random kitchen sink model [239]. This model has seen a surge of interest recently, and a sharp asymptotic analysis was provided in the particular case of uncorrelated Gaussian data $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I}_D)$ and $\boldsymbol{\varphi}_t(\boldsymbol{x}) = \boldsymbol{x}$ in [190, 125] for ridge regression

and generalised by [106, 128] for generic convex losses. Both results can be framed as a Gaussian covariate model with:

$$\Psi = \mathrm{I}_p, \qquad \Phi = \kappa_1 \mathrm{F}^\top, \qquad \Omega = \kappa_0^2 \mathbf{1}_d \mathbf{1}_d^\top + \kappa_1^2 \frac{\mathrm{F}\mathrm{F}^\top}{d} + \kappa_\star^2 \mathrm{I}_d, \tag{7.17}$$

where $\mathbf{1}_d \in \mathbb{R}^d$ is the all-one vector and the constants $(\kappa_0, \kappa_1, \kappa_\star)$ are related to the non-linearity σ:

$$\kappa_0 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(z)\right], \qquad \kappa_1 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[z\sigma(z)\right], \qquad \kappa_\star = \sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(z)^2\right] - \kappa_0^2 - \kappa_1^2}\,. \tag{7.18}$$

In this case, the averages over μ in eq. (7.13) can be directly expressed in terms of the Stieltjes transform associated with the spectral density of $\mathrm{F}\mathrm{F}^\top$. Note, however, that our present framework can accommodate more involved random sinks models, such as when the teacher features are also a random feature model or multi-layer random architectures.

7.2.2 Kernel methods with Gaussian data

Another direct application of our formalism is to kernel methods. Kernel methods admit a dual representation in terms of optimization over feature space [257]. The connection is given by Mercer's theorem, which provides an eigen-decomposition of the kernel and of the target function in the feature basis, effectively mapping kernel regression to a teacher-student problem on feature space. The classical way of studying the performance of kernel methods [272, 55] is then to directly analyse the performance of convex learning in this space. In our notation, the teacher and student feature maps are equal, and we thus set $p = d, \Psi = \Phi = \Omega = \mathrm{diag}(\omega_\mathrm{i})$ where ω_i are the eigenvalues of the kernel and we take the teacher weights $\boldsymbol{\theta}_0$ to be the decomposition of the target function in the kernel feature basis.

There are many results in classical learning theory on this problem for the case of ridge regression (where the teacher is usually called "the source" and the eigenvalues of the kernel matrix the "capacity", see e.g. [272, 235]). However, these are worst case approaches, where no assumption is made on the true distribution of the data. In contrast, here we follow a *typical case* analysis, assuming Gaussianity in feature space. Through Theorem 11, this allows us to go beyond the restriction of the ridge loss. An example for logistic loss is in Fig. 7.2.

For the particular case of kernel ridge regression, Th. 11 provides a rigorous proof of the formula conjectured in [45]. Hard-margin Support Vector Machines (SVMs) have also been studied using the heuristic replica method from statistical physics in [77, 218]. In our framework, this corresponds to the *hinge loss* $g(x, y) = \max(0, 1 - yx)$ when $\lambda \to 0^+$. Our theorem thus puts also these works on rigorous grounds, and extends them to more general losses and regularization.

7.2.3 GAN-generated data and learned teachers

To approach more realistic data sets, we now consider the case in which the input data $\boldsymbol{x} \in \mathcal{X}$ is given by a generative neural network $\boldsymbol{x} = \mathcal{G}(\boldsymbol{z})$, where $\boldsymbol{z}$ is a Gaussian i.i.d. latent vector. Therefore, the covariates $[\boldsymbol{u}, \boldsymbol{v}]$ are the result of the following Markov chain:

$$\boldsymbol{z} \underset{\mathcal{G}}{\mapsto} \boldsymbol{x} \in \mathcal{X} \underset{\varphi_t}{\mapsto} \boldsymbol{u} \in \mathbb{R}^p, \qquad \boldsymbol{z} \underset{\mathcal{G}}{\mapsto} \boldsymbol{x} \in \mathcal{X} \underset{\varphi_s}{\mapsto} \boldsymbol{v} \in \mathbb{R}^d. \tag{7.19}$$

With a model for the covariates, the missing ingredient is the teacher weights $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, which determine the label assignment: $y = f_0(\boldsymbol{u}^\top \boldsymbol{\theta}_0)$. In the experiments that follow, we fit the teacher

weights *from the original data set in which the generative model $\mathcal{G}$ was trained*. Different choices for the fitting yield different teacher weights, and the quality of label assignment can be accessed by the performance of the fit on the test set. The set $(\boldsymbol{\varphi}_t, \boldsymbol{\varphi}_s, \mathcal{G}, \boldsymbol{\theta}_0)$ defines the data generative process. For predicting the learning curves from the iterative eqs. (7.13) we need to sample from the spectral measure μ, which amounts to estimating the *population* covariances (Ψ, Φ, Ω). This is done from the generative process in eq. (7.19) with a Monte Carlo sampling algorithm. This pipeline is explained in detail in Appendix of the original paper. An open source implementation of the algorithms used in the experiments is available online at `https://github.com/IdePHICS/GCMProject`.

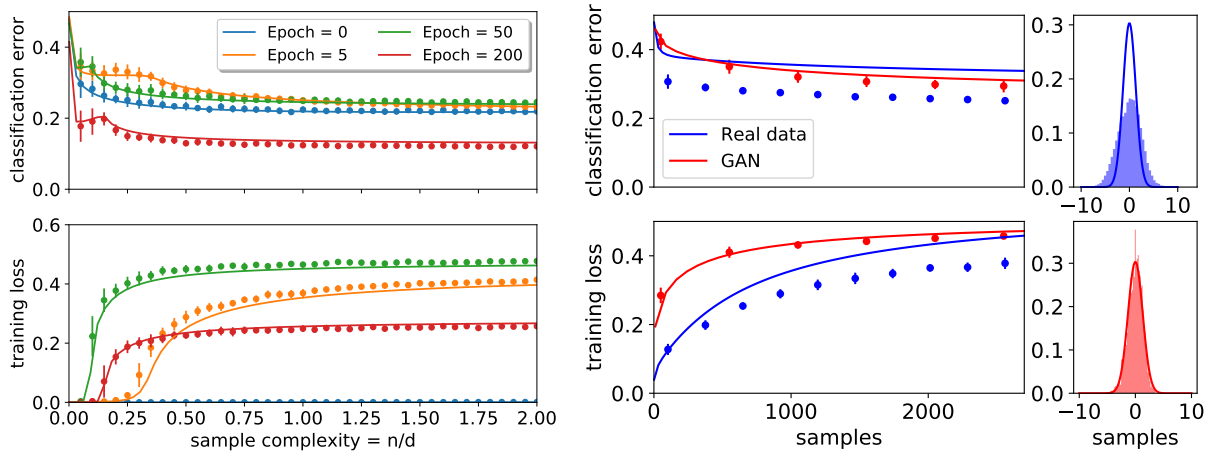


Figure 7.3: **Left:** generalisation classification error (top) and (unregularised) training loss (bottom) vs the sample complexity $\alpha = n/d$ for logistic regression on a learned feature map trained on dcGAN-generated CIFAR10-like images labelled by a teacher fully-connected neural network (see Appendix of the original paper), with vanishing ℓ_2 regularisation. The different curves compare featured maps at different epochs of training. The theoretical predictions based on the Gaussian covariate model (full lines) are in very good agreement with the actual performance (points). **Right:** Test classification error (top) and (unregularised) training loss, (bottom) for logistic regression as a function of the number of samples n for an animal vs not-animal binary classification task with ℓ_2 regularization $\lambda = 10^{-2}$, comparing real CIFAR10 grey-scale images (blue) with dcGAN-generated CIFAR10-like gray-scale images (red). The real-data learning curve was estimated, just as in Figs. 7.4 from the population covariances on the full data set, and it is not in agreement with the theory in this case. On the very right we depict the histograms of the variable $\frac{1}{\sqrt{d}}\boldsymbol{v}^\top \hat{\boldsymbol{w}}$ for a fixed number of samples $n = 2d = 2048$ and the respective theoretical predictions (solid line). Simulations are averaged over 10 independent runs.

Fig. 7.3 shows an example of the learning curves resulting from the pipeline discussed above in a logistic regression task on data generated by a GAN trained on CIFAR10 images. More concretely, we used a pre-trained five-layer deep convolutional GAN (dcGAN) from [237], which maps 100 dimensional i.i.d. Gaussian noise into $k = 32 \times 32 \times 3$ realistic looking CIFAR10-like images: $\mathcal{G} : \boldsymbol{z} \in \mathbb{R}^{100} \mapsto \boldsymbol{x} \in \mathbb{R}^{32\times32\times3}$. To generate labels, we trained a simple fully-connected four-layer neural network on the *real* CIFAR10 data set, on a odd ($y = +1$) vs. even ($y = -1$) task, achieving $\sim 75\%$ classification accuracy on the test set. The teacher weights $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ were taken from the last layer of the network, and the teacher feature map $\boldsymbol{\varphi}_t$ from the three previous layers. For the

student model, we trained a completely independent fully connected 3-layer neural network on the dcGAN-generated CIFAR10-like images and took snapshots of the feature maps $\boldsymbol{\varphi}_s^i$ induced by the 2-first layers during the first $i \in \{0, 5, 50, 200\}$ epochs of training. Finally, once $(\mathcal{G}, \boldsymbol{\varphi}_t, \boldsymbol{\varphi}_s^i, \boldsymbol{\theta}_0)$ have been fixed, we estimated the covariances (Ψ, Φ, Ω) with a Monte Carlo algorithm. Details of the architectures used and of the training procedure can be found in the Appendix of the original paper.

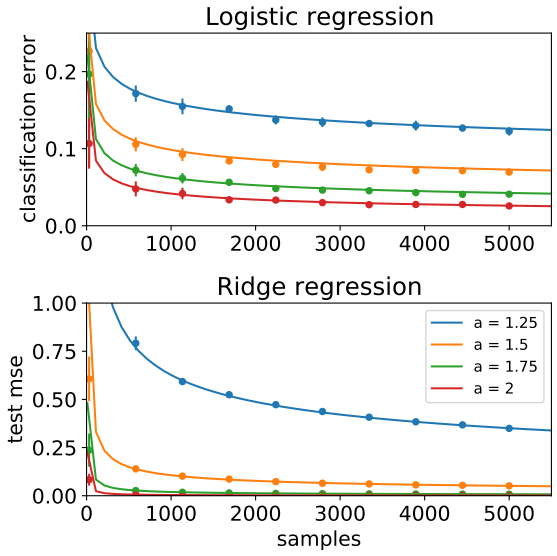


Figure 7.2: Learning in kernel space: Teacher and student live in the same (Hilbert) feature space $\boldsymbol{v} = \boldsymbol{u} \in \mathbb{R}^d$ with $d \gg n$, and the performance only depends on the relative decay between the student spectrum $\omega_i = d\ i^{-2}$ (the capacity) and the teacher weights in feature space $\theta_{0i}^2 \omega_i = d\ i^{-a}$ (the source). Top: a task with sign teacher (in kernel space), fitted with a max-margin support vector machine (logistic regression with vanishing regularisation [246]). Bottom: a task with linear teacher (in kernel space) fitted via kernel ridge regression with vanishing regularisation. Points are simulation that matches the theory (lines). Simulations are averaged over 10 independent runs.

Fig. 7.3 depicts the resulting learning curves obtained by training the last layer of the student. Interestingly, the performance of the feature map at epoch 0 (random initialisation) beats the performance of the learned features during early phases of training in this experiment. Another interesting behaviour is given by the separability threshold of the learned features, i.e. the number of samples for which the training loss becomes larger than 0 in logistic regression. At epoch 50 the learned features are separable at lower sample complexity $\alpha = n/d$ than at epoch 200 - even though in the later the training and generalisation performances are better.

7.2.4 Learning from real data sets

Applying teacher/students to a real data set — Given that the learning curves of realistic-looking inputs can be captured by the Gaussian covariate model, it is fair to ask whether the same might be true for *real data sets*. To test this idea, we first need to cast the real data set into the teacher-student formalism, and then compute the covariance matrices Ω, Ψ, Φ and teacher vector $\boldsymbol{\theta}_0$ required by model (7.1).

Let $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^{n_{\rm tot}}$ denote a real data set, e.g. MNIST or Fashion-MNIST for concreteness, where $n_{\rm tot} = 7 \times 10^4$, $\boldsymbol{x}^\mu \in \mathbb{R}^D$ with $D = 784$. Without loss of generality, we can assume the data is centred. To generate the teacher, let $\boldsymbol{u}^\mu = \boldsymbol{\varphi}_t(\boldsymbol{x}^\mu) \in \mathbb{R}^p$ be a feature map such that data is invertible in feature space, i.e. that $y^\mu = \boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu$ for some teacher weights $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, which should be computed from the samples. Similarly, let $\boldsymbol{v}^\mu = \boldsymbol{\varphi}_s(\boldsymbol{x}^\mu) \in \mathbb{R}^d$ be a feature map we are interested in studying. Then, we can estimate the population covariances (Ψ, Φ, Ω) empirically from the *entire* data set as:

$$\Psi = \sum_{\mu=1}^{n_{\rm tot}} \frac{\boldsymbol{u}^\mu \boldsymbol{u}^{\mu\top}}{n_{\rm tot}}, \qquad \Phi = \sum_{\mu=1}^{n_{\rm tot}} \frac{\boldsymbol{u}^\mu \boldsymbol{v}^{\mu\top}}{n_{\rm tot}}, \qquad \Omega = \sum_{\mu=1}^{n_{\rm tot}} \frac{\boldsymbol{v}^\mu \boldsymbol{v}^{\mu\top}}{n_{\rm tot}}. \tag{7.20}$$

At this point, we have all we need to run the self-consistent equations (7.13). The issue with this approach is that there is not a unique teacher map $\boldsymbol{\varphi}_t$ and teacher vector $\boldsymbol{\theta}_0$ that fit the true labels. However, we can show that *all interpolating linear teachers are equivalent*:

Theorem 13. *(Universality of linear teachers) For any teacher feature map $\boldsymbol{\varphi}_t$, and for any $\boldsymbol{\theta}_0$ that interpolates the data so that $y^\mu = \boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu \, \forall \mu$, the asymptotic predictions of model (7.1) are equivalent.*

Proof. It follows from the fact that the teacher weights and covariances only appear in eq. (7.13) through $\rho = \frac{1}{p}\boldsymbol{\theta}_0^\top \Psi \boldsymbol{\theta}_0$ and the projection $\Phi^\top \boldsymbol{\theta}_0$. Using the estimation (7.20) and the assumption that it exists $y^\mu = \boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu$, one can write these quantities directly from the labels y^μ:

$$\rho = \frac{1}{n_{\rm tot}} \sum_{\mu=1}^{n_{\rm tot}} (y^\mu)^2 \,, \qquad\qquad \Phi^\top \boldsymbol{\theta}_0 = \frac{1}{n_{\rm tot}} \sum_{\mu=1}^{n_{\rm tot}} y^\mu \boldsymbol{v}^\mu \,. \tag{7.21}$$

For linear interpolating teachers, results are thus independent of the choice of the teacher. □

Although this result might seen surprising at first sight, it is quite intuitive. Indeed, the information about the teacher model only enters the Gaussian covariate model (7.1) through the statistics of $\boldsymbol{u}^\top \boldsymbol{\theta}_0$. For a linear teacher $f_0(x) = x$, this is precisely given by the labels.

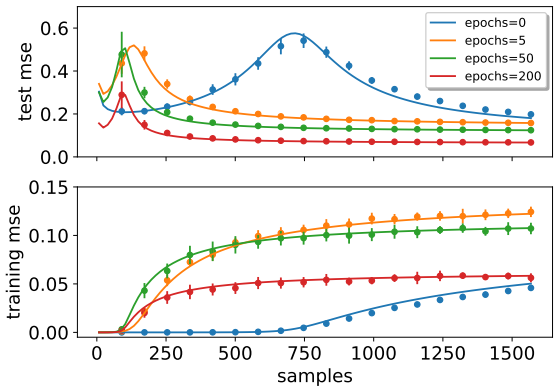


Figure 7.4: Test and training mean-squared errors eqs. (7.15) as a function of the number of samples n for ridge regression. The Fashion-MNIST data set, with vanishing regularisation $\lambda = 10^{-5}$. In this plot, the student feature map $\boldsymbol{\varphi}_s$ is a 3-layer fully-connected neural network with $d = 2352$ hidden neurons trained on the full data set with the square loss. Different curves correspond to the feature map obtained at different stages of training. Simulations are averaged over 10 independent runs. Further details on the simulations are described in the original paper.

Ridge Regression with linear teachers — We now test the prediction of model (7.1) on real data sets, and show that it is surprisingly effective in predicting the learning curves, at least for the ridge regression task. We have trained a 3-layer fully connected neural network with ReLU activations on the full Fashion-MNIST data set to distinguish clothing used above vs. below the waist. The student feature map $\boldsymbol{\varphi}_s : \mathbb{R}^{784} \to \mathbb{R}^d$ is obtained by removing the last layer, see the original paper for a detailed description. In Fig. 7.4 we show the test and training errors of the ridge estimator on a sub-sample of $n < n_{\rm tot}$ on the Fashion-MNIST images. We observe remarkable agreement between the learning curve obtained from simulations and the theoretical prediction by the matching Gaussian covariate model. Note that for the square loss and for $\lambda \ll 1$, the worst performance peak is located at the point in which the linear system becomes invertible. Curiously, Fig. 7.4 shows that the fully-connected network progressively learns a low-rank representation of the data as training proceeds. This can be directly verified by counting the number of zero eigenvalues of Ω, which go from a full-rank matrix to a matrix of rank 380 after 200 epochs of training.

Fig. 7.1 (right) shows a similar experiment on the MNIST data set, but for different out-of-the-box feature maps, such as random features and the scattering transform [52], and we chose the number of random features $d = 1953$ to match the number of features from the scattering transform. Note the characteristic double-descent behaviour [217, 271, 30], and the accurate prediction of the peak where the interpolation transition occurs.

Why is the Gaussian model so effective for describing learning with data that are *not* Gaussian? The point is that ridge regression is sensitive only to second order statistics, and not to the full distribution of the data. It is a classical property (see the appendix of the original paper or the derivation for least-square in the introduction) that the training and generalisation errors are only a function of the spectrum of the *empirical* and *population* covariances, and of their products. Random matrix theory teaches us that such quantities are very robust, and their asymptotic behaviour is universal for a broad class of distributions of $[\boldsymbol{u}, \boldsymbol{v}]$ [17, 159, 91, 174]. The asymptotic behavior of kernel matrices has indeed been the subject of intense scrutiny [92, 63, 230, 190, 96, 261]. Indeed, a universality result akin to Theorem 13 was noted in [133] in the specific case of kernel methods. We thus expect the validity of model (7.1) for ridge regression, with a linear teacher, to go way beyond the Gaussian assumption.

Beyond ridge regression — The same strategy fails beyond ridge regression and mean-squared test error. This suggests a limit in the application of model (7.1) to real (non-Gaussian) data to the universal linear teacher. To illustrate this, consider the setting of Figs. 7.4, and compare the model predictions for the binary classification error instead of the ℓ_2 one. There is a clear mismatch between the simulated performance and prediction given by the theory due to the fact that the classification error does not depends only on the first two moments.

We present an additional experiment in Fig. 7.3. We compare the learning curves of logistic regression on a classification task on the *real* CIFAR10 images with the real labels versus the one on dcGAN-generated CIFAR10-like images and teacher generated labels from Sec. 7.2.3. While the Gaussian theory captures well the behaviour of the later, it fails on the former. A histogram of the distribution of the product $\boldsymbol{u}^\top \hat{\boldsymbol{w}}$ for a fixed number of samples illustrates well the deviation from the prediction of the theory with the real case, in particular on the tails of the distribution. The difference between GAN generated data (that fits the Gaussian theory) and real data is clear. Given that for classification problems there exists a number of choices of "sign" teachers and feature maps that give the exact same labels as in the data set, an interesting open question is: *is there a teacher that allows to reproduce the learning curves more accurately*? This question is left for future works.

Chapter 8

Proofs for the Gaussian covariate model

This section presents the core technical result of this paper in its full generality, along with the required assumptions and its complete proof. For technical reasons, variables different than the ones appearing in the replica calculation are introduced. The proof is nonetheless presented in a self-contained way and the relation with the replica variables are given in section 8.3 , eq.(8.204). We start by reminding the formulation of the problem. Consider the matrices $U \in \mathbb{R}^{n\times p}$ of concatenated vectors $\mathbf{u}$ used by the teacher and $\mathcal{V} \in \mathbb{R}^{n\times d}$ the corresponding one for the student. The estimator may now be defined using potentially non-separable functions:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \left[g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w}, \mathbf{y}\right) + r(\mathbf{w})\right] , \tag{8.1}$$

where the function $g : \mathbb{R}^n \to \mathbb{R}$. The training and generalization errors are reminded as:

$$\mathcal{E}_{\text{train}}(\mathbf{w}) \equiv \frac{1}{n}\mathbb{E}\left[g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w}, \mathbf{y}\right) + \mathbf{F}\left(\mathbf{w}\right)\right] \tag{8.2}$$

$$\mathcal{E}_{\text{gen}}(\mathbf{w}) \equiv \mathbb{E}\left[\hat{g}(\hat{f}\left(\mathbf{v}_{\text{new}}^\top\mathbf{w}\right), y_{\text{new}}\right)\right] \equiv \mathbb{E}\left[\hat{g}\left(\hat{f}(\mathbf{v}_{\text{new}}^\top\mathbf{w}), \boldsymbol{f}_0(\mathbf{u}_{\text{new}}^\top\boldsymbol{\theta}_0)\right)\right] . \tag{8.3}$$

Intuitively, the variables $\mathbf{u}_{\text{new}}^\top\boldsymbol{\theta}_0$ and $\mathbf{v}_{\text{new}}^\top\mathbf{w}$ will play a key role in the analysis. Given an instance of $\boldsymbol{\theta}_0$ and $\mathbf{w}$, the tuple $\left(\frac{1}{\sqrt{p}}\mathbf{u}_{\text{new}}^\top\boldsymbol{\theta}_0, \frac{1}{\sqrt{d}}\mathbf{v}_{\text{new}}^\top\mathbf{w}\right)$ is a bivariate Gaussian with covariance:

$$\begin{bmatrix} \frac{1}{p}\boldsymbol{\theta}_0^\top\Psi\boldsymbol{\theta}_0 & \frac{1}{\sqrt{dp}}(\Phi^\top\boldsymbol{\theta}_0)^\top\mathbf{w} \\ \frac{1}{\sqrt{dp}}(\Phi^\top\boldsymbol{\theta}_0)^\top\mathbf{w} & \frac{1}{d}\mathbf{w}^\top\Omega\mathbf{w} \end{bmatrix} . \tag{8.4}$$

We thus define the following overlaps, that will play a fundamental role in the analysis:

$$\rho = \frac{1}{p}\boldsymbol{\theta}_0^\top\Psi\boldsymbol{\theta}_0, \qquad m = \frac{1}{\sqrt{dp}}(\Phi^\top\boldsymbol{\theta}_0)^\top\mathbf{w}\,, \qquad q = \frac{1}{d}\mathbf{w}^\top\Omega\mathbf{w}, \qquad \chi = \frac{1}{d}\boldsymbol{\theta}_0^\top\Phi\Omega^{-1}\Phi^\top\boldsymbol{\theta}_0\,. \tag{8.5}$$

Note that here, we will not introduce the spectral decomposition 7.7 as it will not simplify the expressions as in the l_2 case. The representations are mathematically equivalent nonetheless. Our main result is that the distribution of the estimator $\hat{\mathbf{w}}$ can be exactly computed in the weak sense from the solution to six scalar fixed point equations with a unique solution.

8.1 Necessary assumptions

We start with a list of the necessary assumptions for the most generic version of the result to hold. We also briefly discuss how they are relevant in a supervised machine learning context.

(A1) The vector $\boldsymbol{\theta}_0$ is pulled from any given distribution $p_{\boldsymbol{\theta}_0} \in \mathbb{R}^p$ (this includes deterministic vectors with bounded norm), and is independent of the matrices U and $\mathcal{V}$. Additionally, the signal is non-vanishing and has finite squared norm, i.e. the following holds almost surely:

$$\lim_{p\to\infty} 0 < \mathbb{E}\left[\frac{\boldsymbol{\theta}_0^\top \boldsymbol{\theta}_0}{p}\right] < +\infty \tag{8.6}$$

(A2) The covariance matrices verify:

$$(\Psi, \Omega) \in \mathbb{S}_p^{++} \times \mathbb{S}_d^{++}, \quad \Omega - \Phi^\top \Psi^{-1} \Phi \succeq 0 \tag{8.7}$$

The spectral distributions of the matrices Φ, Ψ and Ω converge to distributions such that the overlaps defined by equation (8.5) are well-defined. Additionally, the maximum singular values of the covariance matrices are bounded with high probability when $n, p, d \to \infty$.

(A3) The functions $\mathbf{F}$ and g are proper, lower semi-continuous, convex functions. Additionally, we assume that the cost function $\mathbf{F} + g$ is coercive, i.e.:

$$\lim_{\|\mathbf{w}\|_2 \to +\infty} (\mathbf{F} + g)(\mathbf{w}) = +\infty \tag{8.8}$$

and that the following scaling condition holds : for all $n, d \in \mathbb{N}, \mathbf{z} \in \mathbb{R}^n$ and any constant $c > 0$, there exist finite, positive constants C_1, C_2, C_3, such that, for any standard normal random vectors $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{g} \in \mathbb{R}^n$:

$$\|\mathbf{z}\|_2 \leqslant c\sqrt{n} \implies \sup_{\mathbf{x} \in \partial g(\mathbf{z})} \|\mathbf{x}\|_2 \leqslant C_1\sqrt{n}, \qquad \frac{1}{d}\mathbb{E}\left[\mathbf{F}(\mathbf{h})\right] < +\infty, \qquad \frac{1}{n}\mathbb{E}\left[g(\mathbf{g})\right] < +\infty \tag{8.9}$$

(A4) The random elements of the function f_0 are independent of the matrices U and $\mathcal{V}$. Additionally the following limit exists and is finite

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n} f_0(U\boldsymbol{\theta}_0)^\top f_0(U\boldsymbol{\theta}_0)\right] < +\infty$$

(A5) When we send the dimensions n, p, d to infinity, they grow with finite ratios $\alpha = n/d$, $\gamma = p/d$.

(A6) **Additional assumptions for linear finite sample size rates** : the teacher vector $\boldsymbol{\theta}_0$ has sub-Gaussian one dimensional marginals. The functions $\mathbf{F}, g, \phi_1, \phi_2$ are pseudo-Lipschitz of finite order. The eigenvalues of the covariance matrices are bounded with probability one.

(A7) **Additional assumptions for exponential finite sample size rates**: all of the above, and the loss function g is separable and pseudo-Lipschitz of order 2, the regularisation is either a ridge or a Lipschitz function, the functions ϕ_1, ϕ_2 are respectively separable, pseudo-Lipschitz of order 2, and a square or Lipschitz function.

The first assumption (A1) ensures that the teacher distribution is non-vanishing. The positive definiteness in (A2) means the covariance matrices of the blocks U and V are well-specified. Note that the cross-correlation matrix Φ can have singular values equal to zero. The assumption about the limiting spectral distribution is essentially a summability condition which is immediately verified if the limiting spectral distributions have compact support, a common case. The scaling assumptions from (A3) are natural as they imply that non-diverging inputs result in non-diverging outputs in the functions f and g, as well as the sub-differentials. Similar scaling assumptions are encountered in proofs such as [281]. They also allow to show Gaussian concentration of Moreau envelopes, as we will see in Lemma 29. The *coercivity* assumption is verified in most common machine learning setups : any convex loss with ridge regularisation, or any convex loss that is bounded below with a coercive regularisation (LASSO, elastic-net,...), see Corollary 11.15 from [25]. Assumption (A4) is a classical assumption of teacher-student setups, where any correlation between the teacher and the student is modeled by the covariance matrices and not by the label generating function f_0. The summability condition ensures generalization error is well-defined for squared performance measures. Finally, (A5) is the typical *high-dimensional* limit used in statistical physics of learning, random matrix theory and a large recent body of work in high-dimensional statistical learning.

8.2 Main theorem

First, let's define quantities and a scalar optimization problem that will be used to state the asymptotic behaviour of (7.2-7.3):

Definition 13. *(Scalar potentials/replica free energy) Define the following functions of the scalar variables $\tau_1 > 0, \tau_2 > 0, \kappa \geqslant 0, \eta \geqslant 0, \nu, m$:*

$$\mathcal{L}_g(\tau_1, \kappa, m, \eta) = \frac{1}{n}\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa} g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right], \tag{8.10}$$
$$\mathcal{L}_{\mathbf{F}}(\tau_2, \eta, \nu, \kappa) = \frac{1}{d}\mathbb{E}\left[\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}(\nu\mathbf{t} + \kappa\mathbf{g})\right)\right],$$

where $\mathbf{s}, \mathbf{h} \sim \mathcal{N}(0, I_n)$ and $\mathbf{g} \sim \mathcal{N}(0, I_d)$ are random vectors independent of the other quantities, $\mathbf{t} = \Omega^{-1/2}\Phi^\top \boldsymbol{\theta}_0$, $\mathbf{y} = \boldsymbol{f}_0\left(\sqrt{\rho}\mathbf{s}\right)$, and $\mathcal{M}$ denotes the Moreau envelope of a target function.

From these quantities define the following potential:

$$\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m) = \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + m\nu\sqrt{\gamma} - \frac{\tau_2}{2\eta}\frac{m^2}{\rho}$$
$$- \frac{\eta}{2\tau_2}(\nu^2\chi + \kappa^2) + \alpha\mathcal{L}_g(\tau_1, \kappa, m, \eta) + \mathcal{L}_{\mathbf{F}}(\tau_2, \eta, \nu, \kappa)\,. \tag{8.11}$$

Under Assumption (8.1), the previously defined quantities all admit finite limits when $n, p, d \to \infty$.

Proof: This follows directly from Lemma 29.

The next lemma characterizes important properties of the "potential" function $\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$:

Lemma 25. *(Geometry and minimizers of $\mathcal{E}$) The function $\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ is jointly convex in (m, η, τ_1) and jointly concave in (ν, κ, τ_2), and the optimization problem*

$$\min_{m,\eta,\tau_1}\max_{\kappa,\nu,\tau_2}\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m) \tag{8.12}$$

has a unique solution $(\tau_1^*, \tau_2^*, \kappa^*, \eta^*, \nu^*, m^*)$ on $dom(\mathcal{E})$.

Proof: see Appendix 8.2.3. The optimality condition of problem (8.12) yields the set of self-consistent fixed point equations given in Lemma 36 of Chapter 8. Finally, define the following variables:

$$\mathbf{w}^* = \Omega^{-1/2}\text{prox}_{\frac{\eta^*}{\tau_2^*}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta^*}{\tau_2^*}(\nu^*\mathbf{t}+\kappa^*\mathbf{g})\right), \qquad \mathbf{z}^* = \text{prox}_{\frac{\tau_1^*}{\kappa^*}g(.,\mathbf{y})}\left(\frac{m^*}{\sqrt{\rho}}\mathbf{s}+\eta^*\mathbf{h}\right). \tag{8.13}$$

where prox denotes the proximal operator. With these definitions, we can now state our main result:

Theorem 14. *(Training loss and generalisation error) Under Assumption (8.1), there exist constants $C, c, c' > 0$ such that, for any optimal solution $\hat{\mathbf{w}}$ to (7.3), the training loss and generalisation error defined by equation verify, for any $0 < \epsilon < c'$:*

$$\mathbb{P}\left(|\mathcal{E}_{\text{train}}(\hat{\mathbf{w}}) - \mathcal{E}^*_{\text{train}}| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon^2}e^{-cn\epsilon^4}, \tag{8.14}$$
$$\mathbb{P}\left(\left|\mathcal{E}_{\text{gen}}(\hat{\mathbf{w}}) - \mathbb{E}_{\omega,\xi}\left[\hat{g}(f_0(\omega),\hat{f}(\xi))\right]\right| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon^2}e^{-cn\epsilon^4},$$

where $\mathcal{E}^*_{train}$ is defined as follows:

$$\mathcal{E}^*_{\text{train}} = \frac{1}{n}\mathbb{E}\left[g\left(\mathbf{z}^*,\mathbf{y}\right)\right] + \frac{1}{\alpha d}\mathbb{E}\left[\mathbf{F}\left(\mathbf{w}^*\right)\right], \tag{8.15}$$

and the random variables (ω, ξ) are jointly Gaussian with covariance

$$(\omega,\xi) \sim \mathcal{N}\left(0, \begin{bmatrix}\rho & m^* \\ m^* & q^*\end{bmatrix}\right), \; q^* = (\eta^*)^2 + \frac{(m^*)^2}{\rho}. \tag{8.16}$$

Proof: see Appendix 8.2.4. Note that the regularisation may be removed to evaluate the training loss. A more generic result, aiming directly at the estimator $\hat{\mathbf{w}}$, can also be stated:

Theorem 15. *Under Assumption (8.1), for any optimal solution $\hat{\mathbf{w}}$ to (7.3), denote $\hat{\mathbf{z}} = \frac{1}{\sqrt{d}}\mathcal{V}\hat{\mathbf{w}}$. Then, there exist constants $C, c, c' > 0$ such that, for any Lipschitz function $\phi_1 : \mathbb{R}^d \to \mathbb{R}$, and separable, pseudo-Lipschitz function $\phi_2 : \mathbb{R}^n \to \mathbb{R}$ and any $0 < \epsilon < c'$:*

$$\mathbb{P}\left(\left|\phi_1(\frac{\hat{\mathbf{w}}}{\sqrt{d}}) - \mathbb{E}\left[\phi_1\left(\frac{\mathbf{w}^*}{\sqrt{d}}\right)\right]\right| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon^2}e^{-cn\epsilon^4}, \tag{8.17}$$
$$\mathbb{P}\left(\left|\phi_2(\frac{\hat{\mathbf{z}}}{\sqrt{n}}) - \mathbb{E}\left[\phi_2\left(\frac{\mathbf{z}^*}{\sqrt{n}}\right)\right]\right| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon^2}e^{-cn\epsilon^4}. \tag{8.18}$$

Proof: see Appendix 8.2.4. Concentration still holds for a larger class of functions $\phi_{1,2}$, but exponential rates are lost. This is discussed in Appendix 8.1.

8.2.1 Theoretical toolbox

Here we remind a few known results that are used throughout the proof. We also provide proofs of useful, straightforward consequences of theses results that do not appear explicitly in the literature for completeness.

A Gaussian comparison theorem

We start with the Convex Gaussian Min-max Theorem, as presented in [281], which is a tight version of an inequality initially derived in [121].

Theorem 16. *(CGMT) Let $\mathbf{G} \in \mathbb{R}^{m\times n}$ be an i.i.d. standard normal matrix and $\mathbf{g} \in \mathbb{R}^m$, $\mathbf{h} \in \mathbb{R}^n$ two i.i.d. standard normal vectors independent of one another. Let $\mathcal{S}_\mathbf{w}, \mathcal{S}_\mathbf{u}$ be two compact sets such that $\mathcal{S}_\mathbf{w} \subset \mathbb{R}^n$ and $\mathcal{S}_\mathbf{u} \subset \mathbb{R}^m$. Consider the two following optimization problems for any continuous ψ on $\mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$:*

$$\mathbf{C}(\mathbf{G}) := \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \mathbf{u}^\top \mathbf{G}\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}), \tag{8.19}$$

$$\mathcal{C}(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \|\mathbf{w}\|_2 \mathbf{g}^\top \mathbf{u} + \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{w} + \psi(\mathbf{w}, \mathbf{u}) \tag{8.20}$$

then the following holds:

1. *For all $c \in \mathbb{R}$:*
$$\mathbb{P}(\mathbf{C}(\mathbf{G}) < c) \leqslant 2\mathbb{P}(\mathcal{C}(\mathbf{g}, \mathbf{h}) \leqslant c)$$
2. *Further assume that $\mathcal{S}_\mathbf{w}, \mathcal{S}_\mathbf{u}$ are convex sets and ψ is convex-concave on $\mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$. Then, for all $c \in \mathbb{R}$,*
$$\mathbb{P}(\mathbf{C}(\mathbf{G}) > c) \leqslant 2\mathbb{P}(\mathcal{C}(\mathbf{g}, \mathbf{h}) \geqslant c)$$
In particular, for all $\mu \in \mathbb{R}, t > 0, \mathbb{P}(|\mathbf{C}(\mathbf{G}) - \mu| > t) \leqslant 2\mathbb{P}(|\mathcal{C}(\mathbf{g}, \mathbf{h}) - \mu| \geqslant t)$.

Following [281], we will say that any reformulation of a target problem matching the form of (8.19) is an acceptable primary optimization problem (PO), and the corresponding form (8.20) is an acceptable auxiliary problem (AO). The main idea of this approach is to study the asymptotic properties of the (PO) by studying the simpler (AO).

Proximal operators and Moreau envelopes : differentials and useful functions

Here we remind the definition and some important properties of Moreau envelopes and proximal operators, key elements of convex analysis. Other properties will be used throughout the proof but at less crucial stages, thus we don't remind them explicitly. Our main reference for these properties will be [25].
Consider a closed, proper function f such that dom(f)$\subset \mathbb{R}^n$. Its Moreau envelope and proximal operator are respectively defined by :

$$\mathcal{M}_{\tau f}(\mathbf{x}) = \min_{\mathbf{z}\in\text{dom}(f)} \{f(\mathbf{z}) + \frac{1}{2\tau}\|\mathbf{x} - \mathbf{z}\|_2^2\}, \quad \text{prox}_{\tau f}(\mathbf{x}) = \arg\min_{\mathbf{z}\in\text{dom(f)}} \{f(\mathbf{z}) + \frac{1}{2\tau}\|\mathbf{x} - \mathbf{z}\|_2^2\} \tag{8.21}$$

As reminded in [281], the Moreau envelope is jointly convex in $(\tau, \mathbf{x})$ and differentiable almost everywhere, with gradients:

$$\nabla_\mathbf{x} \mathcal{M}_{\tau f}(\mathbf{x}) = \frac{1}{\tau}(\mathbf{x} - \text{prox}_{\tau f}(\mathbf{x})) \tag{8.22}$$

$$\frac{\partial}{\partial\tau} \mathcal{M}_{\tau f}(\mathbf{x}) = -\frac{1}{2\tau^2}\left\|\mathbf{x} - \text{prox}_{\tau f}(\mathbf{x})\right\|_2^2 \tag{8.23}$$

We remind that $\text{prox}_{\tau f}(\mathbf{x})$ is the unique point which solves the strongly convex optimization problem defining the Moreau envelope, i.e.:

$$\mathcal{M}_{\tau f}(\mathbf{x}) = f(\text{prox}_{\tau f}(\mathbf{x})) + \frac{1}{2\tau}\left\|\mathbf{x} - \text{prox}_{\tau f}(\mathbf{x})\right\|_2^2 \tag{8.24}$$

We also remind the definition of order k pseudo-Lipschitz function.

Definition 14. *Pseudo-Lipschitz function For $k \in \mathbb{N}^*$ and any $n, m \in \mathbb{N}^*$, a function $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is called a* pseudo-Lipschitz of order k *if there exists a constant L(k) such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$\|\phi(\mathbf{x}) - \phi(\mathbf{y})\|_2 \leqslant L(k)\left(1 + (\|\mathbf{x}\|_2)^{k-1} + (\|\mathbf{y}\|_2)^{k-1}\right)\|\mathbf{x} - \mathbf{y}\|_2 \tag{8.25}$$

We now give some further properties that will be helpful throughout the proof.

Lemma 26. *(Moreau envelope of pseudo-Lipschitz function) Consider a proper, lower-semicontinuous, convex, pseudo-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ of order k. Then its Moreau envelope is also pseudo-Lipschitz of order k.*

Proof of Lemma 26: For any $\mathbf{x}, \mathbf{y}$ in $\text{dom}(f)$, we have, using the pseudo-Lipschitz property:

$$\begin{aligned}\left|f(\text{prox}_{\tau f}(\mathbf{x})) - f(\text{prox}_{\tau f}(\mathbf{y}))\right| &\leqslant L(k)\left(1 + \left(\left\|\text{prox}_{\tau f}(\mathbf{x})\right\|_2\right)^{k-1} + \left(\left\|\text{prox}_{\tau f}(\mathbf{y})\right\|_2\right)^{k-1}\right) \\ &\qquad \left\|\text{prox}_{\tau f}(\mathbf{x}) - \text{prox}_{\tau f}(\mathbf{y})\right\|_2 \\ &\leqslant L(k)\left(1 + (\|\mathbf{x}\|_2)^{k-1} + (\|\mathbf{y}\|_2)^{k-1}\right)\|\mathbf{x} - \mathbf{y}\|_2 \end{aligned} \tag{8.26}$$

where the second line follows immediately with the same constant $L(k)$ owing to the firm-nonexpansiveness of the proximal operator. Furthermore

$$\begin{aligned}&\left\|\mathbf{x} - \text{prox}_{\tau f}(\mathbf{x})\right\|_2^2 - \left\|\mathbf{y} - \text{prox}_{\tau f}(\mathbf{y})\right\|_2^2 = \\ &\tau\left|\partial f(\text{prox}_{\tau f}(\mathbf{x})) + \partial f(\text{prox}_{\tau f}(\mathbf{y}))\right|\left|\left(\mathbf{x} - \text{prox}_{\tau f}(\mathbf{x}) - \mathbf{y} + \text{prox}_{\tau f}(\mathbf{y})\right)\right| \\ &\leqslant \tau\left\|\partial f(\text{prox}_{\tau f}(\mathbf{x})) + \partial f(\text{prox}_{\tau f}(\mathbf{y}))\right\|_2\left\|\left(\mathbf{x} - \text{prox}_{\tau f}(\mathbf{x}) - \mathbf{y} + \text{prox}_{\tau f}(\mathbf{y})\right)\right\|_2 \end{aligned} \tag{8.27}$$

due to the pseudo-Lipschitz property, one has

$$\partial f(\text{prox}_{\tau f}(\mathbf{x})) \leqslant L(k)\left(1 + 2\left\|\text{prox}_{\tau f}(\mathbf{x})\right\|_2^{k-1}\right) \tag{8.28}$$

This, along with the firm-nonexpansiveness of Id − prox, concludes the proof. □

Lemma 27. *(Useful functions) For any $\mathbf{x} \in \mathbb{R}^n, \tau > 0, \theta \in \mathbb{R}$ and any proper, convex lower semi-*

continuous function f, define the following functions:

$$\begin{aligned} h_1 &: \mathbb{R} \to \mathbb{R} \\ &\theta \mapsto \mathbf{x}^T prox_{\tau f(.)}(\theta \mathbf{x}) \end{aligned} \tag{8.29}$$

$$\begin{aligned} h_2 &: \mathbb{R} \to \mathbb{R} \\ &\tau \mapsto \frac{1}{2\tau^2} \left\| \mathbf{x} - prox_{\tau f(.)}(\mathbf{x}) \right\|_2^2 \end{aligned} \tag{8.30}$$

$$\begin{aligned} h_3 &: \mathbb{R} \to \mathbb{R} \\ &\tau \mapsto \left\| prox_{\frac{f}{\tau}(.)}(\frac{\mathbf{x}}{\tau}) \right\|_2^2 \end{aligned} \tag{8.31}$$

$$\begin{aligned} h_4 &: \mathbb{R} \to \mathbb{R} \\ &\tau \mapsto \left\| \mathbf{x} - prox_{\tau f}(\mathbf{x}) \right\|_2^2 \end{aligned} \tag{8.32}$$

h_1 is nondecreasing, and h_2, h_3, h_4 are nonincreasing.

Proof of Lemma 27: For any $\theta, \tilde{\theta} \in \mathbb{R}$:

$$\begin{aligned} (\theta - \tilde{\theta})(h_1(\theta) - h_1(\tilde{\theta})) &= (\theta \mathbf{x} - \tilde{\theta}\mathbf{x})^\top \left(\text{prox}_{\tau f(.)}(\theta \mathbf{x}) - \text{prox}_{\tau f(.)}(\tilde{\theta}\mathbf{x}) \right) \\ &\geqslant \left\| \text{prox}_{\tau f(.)}(\theta \mathbf{x}) - \text{prox}_{\tau f(.)}(\tilde{\theta}\mathbf{x}) \right\|_2^2 \\ &\geqslant 0 \end{aligned} \tag{8.33}$$

where the inequality comes from the firm non-expansiveness of the proximal operator. Thus h_1 is nondecreasing.
Since the Moreau envelope $\mathcal{M}_{\tau f}(\mathbf{x})$ is convex in τ, we have, for any $\tau, \tilde{\tau}$ in $\mathbb{R}_{++}$

$$(\tau - \tilde{\tau}) \left(\frac{\partial}{\partial \tau} \mathcal{M}_{\tau f}(\mathbf{x}) - \frac{\partial}{\partial \tilde{\tau}} \mathcal{M}_{\tilde{\tau} f}(\mathbf{x}) \right) \geqslant 0, \qquad \iff \qquad (\tau - \tilde{\tau}) \left(h_2(\tilde{\tau}) - h_2(\tau) \right) \geqslant 0 \tag{8.34}$$

which implies that h_2 is non-increasing.
Using the Moreau decomposition, see e.g. [25], we have:

$$h_2(\tau) = \frac{1}{2\tau^2} \left\| \mathbf{x} - \left(\mathbf{x} - \tau \text{prox}_{\frac{f^*}{\tau}} \left(\frac{\mathbf{x}}{\tau} \right) \right) \right\|_2^2 = \left\| \text{prox}_{\frac{f^*}{\tau}} \left(\frac{\mathbf{x}}{\tau} \right) \right\|_2^2 \tag{8.35}$$

which is a nonincreasing function of τ. Since f is convex, we can restart this short process with the conjugate of f to obtain the desired result. Thus h_3 is nonincreasing and $(\tau - \tilde{\tau})(h_3(\tau) - h_3(\tilde{\tau})) \leqslant 0$.
Moving to h_4, proving that it is nonincreasing is equivalent to proving that the following function is increasing

$$h_5(\tau) = \text{prox}_{\tau f}(\mathbf{x})^\top \left(2\mathbf{x} - \text{prox}_{\tau f}(\mathbf{x}) \right) \tag{8.36}$$

using the Moreau decomposition again

$$h_5(\tau) = \left(\mathbf{x} - \tau \text{prox}_{\frac{f^*}{\tau}} \left(\frac{\mathbf{x}}{\tau} \right) \right)^\top \left(\mathbf{x} + \tau \text{prox}_{\frac{f^*}{\tau}} \left(\frac{\mathbf{x}}{\tau} \right) \right) \tag{8.37}$$

then, for any $\tau, \tilde{\tau}$ in $\mathbb{R}_{++}$:

$$(\tau - \tilde{\tau})(h_5(\tau) - h_5(\tilde{\tau})) = (\tau - \tilde{\tau}) \left(\tilde{\tau}^2 \left\| \text{prox}_{\frac{f^*}{\tilde{\tau}}} \left(\frac{\mathbf{x}}{\tilde{\tau}} \right) \right\|_2^2 - \tau^2 \left\| \text{prox}_{\frac{f^*}{\tau}} \left(\frac{\mathbf{x}}{\tau} \right) \right\|_2^2 \right) \tag{8.38}$$

separating the cases $\tau \leqslant \tilde{\tau}$ and $\tau \geqslant \tilde{\tau}$, and using the result on h_3 then gives the desired result. □
The following inequality is similar to one that appeared in one-dimensional form in [281].

Lemma 28. *(A useful inequality) For any proper, lower semi-continuous convex function f, any $\mathbf{x}, \tilde{\mathbf{x}}$ in $dom(f)$, and any $\gamma, \tilde{\gamma} \in \mathbb{R}_{++}$, the following holds:*

$$\begin{aligned} \Big(prox_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) - prox_{\gamma f}(\mathbf{x}) \Big)^\top \left(\frac{\tilde{\mathbf{x}}}{\tilde{\gamma}} - \frac{\mathbf{x}}{\gamma} - \frac{1}{2} \left(\frac{1}{\tilde{\gamma}} - \frac{1}{\gamma} \right) \Big(prox_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) + prox_{\gamma f}(\mathbf{x}) \Big) \right) \\ \geqslant \left(\frac{1}{2\tilde{\gamma}} + \frac{1}{2\gamma} \right) \Big\| \Big(prox_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) - prox_{\gamma f}(\mathbf{x}) \Big) \Big\|_2^2 \end{aligned} \tag{8.39}$$

Proof of Lemma 28 : the subdifferential of a proper convex function is a monotone operator, thus:

$$\Big(\text{prox}_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) - \text{prox}_{\gamma f}(\mathbf{x}) \Big)^\top \Big(\partial f(\text{prox}_{\tilde{\gamma} f}(\tilde{\mathbf{x}})) - \partial f(\text{prox}_{\gamma f}(\mathbf{x})) \Big) \geqslant 0 \tag{8.40}$$

additionally, $\text{prox}_{\gamma f}(\mathbf{x}) = (\text{Id} + \gamma \partial \text{f})^{-1}(\mathbf{x})$, hence:

$$\begin{aligned} &\partial f(\text{prox}_{\tilde{\gamma} f}(\tilde{\mathbf{x}})) - \partial f(\text{prox}_{\gamma f}(\mathbf{x})) = \left(\frac{\tilde{\mathbf{x}}}{\tilde{\gamma}} - \frac{\mathbf{x}}{\gamma} - \frac{1}{\tilde{\gamma}} \text{prox}_{\tilde{\gamma}}(\tilde{\mathbf{x}}) + \frac{1}{\gamma} \text{prox}_{\gamma f}(\mathbf{x}) \right) \\ &= \frac{\tilde{\mathbf{x}}}{\tilde{\gamma}} - \frac{\mathbf{x}}{\gamma} - \frac{1}{\tilde{\gamma}} \text{prox}_{\tilde{\gamma}}(\tilde{\mathbf{x}}) + \frac{1}{\gamma} \text{prox}_{\gamma f}(\mathbf{x}) - \frac{1}{2} \left(\frac{1}{\tilde{\gamma}} - \frac{1}{\gamma} \right) \Big(\text{prox}_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) + \text{prox}_{\gamma f}(\mathbf{x}) \Big) \\ &+ \frac{1}{2} \left(\frac{1}{\tilde{\gamma}} - \frac{1}{\gamma} \right) \Big(\text{prox}_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) + \text{prox}_{\gamma f}(\mathbf{x}) \Big) \\ &= \left(\frac{\tilde{\mathbf{x}}}{\tilde{\gamma}} - \frac{\mathbf{x}}{\gamma} - \frac{1}{2} \left(\frac{1}{\tilde{\gamma}} - \frac{1}{\gamma} \right) \Big(\text{prox}_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) + \text{prox}_{\gamma f}(\mathbf{x}) \Big) \right) - \left(\frac{1}{2\tilde{\gamma}} + \frac{1}{2\gamma} \right) \Big(\text{prox}_{\tilde{\gamma} f}(\tilde{\mathbf{x}}) - \text{prox}_{\gamma f}(\mathbf{x}) \Big) \end{aligned} \tag{8.41}$$

which gives the desired inequality. □

Useful concentration of measure elements

We begin by reminding the Gaussian-Poincaré inequality, see e.g. [47].

Proposition 7. *(Gaussian Poincaré inequality)*
Let $\mathbf{g} \in \mathbb{R}^n$ be a $\mathcal{N}(0, I_n)$ random vector. Then for any continuous, weakly differentiable φ, there exists a constant c such that:

$$Var[\varphi(\mathbf{g})] \leqslant c \, \mathbb{E} \left[\|\nabla \varphi(\mathbf{g})\|_2^2 \right] \tag{8.42}$$

We now use this previous result to show Gaussian concentration of Moreau envelopes of appropriately scaled convex functions.

Lemma 29. *(Gaussian concentration of Moreau envelopes)*
Consider a proper, convex function $f : \mathbb{R}^n \to \mathbb{R}$ verifying the scaling conditions of Assumptions 8.1 and let $\mathbf{g} \in \mathbb{R}^n$ be a standard normal random vector. Then, for any parameter $\tau > 0$ and any $\epsilon > 0$, there exists a constant c such that the following holds:

$$\mathbb{P} \left(\left| \frac{1}{n} \mathcal{M}_{\tau f(.)}(\mathbf{g}) - \mathbb{E} \left[\frac{1}{n} \mathcal{M}_{\tau f(.)}(\mathbf{g}) \right] \right| \geqslant \epsilon \right) \leqslant \frac{c}{n \tau^2 \epsilon^2} \tag{8.43}$$

Proof of Lemma 29:

We start by showing that the Moreau envelope of a proper, convex function $f : \mathbb{R}^n \to \mathbb{R}$ verifying the scaling conditions of Assumptions 8.1 is integrable with respect to the Gaussian measure. Using the convexity of the optimization problem defining the Moreau envelope, and the fact that f is proper, there exists $\mathbf{z}_0 \in \mathbb{R}^n$ and a finite constant $\mathcal{K}$ such that :

$$\begin{aligned}\frac{1}{n}\mathcal{M}_{\tau f(.)}(\mathbf{g}) &\leqslant \frac{1}{n}f(\mathbf{z}_0) + \frac{1}{2n\tau}\|\mathbf{z}_0 - \mathbf{g}\|_2^2 \\ &\leqslant \mathcal{K} + \frac{1}{2n\tau}\|\mathbf{z}_0 - \mathbf{g}\|_2^2 \end{aligned} \tag{8.44}$$

where the second line is integrable under a multivariate Gaussian measure. Then, using Proposition 7, we get:

$$\text{Var}\left[\frac{1}{n}\mathcal{M}_{\tau f(.)}(\mathbf{g})\right] \leqslant \frac{c}{n^2}\mathbb{E}\left[\left\|\nabla_{\mathbf{z}}\mathcal{M}_{\tau f(.)}(\mathbf{g})\right\|_2^2\right] \tag{8.45}$$

$$= \frac{c}{n^2}\mathbb{E}\left[\left\|\frac{1}{\tau}\left(\mathbf{z} - \text{prox}_{\tau f}(\mathbf{g})\right)\right\|_2^2\right] \tag{8.46}$$

Using Proposition 12.27 and Corollary 4.3 from [25], $\mathbf{g} \to \mathbf{z} - \text{prox}_{\tau f}(\mathbf{g})$ is firmly non-expansive and:

$$\left\|\mathbf{g} - \text{prox}_{\tau f}(\mathbf{g})\right\|_2^2 \leqslant \langle \mathbf{g} | \mathbf{g} - \text{prox}_{\tau f}(\mathbf{g})\rangle \quad \text{which implies} \tag{8.47}$$

$$\left\|\mathbf{g} - \text{prox}_{\tau f}(\mathbf{g})\right\|_2^2 \leqslant \|\mathbf{g}\|_2^2 \quad \text{using the Cauchy-Schwarz inequality} \tag{8.48}$$

then

$$\text{Var}\left[\frac{1}{n}\mathcal{M}_{\tau f(.)}(\mathbf{g})\right] \leqslant \frac{c}{n^2\tau^2}\mathbb{E}\left[\|\mathbf{g}\|_2^2\right] = \frac{c}{n\tau^2} \tag{8.49}$$

Chebyshev's inequality then gives, for any $\epsilon > 0$:

$$\mathbb{P}\left(\left|\frac{1}{n}\mathcal{M}_{\tau f(.)}(\mathbf{g}) - \mathbb{E}\left[\frac{1}{n}\mathcal{M}_{\tau f(.)}(\mathbf{g})\right]\right| \geqslant \epsilon\right) \leqslant \frac{c}{n\tau^2\epsilon^2} \tag{8.50}$$

□

Gaussian concentration of pseudo-Lipschitz functions of finite order can also be proven using the Gaussian Poincaré inequality to yield a bound similar to the one obtained for Moreau envelopes. We thus give the result without proof:

Lemma 30. *(Concentration of pseudo-Lipschitz functions) Consider a pseudo-Lipschitz function of finite order k, $f : \mathbb{R}^n \to \mathbb{R}$. Then for any vector $\mathbf{g} \sim \mathcal{N}(0, I_n)$ and any $\epsilon > 0$, there exists a constant $C(k) > 0$ such that*

$$\mathbb{P}\left(\left|f(\frac{\mathbf{g}}{\sqrt{n}}) - \mathbb{E}\left[f(\frac{\mathbf{g}}{\sqrt{n}})\right]\right| \geqslant \epsilon\right) \leqslant \frac{L^2(k)C(k)}{n\epsilon^2} \tag{8.51}$$

We now cite an exponential concentration lemma for separable, pseudo-Lipschitz functions of order 2, taken from [180].

Lemma 31. *(Lemma B.5 from [180]) Consider a separable, pseudo-Lipschitz function of order 2, $f : \mathbb{R}^n \to \mathbb{R}$. Then for any vector $\mathbf{g} \sim \mathcal{N}(0, I_n)$ and any $\epsilon > 0$, there exists constants $C, c, c' > 0$ such that*

$$\mathbb{P}\left(\left|\frac{1}{n}f(\mathbf{g}) - \mathbb{E}\left[\frac{1}{n}f(\mathbf{g})\right]\right| \geqslant c'\epsilon\right) \leqslant Ce^{-cn\epsilon^2} \tag{8.52}$$

where it is understood that $f(\mathbf{g}) = \sum_{i=1}^n f(g_i)$.

8.2.2 Determining a candidate primary problem, auxiliary problem and its solution.

We start with a reformulation of the problem (7.2-7.3) in order to obtain an acceptable primary problem in the framework of Theorem 16. Partitioning the Gaussian distribution, we can rewrite the matrices U and $\mathcal{V}$ in the following way, introducing the standard normal vector:

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} \in \mathbb{R}^{p+d} \sim \mathcal{N}(0, \mathrm{I}_{p+d}) \tag{8.53}$$

We can then rewrite the vectors $\mathbf{u}, \mathbf{v}$ and matrices $U, \mathcal{V}$ as:

$$\mathbf{u} = \Psi^{1/2}\mathbf{a}, \quad U = A\Psi^{1/2} \tag{8.54}$$

$$\mathbf{v} = \Phi^\top\Psi^{-1/2}\mathbf{a} + \left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\mathbf{b}, \quad \mathcal{V} = A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2} \tag{8.55}$$

where the matrices A and B have independent standard normal entries and are independent of $\boldsymbol{\theta}_0$. The learning problem then becomes equivalent to :

$$\text{Generate labels according to :} \quad \mathbf{y} = f_0\left(\frac{1}{\sqrt{p}}A\Psi^{1/2}\boldsymbol{\theta}_0\right) \tag{8.56}$$

$$\text{Learn according to :} \quad \arg\min_{\mathbf{w}} g\left(\frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w}, \mathbf{y}\right) + \mathbf{F}(\mathbf{w}) \tag{8.57}$$

We are then interested in the optimal cost of the following problem

$$\min_{\mathbf{w}} \frac{1}{d}\left[g\left(\frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w}, \mathbf{y}\right) + \mathbf{F}(\mathbf{w})\right] \tag{8.58}$$

Introducing the auxiliary variable $\mathbf{z}$:

$$\min_{\mathbf{w}} g\left(\frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w}, \mathbf{y}\right) + \mathbf{F}(\mathbf{w}) \tag{8.59}$$

$$\iff \min_{\mathbf{w},\mathbf{z}} g(\mathbf{z}, \mathbf{y}) + \mathbf{F}(\mathbf{w})$$

$$\text{s.t. } \mathbf{z} = \frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w} \tag{8.60}$$

Introducing the corresponding Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^n$ and using strong duality, the problem is equivalent to :

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z}, \mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.61}$$

In the remainder of the proof, the preceding cost function will be denoted

$$\mathbf{C}(\mathbf{w},\mathbf{z}) = \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.62}$$

such that the problem reads $\min_{\mathbf{w},\mathbf{z}} \mathbf{C}(\mathbf{w},\mathbf{z})$. Theorem 16 requires working with compact feasibility sets. Adopting similar approaches to the ones from [281, 76], the next lemma shows that the optimization problem (8.61) can be equivalently recast as one over compact sets.

Lemma 32. *(Compactness of feasibility set) Let $\mathbf{w}^*, \mathbf{z}^*, \boldsymbol{\lambda}^*$ be optimal in (8.61). Then there exists positive constants $C_{\mathbf{w}}, C_{\mathbf{z}}$ and $C_{\boldsymbol{\lambda}}$ such that*

$$\mathbb{P}\left(\|\mathbf{w}^*\|_2 \leqslant C_{\mathbf{w}}\sqrt{d}\right) \xrightarrow[d\to\infty]{P} 1,\ \mathbb{P}\left(\|\mathbf{z}^*\|_2 \leqslant C_{\mathbf{z}}\sqrt{n}\right) \xrightarrow[n\to\infty]{P} 1,\ \mathbb{P}\left(\|\boldsymbol{\lambda}^*\|_2 \leqslant C_{\boldsymbol{\lambda}}\sqrt{n}\right) \xrightarrow[n\to\infty]{P} 1 \tag{8.63}$$

Proof of Lemma 32: consider the initial minimisation problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w},\mathbf{y}\right) + \mathbf{F}(\mathbf{w}) \tag{8.64}$$

From assumption (A3), the cost function $g+\mathbf{F}$ is coercive, proper and lower semi-continuous. Since it is proper, there exists $\mathbf{w}_0 \in \mathbb{R}^d$ such that $g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w},\mathbf{y}\right) + \mathbf{F}(\mathbf{w}) \in \mathbb{R}$. The coercivity implies that there exists $\eta \in]0,+\infty[$ such that, for every $\mathbf{w}\in\mathbb{R}^d$ satisfying $\|\mathbf{w}-\mathbf{w}_0\| \geqslant \eta$, $g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w},\mathbf{y}\right)+\mathbf{F}(\mathbf{w}) \geqslant g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w}_0,\mathbf{y}\right) + \mathbf{F}(\mathbf{w}_0)$. Let $S = \{\mathbf{w}\in\mathbb{R}^d | \|\mathbf{w}-\mathbf{w}_0\| \leqslant \eta\}$. Then $S\cap\mathbb{R}^d \neq \emptyset$ and S is compact. Then, there exists $\mathbf{w}^* \in S$ such that $g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w}^*,\mathbf{y}\right) + \mathbf{F}(\mathbf{w}^*) = \inf_{\mathbf{w}\in S} g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w},\mathbf{y}\right) + \mathbf{F}(\mathbf{w}) \leqslant g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w}_0,\mathbf{y}\right) + f(\mathbf{w}_0)$. Thus $g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w}^*,\mathbf{y}\right) + \mathbf{F}(\mathbf{w}^*) \in \inf_{\mathbf{w}\in\mathbb{R}^d} g\left(\frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w},\mathbf{y}\right) + \mathbf{F}(\mathbf{w})$ and the set of minimisers is bounded. Closure is immediately checked by considering a sequence of minimisers converging to $\mathbf{w}^*$.
We conclude that the set of minimisers of problem (8.64) is a non-empy compact set. Then there exists a constant $C_{\mathbf{w}}$ independent of the dimension d, such that:

$$\|\mathbf{w}\|_2 \leqslant C_{\mathbf{w}}\sqrt{d} \tag{8.65}$$

Now consider the equivalent formulation of problem (8.64):

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.66}$$

Its optimality condition reads :

$$\nabla_{\boldsymbol{\lambda}} : \frac{1}{\sqrt{d}}\mathcal{V}\mathbf{w} = \mathbf{z}, \qquad \nabla_{\mathbf{z}} : \boldsymbol{\lambda} \in \partial g(\mathbf{z},\mathbf{y}), \qquad \nabla_{\mathbf{w}} : \frac{1}{\sqrt{d}}\mathcal{V}^\top\boldsymbol{\lambda} \in \partial\mathbf{F}(\mathbf{w}) \tag{8.67}$$

The optimality condition in $\boldsymbol{\lambda}$ gives:

$$\begin{aligned}
\|\mathbf{z}\|_2 &\leqslant \left\|\frac{1}{\sqrt{d}}\mathcal{V}\right\|_{op} \|\mathbf{w}\|_2 \\
&\leqslant \left\|\frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\right\|_{op} \|\mathbf{w}\|_2 \\
&\leqslant \left[\left\|\Psi^{-1/2}\Phi\right\|_{op}\left\|\frac{1}{\sqrt{d}}A\right\|_{op} + \left\|\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right\|_{op}\left\|\frac{1}{\sqrt{d}}B\right\|_{op}\right] \|\mathbf{w}\|_2
\end{aligned} \tag{8.68}$$

According to assumption (A2), the operator norms of the matrices involving the covariance matrices are bounded with high probability and using known results on random matrices, see e.g. [287], the operator norms of $\frac{1}{\sqrt{d}}A$ and $\frac{1}{\sqrt{d}}B$ are bounded by finite constants with high probability when the dimensions go to infinity. Thus there exists a constant $C_{\mathbf{z}}$ also independent of d such that:

$$\mathbb{P}\left(\|\mathbf{z}\|_2 \leqslant C_{\mathbf{Z}}\sqrt{n}\right) \xrightarrow[n\to\infty]{P} 1 \tag{8.69}$$

Finally, the scaling condition from assumption (A3) directly shows that there exists a constant $C_{\boldsymbol{\lambda}}$ such that

$$\mathbb{P}\left(\|\boldsymbol{\lambda}\|_2 \leqslant C_{\boldsymbol{\lambda}}\sqrt{n}\right) \xrightarrow[n\to\infty]{P} 1 \tag{8.70}$$

This concludes the proof of Lemma 32. □

Defining the sets $\mathcal{S}_{\mathbf{w}} = \{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\|_2 \leqslant C_{\mathbf{w}}\sqrt{d}\}, \mathcal{S}_{\mathbf{z}} = \{\mathbf{z} \in \mathbb{R}^n | \|\mathbf{z}\|_2 \leqslant C_{\mathbf{z}}\sqrt{n}\}$ and $\mathcal{S}_{\boldsymbol{\lambda}} = \{\boldsymbol{\lambda} \in \mathbb{R}^n | \|\boldsymbol{\lambda}\|_2 \leqslant C_{\boldsymbol{\lambda}}\sqrt{n}\}$, the optimization problem can now be reduced to:

$$\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\mathbf{z}\in\mathcal{S}_{\mathbf{z}}} \max_{\boldsymbol{\lambda}\in\mathcal{S}_{\boldsymbol{\lambda}}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.71}$$

The rest of this section can then be summarized by the following lemma, the proof of which shows how to find an acceptable (PO) for problem (8.71), the corresponding (AO) and how to reduce the (AO) to a scalar optimization problem. At this point we will assume the teacher vector $\boldsymbol{\theta}_0$ is deterministic, and relax this assumption in paragraph 8.2.5. For this reason we do not add it to the initial list of assumptions in section 8.1.

Lemma 33. *(Scalar equivalent problem) In the framework of Theorem 16, acceptable (AO)s of problem (8.71) can be reduced to the following scalar optimization problems*

$$\text{For } \boldsymbol{\theta}_0 \notin Ker(\Phi^\top) : \max_{\kappa,\nu,\tau_2} \min_{m,\eta,\tau_1} \mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m) \tag{8.72}$$

$$\text{For } \boldsymbol{\theta}_0 \in Ker(\Phi^\top) : \max_{\kappa,\tau_2} \min_{\eta,\tau_1} \mathcal{E}_n^0(\tau_1,\tau_2,\kappa,\eta) \tag{8.73}$$

where

$$\begin{aligned}\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m) &= \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + m\nu\sqrt{\gamma} - \frac{\tau_2}{2\eta}\frac{m^2}{\rho} \\ &- \frac{\eta}{2\tau_2 d}(\nu\mathbf{v} + \kappa\Omega^{1/2}\mathbf{g})^\top\Omega^{-1}(\nu\mathbf{v} + \kappa\Omega^{1/2}\mathbf{g}) - \kappa\mathbf{g}^\top\left(\Sigma^{1/2} - \Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\mathbf{v} \\ &+ \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right) + \frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}\right)\right),\end{aligned} \tag{8.74}$$

$$\mathcal{E}_n^0(\tau_1,\tau_2,\kappa,\nu) = -\frac{\eta\tau_2}{2} + \frac{\kappa\tau_1}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}(\eta\mathbf{h}) + \frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}f(\Omega^{-1/2}.)}(\frac{\eta}{\tau_2}\kappa\mathbf{g}) - \frac{\eta}{2\tau_2 d}\kappa^2\mathbf{g}^\top\mathbf{g} \tag{8.75}$$

and

$$\Sigma = \Omega - \frac{\tilde{\mathbf{v}}\tilde{\mathbf{v}}^T}{\rho p} \quad \tilde{\mathbf{v}} = \Phi^T\boldsymbol{\theta}_0 \quad \rho = \frac{1}{p}\boldsymbol{\theta}_0^\top\Psi\boldsymbol{\theta}_0 \tag{8.76}$$

Proof of Lemma 33: We need to find an i.i.d. Gaussian matrix independent from the rest of the problem in order to use Theorem 16. We thus decompose the mixing matrix A by taking

conditional expectations w.r.t. $\mathbf{y}$, which amounts to conditioning on a linear subset of the Gaussian space generated by A. Dropping the feasibility sets for confort of notation in the following lines:

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \left(\left(\mathbb{E}\left[A|\mathbf{y}\right] + A - \mathbb{E}\left[A|\mathbf{y}\right] \right) \Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2} \right) \mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.77}$$

$$\iff \min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \Bigg(\left(\mathbb{E}\left[A|A\Psi^{1/2}\boldsymbol{\theta}_0\right] + A - \mathbb{E}\left[A|A\Psi^{1/2}\boldsymbol{\theta}_0\right] \right) \Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2} \Bigg) \mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.78}$$

Conditioning in Gaussian spaces amounts to doing orthogonal projections. Denoting $\tilde{\boldsymbol{\theta}}_0 = \Psi^{1/2}\boldsymbol{\theta}_0$ and $\tilde{A}$ a copy of A independent of $\mathbf{y}$, the minimisation problem then becomes:

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \left(\left(A\mathbf{P}_{\tilde{\boldsymbol{\theta}}_0} + \tilde{A}\mathbf{P}^\perp_{\tilde{\boldsymbol{\theta}}_0} \right) \Psi^{-1/2}\Phi + B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2} \right) \mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.79}$$

$$\iff \min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} A\mathbf{P}_{\tilde{\boldsymbol{\theta}}_0}\Psi^{-1/2}\Phi\mathbf{w} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \tilde{A}\mathbf{P}^\perp_{\tilde{\boldsymbol{\theta}}_0}\Psi^{-1/2}\Phi\mathbf{w} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.80}$$

$$\iff \min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \mathbf{s}\frac{\tilde{\boldsymbol{\theta}}_0^\top}{\left\|\tilde{\boldsymbol{\theta}_0}\right\|_2}\Psi^{-1/2}\Phi\mathbf{w} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \tilde{A}\mathbf{P}^\perp_{\tilde{\boldsymbol{\theta}}_0}\Psi^{-1/2}\Phi\mathbf{w} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} B\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.81}$$

where we used $\mathbf{P}_{\tilde{\boldsymbol{\theta}}_0} = \frac{\tilde{\boldsymbol{\theta}}_0\tilde{\boldsymbol{\theta}}_0^\top}{\left\|\tilde{\boldsymbol{\theta}}_0\right\|_2^2}$ and $\mathbf{s} = A\frac{\tilde{\boldsymbol{\theta}_0}}{\left\|\tilde{\boldsymbol{\theta}_0}\right\|_2}$. Knowing that $\tilde{A}, B$ are independent standard Gaussian matrices, and independent from $\mathbf{A}, \mathbf{y}, f_0$, we can rewrite the problem as :

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} \mathbf{s}\frac{\boldsymbol{\theta}_0^\top}{\left\|\Psi^{1/2}\boldsymbol{\theta}_0\right\|}\Phi\mathbf{w} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} Z\Sigma^{1/2}\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.82}$$

where $\Sigma = \Phi^\top\Psi^{-1/2}\mathbf{P}^\perp_{\tilde{\boldsymbol{\theta}}_0}\Psi^{-1/2}\Phi + \Omega - \Phi^\top\Psi^{-1}\Phi = \Omega - \Phi^\top\Psi^{-1/2}\mathbf{P}_{\tilde{\theta}_0}\Psi^{-1/2}\Phi$, and Z is a standard Gaussian matrix independent of $\mathbf{A}, \mathbf{y}, f_0$. Recall $\rho = \frac{1}{p}\boldsymbol{\theta}_0^\top\Psi\boldsymbol{\theta}_0$ from the main text. Replacing with the expression of $\tilde{\boldsymbol{\theta}}_0$ and letting $\tilde{\mathbf{v}} = \Phi^\top\boldsymbol{\theta}_0$, we have

$$\Sigma = \Omega - \phi^\top\Psi^{-1/2}\tilde{\boldsymbol{\theta}}_0\tilde{\boldsymbol{\theta}_0}^\top\Psi^{-1/2}\Phi\frac{1}{\left\|\tilde{\theta}_0\right\|_2^2} = \Omega - \frac{\phi^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi}{\boldsymbol{\theta}_0^\top\Psi\boldsymbol{\theta}_0} \tag{8.83}$$

$$= \Omega - \frac{\tilde{\mathbf{v}}\tilde{\mathbf{v}}^\top}{p\rho} \tag{8.84}$$

The problem then becomes

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{dp}} \mathbf{s}\frac{\tilde{\mathbf{v}}^\top}{\sqrt{\rho}}\mathbf{w} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} Z\Sigma^{1/2}\mathbf{w} - \boldsymbol{\lambda}^\top\mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) \tag{8.85}$$

Two cases must now be considered, $\boldsymbol{\theta}_0 \notin \mathrm{Ker}(\phi^\top)$ and $\boldsymbol{\theta}_0 \in \mathrm{Ker}(\phi^\top)$. Another possible case is $\Phi = 0_{p\times d}$, however it leads to the same steps as the case $\boldsymbol{\theta}_0 \in \mathrm{Ker}(\Phi^\top)$.

Case 1: $\boldsymbol{\theta}_0 \notin \mathbf{Ker}(\Phi^\top)$

It is tempting to invert the matrix $\Sigma^{1/2}$ to make the change of variable $\mathbf{w}_\perp = \Sigma^{1/2}\mathbf{w}$ and continue the calculation. However there is no guarantee that Σ is invertible : it is only semi-positive definite. Taking identities everywhere gives for examples $\mathbf{P}^\perp_{\tilde{\theta}_0}$ which is non-invertible. We thus introduce an additional variable:

$$\min_{\mathbf{w},\mathbf{z},\mathbf{p}} \max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \boldsymbol{\lambda}^\top \frac{1}{\sqrt{dp}} \mathbf{s} \frac{\tilde{\mathbf{v}}^\top}{\sqrt{\rho}} \mathbf{w} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} Z\mathbf{p} - \boldsymbol{\lambda}^\top \mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}(\mathbf{w}) + \boldsymbol{\mu}^\top \left(\Sigma^{1/2}\mathbf{w} - \mathbf{p}\right) \tag{8.86}$$

Here the minimisation on f and g is linked by the bilinear form $\boldsymbol{\lambda}^\top \mathbf{s}\tilde{\mathbf{v}}^\top \mathbf{w}$. We wish to separate them in order for the Moreau envelopes to appear later on in simple fashion. To do so, we introduce the orthogonal decomposition of $\mathbf{w}$ on the direction of $\tilde{\mathbf{v}}$:

$$\begin{aligned}
\mathbf{w} &= \left(\mathbf{P}_{\tilde{\mathbf{v}}} + \mathbf{P}^\perp_{\tilde{\mathbf{v}}}\right)\mathbf{w} = \frac{\tilde{\mathbf{v}}^\top \mathbf{w}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{P}^\perp_{\tilde{\mathbf{v}}}\mathbf{w} \\
&= \frac{\tilde{\mathbf{v}}^\top \mathbf{w}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp \text{ where } \mathbf{w}_\perp \perp \tilde{\mathbf{v}} \\
&= \frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp \quad \text{where } m = \tfrac{1}{\sqrt{dp}}\tilde{\mathbf{v}}^\top \mathbf{w}
\end{aligned} \tag{8.87}$$

where the parameter m corresponds to the one defined in (8.5). This gives the following, after introducing the scalar Lagrange multiplier $\nu \in \mathbb{R}$ to enforce the constraint $\mathbf{w}_\perp \perp \tilde{\mathbf{v}}$. Note that several methods can be used to express the orthogonality constraint, as in e.g. [76], but the one chosen here allows to complete the proof and match the replica prediction. Reintroducing the normalization, we then have the equivalent form for (8.58):

$$\begin{aligned}
\min_{m,\mathbf{w}_\perp,\mathbf{z},\mathbf{p}} \max_{\boldsymbol{\lambda},\boldsymbol{\mu},\nu} \frac{1}{d}\Bigg[&\boldsymbol{\lambda}^\top \frac{m}{\sqrt{\rho}}\mathbf{s} + \boldsymbol{\lambda}^\top \frac{1}{\sqrt{d}} Z\mathbf{m} - \boldsymbol{\lambda}^\top \mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) \\
&+ \boldsymbol{\mu}^\top \left(\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) - \mathbf{p}\right) - \nu\tilde{\mathbf{v}}^\top \mathbf{w}_\perp \Bigg]
\end{aligned} \tag{8.88}$$

A follow-up of the previous equations shows that the feasibility set now reads :

$$\begin{aligned}
\mathcal{S}_{m,\mathbf{w}_\perp,\mathbf{z},\mathbf{p},\boldsymbol{\lambda},\boldsymbol{\mu},\nu} = \Bigg\{ &m \in \mathbb{R}, \mathbf{w}_\perp \in \mathbb{R}^{d-1}, \mathbf{z} \in \mathbb{R}^n, \mathbf{p} \in \mathbb{R}^d, \boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^d, \nu \in \mathbb{R} \mid \\
&\sqrt{m^2 + \frac{\|\mathbf{w}_\perp\|_2^2}{d}} \leqslant C_{\mathbf{w}}, \|\mathbf{z}\|_2 \leqslant C_{\mathbf{z}}\sqrt{n}, \|\mathbf{p}\|_2 \leqslant \sigma_{max}(\Sigma^{1/2})C_{\mathbf{w}}\sqrt{d}, \|\boldsymbol{\lambda}\|_2 \leqslant C_{\boldsymbol{\lambda}}\sqrt{n} \Bigg\}
\end{aligned} \tag{8.89}$$

where the boundedness of $\|\mathbf{p}\|_2$ follows immediately from the assumptions on the covariance matrices and Lemma 32. We denote $\mathcal{S}_{\mathbf{p}} = \{\mathbf{p} \in \mathbb{R}^d | \|\mathbf{p}\|_2 \leqslant C_{\mathbf{p}}\}$ for some constant $C_{\mathbf{p}} \geqslant \sigma_{max}(\Sigma^{1/2})C_{\mathbf{w}}$.

The set $\mathcal{S}_{\mathbf{p}} \times \mathcal{S}_{\boldsymbol{\lambda}}$ is compact and the matrix Z is independent of all other random quantities of the problem, thus problem (8.88) is an acceptable (PO). We can now write the auxiliary optimization problem (AO) corresponding to the primary one (8.88), dropping the feasibility sets again for

convenience:

$$\min_{m,\mathbf{w}_\perp,\mathbf{z},\mathbf{p}} \max_{\boldsymbol{\lambda},\boldsymbol{\mu},\nu} \frac{1}{d}\Big[\boldsymbol{\lambda}^\top \frac{m}{\sqrt{\rho}}\mathbf{s} + \frac{1}{\sqrt{d}}\|\boldsymbol{\lambda}\|_2 \mathbf{g}^\top \mathbf{p} + \frac{1}{\sqrt{d}}\|\mathbf{p}\|_2 \mathbf{h}^\top \boldsymbol{\lambda} - \boldsymbol{\lambda}^\top \mathbf{z} + g(\mathbf{z},\mathbf{y}) + \mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) + \boldsymbol{\mu}^\top\left(\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) - \mathbf{p}\right) - \nu\tilde{\mathbf{v}}^\top \mathbf{w}_\perp\Big] \tag{8.90}$$

We now turn to the simplification of this problem.

The variable $\boldsymbol{\lambda}$ only appears in linear terms, we can thus directly optimize over its direction, introducing the positive scalar variable $\kappa = \|\boldsymbol{\lambda}\|_2/\sqrt{d}$:

$$\min_{m,\mathbf{w}_\perp,\mathbf{z},\mathbf{p}} \max_{\kappa,\boldsymbol{\mu},\nu} \frac{1}{d}\Big[\kappa \mathbf{g}^\top \mathbf{p} + \kappa\left\|\frac{m}{\sqrt{\rho}}\sqrt{d}\mathbf{s} + \|\mathbf{p}\|_2\mathbf{h} - \sqrt{d}\mathbf{z}\right\|_2 + g(\mathbf{z},\mathbf{y}) + \mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) + \boldsymbol{\mu}^\top\left(\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) - \mathbf{p}\right) - \nu\tilde{\mathbf{v}}^\top \mathbf{w}_\perp\Big] \tag{8.91}$$

The previous expression may not be convex-concave because of the term $\|\mathbf{p}\|_2\mathbf{h}$. However, it was shown in [281] that the order of the min and max can still be inverted in this case, because of the convexity of the original problem. As the proof would be very similar, we do not reproduce it. Inverting the max-min order and performing the linear optimization on $\mathbf{p}$ with $\eta = \|\mathbf{p}\|_2/\sqrt{d}$:

$$\max_{\kappa,\boldsymbol{\mu},\nu} \min_{m,\mathbf{w}_\perp,\mathbf{z},\eta}\Big\{ -\frac{\eta}{\sqrt{d}}\|\boldsymbol{\mu} + \kappa\mathbf{g}\|_2 + \frac{\kappa}{\sqrt{d}}\left\|\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h} - \mathbf{z}\right\|_2 + \frac{1}{d}\Big[g(\mathbf{z},\mathbf{y}) + \mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) + \boldsymbol{\mu}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) - \nu\tilde{\mathbf{v}}^\top\mathbf{w}_\perp\Big]\Big\} \tag{8.92}$$

using the following representation of the norm, as in [281], for any vector t, $\|t\|_2 = \min_{\tau>0} \frac{\tau}{2} + \frac{\|t\|_2^2}{2\tau}$:

$$\max_{\kappa,\boldsymbol{\mu},\nu,\tau_2} \min_{m,\mathbf{w}_\perp,\mathbf{z},\eta,\tau_1}\Big\{ \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} - \frac{\eta}{2\tau_2 d}\|\boldsymbol{\mu} + \kappa\mathbf{g}\|_2^2 + \frac{\kappa}{2\tau_1 d}\left\|\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h} - \mathbf{z}\right\|_2^2 + \frac{1}{d}\Big[g(\mathbf{z},\mathbf{y}) + \mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) + \boldsymbol{\mu}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) - \nu\tilde{\mathbf{v}}^\top\mathbf{w}_\perp\Big]\Big\} \tag{8.93}$$

performing the minimisation over $\mathbf{z}$ and recognizing the Moreau envelope of $g(.,\mathbf{y})$:

$$\max_{\kappa,\boldsymbol{\mu},\nu,\tau_2} \min_{m,\mathbf{w}_\perp,\eta,\tau_1}\Big\{ \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \beta\mathbf{h}\right) - \frac{\eta}{2\tau_2 d}\|\boldsymbol{\mu} + \kappa\mathbf{g}\|_2^2 + \frac{1}{d}\Big[\mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) + \boldsymbol{\mu}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right) - \nu\tilde{\mathbf{v}}^\top\mathbf{w}_\perp\Big]\Big\} \tag{8.94}$$

At this point we have a convex-concave problem. Inverting the min-max order, $\boldsymbol{\mu}$ appears in a well defined strictly convex least-square problem.

$$\max_{\kappa,\nu,\tau_2} \min_{m,\mathbf{w}_\perp,\eta,\tau_1} \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right) - \frac{\nu}{d}\tilde{\mathbf{v}}^\top\mathbf{w}_\perp + \frac{1}{d}\mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)$$
$$+\frac{1}{d}\max_{\boldsymbol{\mu}}\left\{-\frac{\eta}{2\tau_2}\|\boldsymbol{\mu}+\kappa\mathbf{g}\|_2^2 + \boldsymbol{\mu}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)\right\} \tag{8.95}$$

Solving it:

$$\max_{\boldsymbol{\mu}}\left\{-\frac{\eta}{2\tau_2}\|\boldsymbol{\mu}+\kappa\mathbf{g}\|_2^2 + \boldsymbol{\mu}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)\right\}$$
$$\boldsymbol{\mu}^* = \frac{\tau_2}{\eta}\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right) - \kappa\mathbf{g}$$
$$\text{with optimal cost } \frac{\tau_2}{2\eta}\left\|\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)\right\|_2^2 - \kappa\mathbf{g}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right) \tag{8.96}$$

remembering that $\Sigma = \Omega - \tilde{\mathbf{v}}\tilde{\mathbf{v}}^\top/(p\rho)$ and $\mathbf{w}_\perp \perp \tilde{\mathbf{v}}$, the optimal cost of this least-square problem simplifies to:

$$c^* = \frac{\tau_2}{2\eta}\left(\left\|\Omega^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)\right\|_2^2 - \frac{m^2}{\rho}d\right) - \kappa\mathbf{g}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right) \tag{8.97}$$

The (AO) then reads :

$$\max_{\kappa,\nu,\tau_2} \min_{m,\mathbf{w}_\perp,\eta,\tau_1}\left\{\frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right) - \frac{\tau_2}{2\eta}\frac{m^2}{\rho} - \frac{\nu}{d}\tilde{\mathbf{v}}^\top\mathbf{w}_\perp\right.$$
$$\left.+\frac{1}{d}\mathbf{F}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right) + \frac{\tau_2}{2\eta d}\left\|\Omega^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)\right\|_2^2 - \frac{\kappa}{d}\mathbf{g}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)\right\} \tag{8.98}$$

We now need to solve in $\mathbf{w}_\perp$. To do so, we can replace $\mathbf{F}$ with its convex conjugate and solve the least-square problem in $\mathbf{w}_\perp$. This will lead to a Moreau envelope of $\mathbf{F}^*$ in the introduced dual variable, which can be linked to the Moreau envelope of $\mathbf{F}$ by Moreau decomposition. Intuitively, it is natural to think that the corresponding primal variable will be $\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp = \mathbf{w}$ for any feasible $m, \mathbf{w}_\perp$. However, we would like to have an explicit follow-up of the variables we optimize on, as we had for the Moreau envelpe of g which is defined with $\mathbf{z}$, so we prefer to introduce a slack variable $\mathbf{w}' = \frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp$ with corresponding dual parameter $\boldsymbol{\eta}$ to show that the (AO) can be reformulated in terms of the original variable $\mathbf{w}$. Note that the feasibility set on $\mathbf{w}'$ is almost surely compact.

$$\max_{\kappa,\nu,\tau_2,\boldsymbol{\eta}} \min_{m,\mathbf{w}_\perp,\mathbf{w}',\eta,\tau_1} \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right) + \frac{1}{d}\mathbf{F}(\mathbf{w}') - \frac{1}{d}\boldsymbol{\eta}^T\mathbf{w}' - \frac{\tau_2}{2\eta}\frac{m^2}{\rho}$$
$$-\frac{\nu}{d}\tilde{\mathbf{v}}^\top\mathbf{w}_\perp + \frac{\tau_2}{2\eta d}\left\|\Omega^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)\right\|_2^2 - \frac{\kappa}{d}\mathbf{g}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right) + \frac{1}{d}\boldsymbol{\eta}^\top\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right) \tag{8.99}$$

Isolating the terms depending on $\mathbf{w}_\perp$, we get a strictly convex least-square problem, remembering that $\Omega \in \mathcal{S}_d^{++}$:

$$\max_{\kappa,\nu,\tau_2,\boldsymbol{\eta}} \min_{m,\mathbf{w}_\perp,\mathbf{w}',\eta,\tau_1} \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right) + \frac{1}{d}\mathbf{F}(\mathbf{w}') - \frac{1}{d}\boldsymbol{\eta}^T\mathbf{w}' - \frac{\tau_2}{2\eta}\frac{m^2}{\rho} + \boldsymbol{\eta}^\top\frac{m\sqrt{\kappa_2}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}$$
$$- \kappa\mathbf{g}^\top\Sigma^{1/2}\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} - \frac{\nu}{d}\tilde{\mathbf{v}}^\top\mathbf{w}_\perp + \frac{\tau_2}{2\eta d}\left\|\Omega^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right)\right\|_2^2 - \frac{\kappa}{d}\mathbf{g}^\top\Sigma^{1/2}\mathbf{w}_\perp + \frac{1}{d}\boldsymbol{\eta}^\top\mathbf{w}_\perp \quad (8.100)$$

$$\max_{\kappa,\nu,\tau_2,\boldsymbol{\eta}} \min_{m,\mathbf{w}',\eta,\tau_1} \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right) + \frac{1}{d}\mathbf{F}(\mathbf{w}') - \frac{1}{d}\boldsymbol{\eta}^T\mathbf{w}' - \frac{\tau_2}{2\eta}\frac{m^2}{\rho} + \boldsymbol{\eta}^\top\frac{m\sqrt{\kappa_2}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}$$
$$- \kappa\mathbf{g}^\top\Sigma^{1/2}\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \frac{1}{d}\Big[\min_{\mathbf{w}_\perp}\frac{\tau_2}{2\eta}\left\|\Omega^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right)\right\|_2^2 - \mathbf{w}_\perp^\top\left(\kappa\Sigma^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\tilde{\mathbf{v}}\right)\Big] \quad (8.101)$$

The quantity $\mathbf{g}^\top\Sigma^{1/2}\mathbf{w}_\perp$ is a Gaussian random variable with variance $\left\|\Sigma^{1/2}\mathbf{w}_\perp\right\|_2^2 = \mathbf{w}_\perp^\top(\Omega - \tilde{\mathbf{v}}\tilde{\mathbf{v}}^\top/(p\rho))\mathbf{w}_\perp = \mathbf{w}_\perp\Omega\mathbf{w}_\perp = \left\|\Omega^{1/2}\mathbf{w}_\perp\right\|_2^2$ using the expression of Σ and the orthogonality of $\mathbf{w}_\perp$ with respect to $\tilde{\mathbf{v}}$. We can thus change $\Sigma^{1/2}$ for $\Omega^{1/2}$ in front of $\mathbf{w}_\perp$ combined with $\mathbf{g}$. The least-square problem, its solution and optimal cost then read:

$$\min_{\mathbf{w}_\perp}\frac{\tau_2}{2\eta}\left\|\Omega^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} + \mathbf{w}_\perp\right)\right\|_2^2 - \mathbf{w}_\perp^\top\left(\kappa\Omega^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\tilde{\mathbf{v}}\right) \quad (8.102)$$

$$\mathbf{w}_\perp^* = \frac{\eta}{\tau_2}\Omega^{-1}\left(\kappa\Omega^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\mathbf{v}\right) - \frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} \quad (8.103)$$

$$\text{with optimal cost } -\frac{\eta}{2\tau_2}\left(\kappa\Omega^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\tilde{\mathbf{v}}\right)^\top\Omega^{-1}\left(\kappa\Omega^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\tilde{\mathbf{v}}\right) + \frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}^\top\left(\kappa\Omega^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\tilde{\mathbf{v}}\right) \quad (8.104)$$

replacing in the (AO) and simplifying :

$$\iff \max_{\kappa,\nu,\tau_2,\boldsymbol{\eta}} \min_{m,\mathbf{w}',\eta,\tau_1} \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right) + \frac{1}{d}\mathbf{F}(\mathbf{w}') - \frac{1}{d}\boldsymbol{\eta}^T\mathbf{w}' - \frac{\tau_2}{2\eta}\frac{m^2}{\rho}$$
$$- \kappa\mathbf{g}^\top\left(\Sigma^{1/2} - \Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} - \frac{\eta}{2\tau_2 d}\left(\kappa\Omega^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\tilde{\mathbf{v}}\right)^\top\Omega^{-1}\left(\kappa\Omega^{1/2}\mathbf{g} - \boldsymbol{\eta} + \nu\tilde{\mathbf{v}}\right) + m\nu\sqrt{\gamma} \quad (8.105)$$

Another strictly convex least-square problem appears on $\boldsymbol{\eta}$, the solution and optimal value of which read

$$\boldsymbol{\eta}^* = -\frac{\tau_2}{\eta}\Omega\mathbf{w}' + (\kappa\Omega^{1/2}\mathbf{g} + \nu\tilde{\mathbf{v}}) \quad (8.106)$$

$$\text{with optimal cost } \frac{\tau_2}{2\eta d}\mathbf{w}'^\top\Omega\mathbf{w}' - \mathbf{w}'^\top(\kappa\Omega^{1/2}\mathbf{g} + \nu\tilde{\mathbf{v}}) \quad (8.107)$$

At this point we have expressed feasible solutions of $\boldsymbol{\eta}, \mathbf{w}_\perp$ as functions of the remaining variables. For any feasible solution in those variables, $\mathbf{w}$ and $\mathbf{w}'$ are the same. Replacing in the (AO) and a

completion of squares leads to

$$\max_{\kappa,\nu,\tau_2}\min_{m,\eta,\tau_1}\frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}+m\nu\sqrt{\gamma}-\frac{\tau_2}{2\eta}\frac{m^2}{\rho}-\frac{\eta}{2\tau_2 d}(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g})^\top\Omega^{-1}(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g})$$
$$-\kappa\mathbf{g}^\top\left(\Sigma^{1/2}-\Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\min_{\mathbf{w}'}\left\{\mathbf{F}(\mathbf{w}')+\frac{\tau_2}{2\eta}\left\|\Omega^{1/2}\mathbf{w}'-\frac{\eta}{\tau_2}(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}))\right\|_2^2\right\} \tag{8.108}$$

Recognizing the Moreau envelope of f and introducing the variable $\tilde{\mathbf{w}}=\Omega^{1/2}\mathbf{w}'=\Omega^{1/2}\mathbf{w}$, it follows:

$$\max_{\kappa,\nu,\tau_2}\min_{m,\eta,\tau_1}\frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}+m\nu\sqrt{\gamma}-\frac{\tau_2}{2\eta}\frac{m^2}{\rho}-\frac{\eta}{2\tau_2 d}(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g})^\top\Omega^{-1}(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g})$$
$$-\kappa\mathbf{g}^\top\left(\Sigma^{1/2}-\Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)+\frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right) \tag{8.109}$$

where the Moreau envelopes of f and g are respectively defined w.r.t. the variables $\mathbf{w}''$ and $\mathbf{z}$. At this point we have reduced the initial high-dimensional minimisation problem (8.90) to a scalar problem over six parameters. Another follow-up of the feasibility set shows that there exist positive constants C_m, C_κ, C_η independent of n, p, d such that $0\leqslant\kappa\leqslant C_\kappa$, $0\leqslant\eta\leqslant C_\eta$ and $0\leqslant m\leqslant C_m$.

Case 2: $\boldsymbol{\theta}_0\in\mathbf{Ker}(\Phi^\top)$ In this case, the min-max problem (8.85) becomes:

$$\min_{\mathbf{w},\mathbf{z}}\max_{\boldsymbol{\lambda}}\boldsymbol{\lambda}^\top\frac{1}{\sqrt{d}}Z\Omega^{1/2}\mathbf{w}-\boldsymbol{\lambda}^\top\mathbf{z}+g(\mathbf{z},\mathbf{y})+f(\mathbf{w}) \tag{8.110}$$

Since Ω is positive definite, we can define $\tilde{\mathbf{w}}=\Omega^{1/2}\mathbf{w}$ and write the equivalent problem:

$$\min_{\tilde{\mathbf{w}},\mathbf{z}}\max_{\boldsymbol{\lambda}}\boldsymbol{\lambda}^\top\frac{1}{\sqrt{d}}Z\tilde{\mathbf{w}}-\boldsymbol{\lambda}^\top\mathbf{z}+g(\mathbf{z},\mathbf{y})+f(\Omega^{-1/2}\tilde{\mathbf{w}}) \tag{8.111}$$

where the compactness of the feasibility set is preserved almost surely from the almost sure boundedness of the eigenvalues of Ω. We can thus write the corresponding auxiliary optimization problem, reintroducing the normalization by d:

$$\min_{\tilde{\mathbf{w}},\mathbf{z}}\max_{\boldsymbol{\lambda}}\frac{1}{d}\left[\|\boldsymbol{\lambda}\|_2\frac{1}{\sqrt{d}}\mathbf{g}^\top\tilde{\mathbf{w}}+\|\tilde{\mathbf{w}}\|_2\frac{1}{\sqrt{d}}\mathbf{h}^\top\boldsymbol{\lambda}-\boldsymbol{\lambda}^\top\mathbf{z}+g(\mathbf{z},\mathbf{y})+f(\Omega^{-1/2}\tilde{\mathbf{w}})\right] \tag{8.112}$$

introducing the convex conjugate of f with dual parameter $\boldsymbol{\eta}$:

$$\min_{\tilde{\mathbf{w}},\mathbf{z}}\max_{\boldsymbol{\lambda},\boldsymbol{\eta}}\frac{1}{d}\left[\|\boldsymbol{\lambda}\|_2\frac{1}{\sqrt{d}}\mathbf{g}^\top\tilde{\mathbf{w}}+\|\mathbf{w}_\perp\|_2\frac{1}{\sqrt{d}}\mathbf{h}^\top\boldsymbol{\lambda}-\boldsymbol{\lambda}^\top\mathbf{z}+g(\mathbf{z},\mathbf{y})+\boldsymbol{\eta}^\top\Omega^{-1/2}\tilde{\mathbf{w}}-f^*(\boldsymbol{\eta})\right] \tag{8.113}$$

We then define the scalar quantities $\kappa=\frac{\|\boldsymbol{\lambda}\|_2}{\sqrt{d}}$ and $\eta=\frac{\|\tilde{\mathbf{w}}\|_2}{\sqrt{d}}$ and perform the linear optimization on $\boldsymbol{\lambda},\tilde{\mathbf{w}}$, giving the equivalent:

$$\min_{\mathbf{z},\eta\geqslant 0}\max_{\boldsymbol{\eta},\kappa\geqslant 0}-\frac{\eta}{\sqrt{d}}\left\|\kappa\mathbf{g}-\Omega^{-1/2}\boldsymbol{\eta}\right\|_2+\frac{\kappa}{\sqrt{d}}\|\eta\mathbf{h}-\mathbf{z}\|_2+\frac{1}{d}g(\mathbf{z},\mathbf{y})-\frac{1}{d}f^*(\boldsymbol{\eta}) \tag{8.114}$$

Using the square root trick with parameters τ_1,τ_2:

$$\min_{\tau_1>0,\mathbf{z},\eta\geqslant 0}\max_{\tau_2>0,\boldsymbol{\eta},\kappa\geqslant 0}-\frac{\eta\tau_2}{2}-\frac{\eta}{2\tau_2 d}\left\|\kappa\mathbf{g}-\Omega^{-1/2}\boldsymbol{\eta}\right\|_2^2+\frac{\kappa\tau_1}{2}+\frac{\kappa}{2\tau_1 d}\|\eta\mathbf{h}-\mathbf{z}\|_2^2+\frac{1}{d}g(\mathbf{z},\mathbf{y})-\frac{1}{d}f^*(\boldsymbol{\eta}) \tag{8.115}$$

performing the optimizations on $\mathbf{z}, \boldsymbol{\eta}$ and recognizing the Moreau envelopes, the problem becomes:

$$\min_{\tau_1>0,\eta\geqslant 0}\max_{\tau_2>0,\kappa\geqslant 0} -\frac{\eta\tau_2}{2}+\frac{\kappa\tau_1}{2}+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}(\eta\mathbf{h})-\frac{1}{d}\mathcal{M}_{\frac{\tau_2}{\eta}f^*(\Omega^{1/2}.)}(\kappa\mathbf{g}) \tag{8.116}$$

$$\iff \min_{\tau_1>0,\eta\geqslant 0}\max_{\tau_2>0,\kappa\geqslant 0} -\frac{\eta\tau_2}{2}+\frac{\kappa\tau_1}{2}+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}(\eta\mathbf{h})+\frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}f(\Omega^{-1/2}.)}(\frac{\eta}{\tau_2}\kappa\mathbf{g})-\frac{\eta}{2\tau_2 d}\kappa^2\mathbf{g}^\top\mathbf{g} \tag{8.117}$$

This concludes the proof of Lemma 33. □

8.2.3 Study of the scalar equivalent problem : geometry and asymptotics.

Here we study the geometry, solutions and asymptotics of the scalar optimization problem (8.109). We will focus on the case $\boldsymbol{\theta}_0 \notin \mathrm{Ker}(\Phi^\top)$ as the other case simply shows that no learning is performed (see the remark at the end of this section). The following lemma characterizes the continuity and geometry of the cost function $\mathcal{E}_n$.

Lemma 34. *(Geometry of $\mathcal{E}_n$) Recall the function:*

$$\begin{aligned}\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m) &= \frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}+m\nu\sqrt{\gamma}-\frac{\tau_2}{2\eta}\frac{m^2}{\rho}-\frac{\eta}{2\tau_2 d}(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g})^\top\Omega^{-1}(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g})\\ &-\kappa\mathbf{g}^\top\left(\Sigma^{1/2}-\Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)+\frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right)\end{aligned} \tag{8.118}$$

Then $\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m)$ is continuous on its domain, jointly convex in (m,η,τ_1) and jointly concave in (κ,ν,τ_2).

Proof of Lemma 34 : $\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m)$ is a linear combination of linear and quadratic terms with Moreau envelopes, which are all continuous on their domain. Remembering the formulation

$$\begin{aligned}\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m) &= \frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}+m\nu\sqrt{\gamma}-\frac{\tau_2}{2\eta}\frac{m^2}{\rho}-\kappa\mathbf{g}^\top\left(\Sigma^{1/2}-\Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}\\ &+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)-\frac{1}{d}\mathcal{M}_{\frac{\tau_2}{\eta}f^*(\Omega^{1/2}.)}\left(\Omega^{-1/2}\left(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g}\right)\right)\end{aligned} \tag{8.119}$$

and using the properties of Moreau envelopes, $\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)$ is jointly convex in (κ,τ_1,m,η) as a composition of convex functions of those arguments. The same applies for

$$\mathcal{M}_{\frac{\tau_2}{\eta}f^*(\Omega^{1/2}.)}\left(\Omega^{-1/2}\left(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g}\right)\right) \tag{8.120}$$

jointly convex in (τ_2,η,ν,κ) and its opposite is jointly concave in those parameters. The remaining terms being linear in τ_1,τ_2,ν, we conclude that $\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m)$ is jointly concave in (ν,τ_2) and convex in τ_1 whatever the values of (κ,η,m). Going back to equation (8.93), we can write

$$\begin{aligned}\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m) &= \max_{\boldsymbol{\mu}}\min_{\mathbf{z},\mathbf{w}_\perp}\frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}-\frac{\eta}{2\tau_2 d}\|\boldsymbol{\mu}+\kappa\mathbf{g}\|_2^2+\frac{\kappa}{2\tau_1 d}\left\|\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}-\mathbf{z}\right\|_2^2\\ &+\frac{1}{d}\left[g(\mathbf{z},\mathbf{y})+f\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)+\boldsymbol{\mu}^\top\Sigma^{1/2}\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)-\nu\tilde{\mathbf{v}}^\top\mathbf{w}_\perp\right]\end{aligned} \tag{8.121}$$

The squared term in $m, \eta, \mathbf{z}$ can be written as

$$\frac{\kappa}{2\tau_1 d}\left\|\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}-\mathbf{z}\right\|_2^2=\tau_1\frac{\kappa}{2d}\left\|\frac{m}{\tau_1\sqrt{\rho}}\mathbf{s}+\frac{\eta}{\tau_1}\mathbf{h}-\frac{\mathbf{z}}{\tau_1}\right\|_2^2 \tag{8.122}$$

which is the perspective function with parameter τ_1 of a function jointly convex in $(\mathbf{z}, m, \eta)$. Thus it is jointly convex in $(\tau_1, \mathbf{z}, m, \eta)$. Furthermore, the term $f\left(\frac{m\sqrt{dp}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}+\mathbf{w}_\perp\right)$ is a composition of a convex function with a linear one, thus it is jointly convex in $(m, \mathbf{w}_\perp)$. The remaining terms in τ_1, η, m are linear. Since minimisation on convex sets preserves convexity, minimizing with respect to $\mathbf{z}, \mathbf{w}_\perp$ will lead to a jointly convex function in (τ_1, η, m). Similarly, the term $-\frac{\eta}{2\tau_2 d}\|\boldsymbol{\mu}+\kappa\mathbf{g}\|_2^2$ is jointly concave in $\tau_2, \kappa, \boldsymbol{\mu}$, and maximizing over $\boldsymbol{\mu}$ will result in a jointly concave function in (τ_2, ν, κ). We conclude that $\mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ is jointly convex in (τ_1, m, η) and jointly concave in (κ, ν, τ_2). □

The next lemma then characterizes the infinite dimensional limit of the scalar optimization problem (8.109), along with the consistency of its optimal value.

Lemma 35. *(Asymptotics of $\mathcal{E}_n$) Recall the following quantities:*

$$\mathcal{L}_g(\tau_1,\kappa,m,\eta)=\frac{1}{n}\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)\right]\ \textit{where}\ \mathbf{y}=f_0(\sqrt{\rho_p}\mathbf{s}),\ \mathbf{s}\sim\mathcal{N}(0,\mathbb{I}_n) \tag{8.123}$$

$$\mathcal{L}_{\mathbf{F}}(\tau_2,\eta,\nu,\kappa)=\frac{1}{d}\mathbb{E}\left[\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right)\right]\ \textit{where}\ \tilde{\mathbf{v}}=\Phi^\top\boldsymbol{\theta}_0 \tag{8.124}$$

$$\chi=\frac{1}{d}\boldsymbol{\theta}_0^\top\Phi\Omega^{-1}\Phi^\top\boldsymbol{\theta}_0 \tag{8.125}$$

$$\rho=\frac{1}{p}\boldsymbol{\theta}_0^\top\Psi\boldsymbol{\theta}_0 \tag{8.126}$$

and the potential:

$$\mathcal{E}(\tau_1,\tau_2,\kappa,\eta,\nu,m)=\frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}+m\nu\sqrt{\gamma}-\frac{\tau_2}{2\eta}\frac{m^2}{\rho}-\frac{\eta}{2\tau_2}(\nu^2\chi+\kappa^2)+\alpha\mathcal{L}_g(\tau_1,\kappa,m,\eta)+\mathcal{L}_{\mathbf{F}}(\tau_2,\eta,\nu,\kappa) \tag{8.127}$$

Then:

$$\max_{\kappa,\nu,\tau_2}\min_{m,\eta,\tau_1}\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m)\xrightarrow[n,p,d\to\infty]{P}\max_{\kappa,\nu,\tau_2}\min_{m,\eta,\tau_1}\mathcal{E}(\tau_1,\tau_2,\kappa,\eta,\nu,m) \tag{8.128}$$

and $\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ is continuously differentiable on its domain, jointly convex in (m, η, τ_1) and jointly concave in (κ, ν, τ_2).

Proof of Lemma 35: The strong law of large numbers, see e.g. [88] gives $\frac{1}{d}\mathbf{g}^\top\mathbf{g}\xrightarrow[d\to\infty]{a.s.}1$. Additionally, using assumption (A2) on the summability of $\boldsymbol{\theta}_0$ and (A3) on the boundedness of the spectrum of the covariance matrices, the quantity $\chi=\lim_{d\to\infty}\frac{1}{d}\boldsymbol{\theta}_0^\top\Phi\Omega^{-1}\Phi^\top\boldsymbol{\theta}_0$ exists and is finite. Since $\boldsymbol{\theta}_0\notin\mathrm{Ker}(\Phi^\top)$ and using the non-vanishing signal hypothesis, the quantity $\rho_{\tilde{\mathbf{v}}}=\lim_{d\to\infty}\frac{1}{d}\tilde{\mathbf{v}}^\top\tilde{\mathbf{v}}$ exists, is finite and strictly positive. Then $\kappa\mathbf{g}^\top\left(\Sigma^{1/2}-\Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\mathbf{v}$ is a centered Gaussian random

variable with variance verifying:

$$\begin{aligned}\mathrm{Var}\left[\kappa \mathbf{g}^\top \left(\Sigma^{1/2} - \Omega^{1/2}\right) \frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}\right] &\leqslant \kappa^2 \sigma_{max}^2 \left(\Sigma^{1/2} - \Omega^{1/2}\right) \frac{m^2\gamma}{\|\tilde{\mathbf{v}}\|_2^2} \\ &= \kappa^2 \sigma_{max}^2 \left(\Sigma^{1/2} - \Omega^{1/2}\right) \frac{m^2\gamma}{d\rho_{\tilde{\mathbf{v}}}} \end{aligned} \tag{8.129}$$

Using lemma 32, κ and m are finitely bounded independently of the dimension d. $\gamma, \sigma_{max}\left(\Sigma^{1/2} - \Omega^{1/2}\right)$ are finite. Thus there exists a finite constant C such that the standard deviation of

$$\kappa \mathbf{g}^\top \left(\Sigma^{1/2} - \Omega^{1/2}\right) \frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} \tag{8.130}$$

is smaller than $\sqrt{C}/\sqrt{d}$. Then, for any $\epsilon > 0$:

$$\begin{aligned}\mathbb{P}\left(\left|\kappa \mathbf{g}^\top \left(\Sigma^{1/2} - \Omega^{1/2}\right) \frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}}\right| \geqslant \epsilon\right) &\leqslant \mathbb{P}\left(|\mathcal{N}(0,1)| \geqslant \epsilon\sqrt{d}/\sqrt{C}\right) \\ &\leqslant \frac{\sqrt{C}}{\epsilon\sqrt{d}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\epsilon^2 d}{C}\right) \end{aligned} \tag{8.131}$$

using the Gaussian tail. The Borel-Cantelli lemma and summability of this tail gives

$$\kappa \mathbf{g}^\top \left(\Sigma^{1/2} - \Omega^{1/2}\right) \frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\tilde{\mathbf{v}} \xrightarrow[d\to\infty]{a.s.} 0 \tag{8.132}$$

Concentration of the Moreau envelopes of both f and g follows directly from lemma 29.
We thus have the pointwise convergence:

$$\mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m) \xrightarrow[n,p,d\to\infty]{P} \mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m) \tag{8.133}$$

Since pointwise convergence preserves convexity, $\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ is jointly convex in (m, η, τ_1) and jointly concave in (κ, ν, τ_2).
Now recall the expression of $\mathcal{E}$

$$\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m) = \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + m\nu\sqrt{\gamma} - \frac{\tau_2}{2\eta}\frac{m^2}{\rho} - \frac{\eta}{2\tau_2}(\nu^2\chi + \kappa^2) + \alpha\mathcal{L}_g(\tau_1, \kappa, m, \eta) + \mathcal{L}_f(\tau_2, \eta, \nu, \kappa) \tag{8.134}$$

The feasibility sets of κ, η, m are compact from Lemma 32 and the subsequent follow-up of the feasibility sets. Then, using Proposition 12.32 from [25], for fixed $(\tau_2, \kappa, \eta, \nu, m)$, we have:

$$\lim_{\tau_1\to+\infty} \frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right) = \frac{1}{d}\inf_{\mathbf{z}\in\mathbb{R}^n} g(\mathbf{z}, \mathbf{y}) \tag{8.135}$$

which is a finite quantity since $g(., \mathbf{y})$ is a proper, convex function verifying the scaling assumptions 8.1. Then, since $\kappa > 0$, we have:

$$\lim_{\tau_1\to+\infty} \mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m) = +\infty \tag{8.136}$$

Similarly, for fixed $(\tau_1, \kappa, \eta, \nu, m)$ and noting that composing f with the positive definite matrix $\Omega^{-1/2}$ does not change its convexity, or it being proper and lower semi-continuous, we get:

$$\lim_{\tau_2 \to +\infty} \frac{1}{d} \mathcal{M}_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)} \left(\frac{\eta}{\tau_2} \left(\nu \Omega^{-1/2} \tilde{\mathbf{v}} + \kappa \mathbf{g} \right) \right) = \frac{1}{d} f(0_d) \tag{8.137}$$

which is also a bounded quantity from the scaling assumptions made on f. Since $\beta > 0$, we then have:

$$\lim_{\tau_2 \to +\infty} \mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m) = -\infty \tag{8.138}$$

Finally, the limit $\lim_{\nu \to +\infty} \mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ needs to be checked for both $+\infty$ and $-\infty$ since there is no restriction on the sign of ν. From the definition of the Moreau envelope, we can write:

$$\frac{1}{d} \mathcal{M}_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)} \left(\frac{\eta}{\tau_2} \left(\nu \Omega^{-1/2} \tilde{\mathbf{v}} + \kappa \mathbf{g} \right) \right) \leqslant \frac{1}{d} f(0_d) + \frac{\tau_2}{2\eta} \left\| \frac{\eta}{d\tau_2} \left(\nu \Omega^{-1/2} \tilde{\mathbf{v}} + \kappa \mathbf{g} \right) \right\|_2^2 \tag{8.139}$$

Thus, for any fixed $(\tau_1, \tau_2, m, \kappa, \eta)$:

$$\mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m) \leqslant \frac{\kappa \tau_1}{2} - \frac{\eta \tau_2}{2} + m\nu\sqrt{\gamma} - \kappa \mathbf{g}^\top \left(\Sigma^{1/2} - \Omega^{1/2} \right) \frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2} \mathbf{v} + \frac{1}{d} \mathcal{M}_{\frac{\tau_1}{\kappa} g(.,\mathbf{y})} \left(\frac{m}{\sqrt{\rho}} \mathbf{s} + \eta \mathbf{h} \right) + \frac{1}{d} f(0_d) \tag{8.140}$$

which immediately gives $\lim_{\nu \to -\infty} \mathcal{E}_n = -\infty$. Turning to the other limit, remembering that $\mathcal{E}_n$ is continuously differentiable on its domain, we have:

$$\frac{\partial \mathcal{E}_n}{\partial \nu}(\tau_1, \tau_2, \kappa, \eta, \nu, m) = m\sqrt{\gamma} - \frac{1}{d} \tilde{\mathbf{v}}^\top \Omega^{-1/2} \mathrm{prox}_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)} \left(\frac{\eta}{\tau_2} \left(\nu \Omega^{-1/2} \tilde{\mathbf{v}} + \kappa \mathbf{g} \right) \right) \tag{8.141}$$

Thus $\lim_{\nu \to +\infty} \frac{\partial \mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m)}{\partial \nu} \to -\infty$. Since $\mathcal{E}_n$ is continuously differentiable in ν on $[0, +\infty[$, and from the short argument led above, we have shown

$$\lim_{|\nu| \to +\infty} \mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m) = -\infty \tag{8.142}$$

Using similar arguments as in the proof of Lemma 32, we can now reduce the feasibility set of τ_1, τ_2, ν to a compact one. Then, using the fact that convergence of convex functions on compact sets implies uniform convergence [6], we obtain

$$\max_{\kappa, \nu, \tau_2} \min_{m, \eta, \tau_1} \mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m) \xrightarrow[n,p,d \to +\infty]{P} \max_{\kappa, \nu, \tau_2} \min_{m, \eta, \tau_1} \mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m) \tag{8.143}$$

which is the desired result. □

At this point, it is necessary to characterize the set of solutions of the asymptotic minimisation problem (8.12). We start with the explicit form of the optimality condition associated to any solution.

Lemma 36. *(Fixed point equations) The zero-gradient condition of the optimization problem (8.12) prescribes the following set of fixed point equations for any feasible solution:*

$$\partial_\kappa : \tau_1 = \frac{1}{d}\mathbb{E}\left[\mathbf{g}^\top prox_{\frac{\eta}{\tau_2}f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right)\right] \tag{8.144}$$

$$\partial_\nu : m\sqrt{\gamma} = \frac{1}{d}\mathbb{E}\left[\tilde{\mathbf{v}}^\top\Omega^{-1/2} prox_{\frac{\eta}{\tau_2}f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right)\right] \tag{8.145}$$

$$\partial_\eta : \tau_2 = \alpha\frac{\kappa}{\tau_1}\eta - \frac{\kappa\alpha}{\tau_1 n}\mathbb{E}\left[\mathbf{h}^\top prox_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)\right] \tag{8.146}$$

$$\partial_{\tau_2} : \frac{1}{2d}\frac{\tau_2}{\eta}\mathbb{E}\left[\left\|\frac{\eta}{\tau_2}(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}) - prox_{\frac{\eta}{\tau_2}f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right)\right\|_2^2\right] =$$

$$\frac{\eta}{2\tau_2}(\nu^2\chi+\kappa^2) - m\nu\sqrt{\gamma} - \kappa\tau_1 + \frac{\eta\tau_2}{2} + \frac{\tau_2}{2\eta}\frac{m^2}{\rho} \tag{8.147}$$

$$\partial_m : \nu\sqrt{\gamma} = \alpha\frac{\kappa}{n\tau_1}\mathbb{E}\left[(\frac{m}{\eta\rho}\mathbf{h} - \frac{\mathbf{s}}{\sqrt{\rho}})^\top prox_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)\right] \tag{8.148}$$

$$\partial_{\tau_1} : \frac{\tau_1^2}{2} = \frac{1}{2}\alpha\frac{1}{n}\mathbb{E}\left[\left\|\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h} - prox_{\frac{\tau_1}{\kappa}g(.,y)}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)\right\|_2^2\right] \tag{8.149}$$

This set of equations can be converted to the replica notations using the table (8.204).

Proof of Lemma 36: Using arguments similar to the ones in the proof of Lemma 29, Moreau envelopes and their derivatives verify the necessary conditions of the dominated convergence theorem. Additionally, uniform convergence of the sequence of derivatives can be verified in a straightforward manner as all involved functions are firmly non-expansive and integrated w.r.t. Gaussian measures. We can therefore invert the limits and derivatives, and invert expectations and derivatives. We can now write explicitly the optimality condition for the scalar problem (8.127), using the expressions for derivatives of Moreau envelopes from Appendix 8.2.1. Some algebra and replacing with prescriptions obtained from each partial derivative leads to the set of equations above. □

Remark : Here we see that the potential function (8.127) can be further studied using the fixed point equations (36) and the relation (8.24). For any optimal $(\tau_1, \tau_2, \kappa, \eta, \nu, m)$, it holds that

$$\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$$
$$= \alpha\frac{1}{n}\mathbb{E}\left[g\left(\text{prox}_{\frac{\tau_1}{\kappa}g(.,y)}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right),\mathbf{y}\right)\right] + \frac{1}{d}\mathbb{E}\left[f\left(\Omega^{-1/2}\text{prox}_{\frac{\eta}{\tau_2}f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right)\right)\right] \tag{8.150}$$

Finally, we give a strict-convexity and strict-concavity property of the asymptotic potential $\mathcal{E}$ which will be helpful to prove Lemma 25.

Lemma 37. *(Strict convexity and strict concavity near minimisers) Consider the asymptotic potential function $\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$. Then for any fixed (η, m, τ_1) in their feasibility sets, the function*

$$\tau_2, \kappa, \nu \to \mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m) \tag{8.151}$$

is jointly strictly concave in (τ_2, κ, ν).
Additionally, consider the set $\mathcal{S}_{\partial_{\nu,\tau_2}}$ *defined by:*

$$\mathcal{S}_{\partial_{\nu,\tau_2}} = \left\{ \tau_1, \tau_2, \kappa, \eta, \nu, m \mid m\sqrt{\gamma} = \frac{1}{d}\mathbb{E}\left[\tilde{\mathbf{v}}^T \Omega^{-1/2} prox_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}\right)\right)\right], \right.$$
$$\left. \frac{1}{2d}\frac{1}{\eta}\mathbb{E}\left[\left\| prox_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}\right)\right)\right\|_2^2\right] = \frac{\eta}{2} + \frac{1}{2\eta}\frac{m^2}{\rho} \right\} \quad (8.152)$$

then for any fixed τ_2, κ, ν *in* $\mathcal{S}_{\partial_{\nu,\tau_2}}$, *the function* $(\eta, m, \tau_1) \to \mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ *is jointly strictly convex in* (η, m, τ_1) *on* $\mathcal{S}_{\partial_{\nu,\tau_2}}$

Proof of Lemma 37: We will use the following first order characterization of strictly convex functions: f is strictly convex $\iff$ $\langle \mathbf{x} - \mathbf{y} | \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{y}) \rangle > 0 \ \forall \mathbf{x} \neq \mathbf{y} \in \mathrm{dom}(f)$. To simplify notations, we will write, for any fixed (m, η, τ_1)

$$(\nabla_{\kappa,\nu,\tau_2}\mathcal{E}) = \left((\partial_\kappa \mathcal{E}, \partial_\nu \mathcal{E}, \partial_{\tau_2}\mathcal{E})(\tau_1, \tau_2, \kappa, \eta, \nu, m)\right)_i \quad (8.153)$$

as the i-th component of the gradient of $\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ with respect to (κ, ν, τ_2) for any fixed (m, η, τ_1) in the feasibility set. Then for any distinct triplets $(\kappa, \nu, \tau_2), (\tilde{\kappa}, \tilde{\nu}, \tilde{\tau}_2)$ and fixed (η, m, τ_1) in the feasibility set, determining the partial derivatives of $\mathcal{E}$ in similar fashion as is implied in the proof of Lemma 36, we have:

$$\begin{aligned}
&((\kappa, \nu, \tau_2) - (\tilde{\kappa}, \tilde{\nu}, \tilde{\tau}_2))^\top (\nabla \mathcal{E}_{\kappa,\nu,\tau_2} - \nabla \mathcal{E}_{\tilde{\kappa},\tilde{\nu},\tilde{\tau}_2}) \\
&= (\kappa - \tilde{\kappa})\alpha \frac{1}{2\tau_1}\frac{1}{n}\left(\mathbb{E}\left[\left\|\mathbf{r}_1 - \mathrm{prox}_{\frac{\tau_1}{\kappa} g(.,\mathbf{y})}(\mathbf{r}_1)\right\|_2^2 - \left\|\mathbf{r}_1 - \mathrm{prox}_{\frac{\tau_1}{\tilde{\kappa}} g(.,\mathbf{y})}(\mathbf{r}_1)\right\|_2^2\right]\right) \\
&+ \left(\mathrm{prox}_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\mathbf{r}_2\right) - \mathrm{prox}_{\frac{\eta}{\tilde{\tau}_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tilde{\tau}_2}\tilde{\mathbf{r}}_2\right)\right)^\top \left(\tilde{\mathbf{r}}_2 - \mathbf{r}_2 \right. \\
&\qquad \left. + \frac{\tau_2 - \tilde{\tau}_2}{2\eta d}\left(\mathrm{prox}_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\mathbf{r}_2\right) + \mathrm{prox}_{\frac{\eta}{\tilde{\tau}_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tilde{\tau}_2}\tilde{\mathbf{r}}_2\right)\right)\right) \\
&\leqslant (\kappa - \tilde{\kappa})\alpha \frac{1}{2\tau_1}\frac{1}{n}\left(\mathbb{E}\left[\left\|\mathbf{r}_1 - \mathrm{prox}_{\frac{\tau_1}{\kappa} g(.,y)}(\mathbf{r}_1)\right\|_2^2 - \left\|\mathbf{r}_1 - \mathrm{prox}_{\frac{\tau_1}{\tilde{\kappa}} g(.,y)}(\mathbf{r}_1)\right\|_2^2\right]\right) \\
&\qquad + \left(\frac{(\tau_2 + \tilde{\tau}_2)}{2\eta d}\mathbb{E}\left[-\left\|\mathrm{prox}_{\frac{\eta}{\tau_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\mathbf{r}_2\right) - \mathrm{prox}_{\frac{\eta}{\tilde{\tau}_2} f(\Omega^{-1/2}.)}\left(\frac{\eta}{\tilde{\tau}_2}\tilde{\mathbf{r}}_2\right)\right\|_2^2\right]\right)
\end{aligned} \quad (8.154)$$

where the last line follows from the inequality in Lemma 28, and we defined the shorthands, $\mathbf{r}_1 = \frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}$, $\mathbf{r}_2 = \nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}, \tilde{\mathbf{r}}_2 = \tilde{\nu}\Omega^{-1/2}\mathbf{v} + \tilde{\kappa}\mathbf{g}$. Using Lemma 27, the first term of the r.h.s of the last inequality is also negative as an increment of a nonincreasing function. Thus, both expectations are taken on negative functions. If those functions are not zero almost everywhere with respect to the Lebesgue measure, then the result will be strictly negative. Moreover, the functional taking each operator T to its resolvent $(\mathrm{Id} + \mathrm{T})^{-1}$ is a bijection on the set of non-trivial, maximally monotone operators, see e.g. [25] Proposition 23.21 and the subsequent discussion. The subdifferential of a proper, closed, convex function being maximally monotone, for two different parameters the corresponding proximal operator cannot be equal almost everywhere. The previously studied increment $((\kappa, \nu, \tau_2) - (\tilde{\kappa}, \tilde{\nu}, \tilde{\tau}_2))^\top (\nabla \mathcal{E}_{\kappa,\nu,\tau_2} - \nabla \mathcal{E}_{\tilde{\kappa},\tilde{\nu},\tilde{\tau}_2})$ is therefore strictly negative, giving the desired strict concavity in (κ, ν, τ_2). Restricting ourselves to the set $\mathcal{S}_{\partial\nu,\tau_2}$, the increment in (m, η, τ_1) can be written similarly. Note that $Id - \mathrm{prox}$ will appear in the expressions instead

of prox. The appropriate terms can then be brought to the form of the inequality from Lemma 28 using Moreau's decomposition. Using the definitions of the set $\mathcal{S}_{\partial\nu,\tau_2}$ and the increments from Lemma 27, a similar argument as the previous one can be carried out. The lemma is proved. □

What is now left to do is link the properties of the scalar optimization problem (8.12) to the original learning problem (7.3) using the tight inequalities from Theorem 16.

Remark: in the case $\boldsymbol{\theta}_0 \in \mathrm{Ker}(\Phi^T)$, the cost function $\mathcal{E}_n^0$ will uniformly converge to the following potential:

$$-\frac{\eta\tau_2}{2}+\frac{\kappa\tau_1}{2}-\frac{\eta}{2\tau_2}\kappa^2+\frac{\alpha}{n}\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}(\eta\mathbf{h})\right]+\frac{1}{d}\mathbb{E}\left[\mathcal{M}_{\frac{\eta}{\tau_2}f(\Omega^{-1/2}.)}(\frac{\eta}{\tau_2}\kappa\mathbf{g})\right] \tag{8.155}$$

As we will see in the next section, this will lead to estimators solely based on noise.

8.2.4 Back to the original problem : proof of Theorem 14 and 15

We begin this part by considering that the "necessary assumptions for exponential rates" from the set of assumptions 8.1 are verified. In the end we will discuss how relaxing these assumptions modifies the convergence speed. We closely follow the analysis introduced in [204] and further developed in [57]. The main difference resides in checking the concentration properties of generic Moreau envelopes depending on the regularity of the target function instead of specific instances such as the LASSO. Since the dimensions n, p, d are linked by multiplicative constants, we can express the rates with any of the three. Recall the original reformulation of the problem defining the student.

$$\max_{\boldsymbol{\lambda}}\min_{\mathbf{w},\mathbf{z}} g(\mathbf{z},\mathbf{y})+f(\mathbf{w})+\boldsymbol{\lambda}^\top\left(\frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi+B\left(\Omega-\Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\mathbf{w}-\mathbf{z}\right) \tag{8.156}$$

Introducing the variable $\tilde{\mathbf{w}}=\Omega^{1/2}\mathbf{w}$ it can be equivalently written, since Ω is almost surely invertible and the problem is convex concave with a closed convex feasibility set on $\tilde{\mathbf{w}},\mathbf{z}$.

$$\min_{\tilde{\mathbf{w}},\mathbf{z}}\max_{\boldsymbol{\lambda}} g(\mathbf{z},\mathbf{y})+f(\Omega^{-1/2}\tilde{\mathbf{w}})+\boldsymbol{\lambda}^\top\left(\frac{1}{\sqrt{d}}\left(A\Psi^{-1/2}\Phi+B\left(\Omega-\Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\Omega^{-1/2}\tilde{\mathbf{w}}-\mathbf{z}\right) \tag{8.157}$$

Recall the equivalent scalar auxiliary problem at finite dimension $\mathcal{E}_n$ and its asymptotic counterpart $\mathcal{E}$ both defined on the same variables as the original problem $\tilde{\mathbf{w}},\mathbf{z}$ through the Moreau envelopes of g and r:

$$\mathcal{E}(\tau_1,\tau_2,\kappa,\eta,\nu,m)=\frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}+m\nu\sqrt{\gamma}-\frac{\tau_2}{2\eta}\frac{m^2}{\rho}-\frac{\eta}{2\tau_2}(\nu^2\chi+\kappa^2)+\alpha\mathcal{L}_g(\tau_1,\kappa,m,\eta)+\mathcal{L}_f(\tau_2,\eta,\nu,\kappa) \tag{8.158}$$

$$\begin{aligned}\mathcal{E}_n(\tau_1,\tau_2,\kappa,\eta,\nu,m)=&\frac{\kappa\tau_1}{2}-\frac{\eta\tau_2}{2}+m\nu\sqrt{\gamma}-\frac{\tau_2}{2\eta}\frac{m^2}{\rho}-\frac{\eta}{2\tau_2 d}(\nu\tilde{\mathbf{v}}+\kappa\Omega^{1/2}\mathbf{g})^\top\Omega^{-1}(\nu\mathbf{v}+\kappa\Omega^{1/2}\mathbf{g})\\&-\kappa\mathbf{g}^\top\left(\Sigma^{1/2}-\Omega^{1/2}\right)\frac{m\sqrt{\gamma}}{\|\tilde{\mathbf{v}}\|_2^2}\mathbf{v}+\frac{1}{d}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s}+\eta\mathbf{h}\right)+\frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}}+\kappa\mathbf{g}\right)\right)\end{aligned} \tag{8.159}$$

Recall the variables:

$$\tilde{\mathbf{w}}^* = \text{prox}_{\frac{\eta^*}{\tau_2^*} f(\Omega^{-1/2}.)}(\frac{\eta^*}{\tau_2^*}(\nu^*\mathbf{t} + \kappa^*\mathbf{g})), \qquad \mathbf{z}^* = \text{prox}_{\frac{\tau^*}{\frac{1}{\kappa^*}g(.,\mathbf{y})}}\left(\frac{m^*}{\sqrt{\rho}}\mathbf{s} + \eta^*\mathbf{h}\right) \tag{8.160}$$

Denote $(\tau_1^*, \tau_2^*, \kappa^*, \eta^*, \nu^*, m^*)$ the unique solution to the optimization problem (8.12) and $\mathcal{E}^*$ the corresponding optimal cost. $\mathcal{E}^*$ defines a strongly convex optimization problem (due to the Moreau envelopes) on $\tilde{\mathbf{w}}, \mathbf{z}$ whose solution is given by Eq.(8.160). Similarly, denote $(\tau_{1,n}^*, \tau_{2,n}^*, \kappa_n^*, \eta_n^*, \nu_n^*, m_n^*)$ any solution to the optimization problem on $\mathcal{E}_n$ and $\mathcal{E}_n^*$ the corresponding optimal value. Finally, we write $E_n(\tilde{\mathbf{w}}, \mathbf{z})$ the cost function of the optimization problem on $\tilde{\mathbf{w}}, \mathbf{z}$ defined by $\mathcal{E}_n^*$ for any optimal solution $(\tau_{1,n}^*, \tau_{2,n}^*, \kappa_n^*, \eta_n^*, \nu_n^*, m_n^*)$, such that:

$$\mathcal{E}_n^* = \min_{\tilde{\mathbf{w}}, \mathbf{z}} E_n(\tilde{\mathbf{w}}, \mathbf{z}) \tag{8.161}$$

By the definition of Moreau envelopes, we have that $E_n(\tilde{\mathbf{w}}, \mathbf{z})$ is $\frac{\kappa_n^*}{2d\tau_{1,n}^*}$ strongly convex in $\mathbf{z}$ and $\frac{\tau_{2,n}^*}{2d\eta_n^*}$ strongly convex in $\tilde{\mathbf{w}}$. The following lemma ensures that these strong convexity constants are non-zero for any finite n.

Lemma 38. *Consider the finite size scalar optimization problem*

$$\max_{\kappa,\nu,\tau_2} \min_{m,\eta,\tau_1} \mathcal{E}_n(\tau_1, \tau_2, \kappa, \eta, \nu, m) \tag{8.162}$$

where the feasibility set of $(\tau_1, \tau_2, \kappa, \eta, \nu, m)$ is compact and $\tau_1 > 0, \tau_2 > 0$. Then any optimal values κ^*, τ_2^* verify:

$$\kappa^* \neq 0 \quad \tau_2^* \nrightarrow 0 \tag{8.163}$$

Proof of Lemma 38: from the analysis carried out in the proof of Lemma 35, the feasibility set of the optimization problem is compact. Suppose $\kappa^* = 0$. Then the value of m minimizing the cost function is $-\infty$, which contradicts the compactness of the feasibility set. A similar argument holds for τ_2. □

The next lemma characterizes the speed of convergence of the optimal value of the finite dimensional scalar optimization problem to its asymptotic counterpart, which has a unique solution in $\tau_1, \tau_2, \kappa, \eta, \nu, m$. The intuition is that, using the strong convexity of the auxiliary problems, we can show that the solution in $\tilde{\mathbf{w}}, \tilde{\mathbf{z}}$ to the finite size problem $\mathcal{E}_n^*$ converges to the solution $\tilde{\mathbf{w}}^*, \tilde{\mathbf{z}}^*$ of the asymptotic problem $\mathcal{E}^*$, with convergence rates governed by those of the finite size cost towards its asymptotic counterpart.

Lemma 39. *For any $\epsilon > 0$, there exist constants C, c, γ such that:*

$$\mathbb{P}\left(|\mathcal{E}_n^* - \mathcal{E}^*| \geqslant \gamma\epsilon\right) \leqslant \frac{C}{\epsilon}\exp^{-cn\epsilon^2} \tag{8.164}$$

which is equivalent to

$$\mathbb{P}\left(\left|\min_{\tilde{\mathbf{w}}, \mathbf{z}} E_n(\tilde{\mathbf{w}}, \mathbf{z}) - \mathcal{E}^*\right| \geqslant \gamma\epsilon\right) \leqslant \frac{C}{\epsilon}\exp^{-cn\epsilon^2} \tag{8.165}$$

Proof of Lemma 39: for any fixed $(\tau_1, \tau_2, \kappa, \nu, \eta, m)$, we can determine the rates of convergence of all the random quantities in $\mathcal{E}_n$. The linear terms involving $\frac{1}{d}\mathbf{g}^T\mathbf{v}$ are sub-Gaussian with sub-Gaussian norm bounded by C/d for some constant $C > 0$. Thus we can find constants, $C, c > 0$ such that, for any $\epsilon > 0$:

$$\mathbb{P}\left(\left|\frac{1}{d}\mathbf{g}^T\tilde{\mathbf{v}}\right| \geqslant \epsilon\right) \leqslant Ce^{-cn\epsilon^2} \tag{8.166}$$

The term involving $\mathbf{v}^T\Omega\mathbf{v}$ is deterministic in this setting. We will see in section 8.2.5 how a random $\boldsymbol{\theta}_0$ affects the convergence rates. The term involving $\frac{1}{d}\mathbf{g}^T\mathbf{g}$ is a weighted sum of sub-exponential random variables, the tail of which can be determined using Bernstein's inequality, see e.g. [288] Corollary 2.8.3, which gives a sub-Gaussian tail for small deviations and a sub-exponential tail for large deviations. Parametrizing the deviation ϵ with a scalar variable c', we thus get the following bound : for any $\epsilon > 0$, there exists constants $C, c, c' > 0$ such that:

$$\mathbb{P}\left(\left|\frac{1}{d}\mathbf{g}^T\mathbf{g} - 1\right| \geqslant c'\epsilon\right) \leqslant Ce^{-cn\epsilon^2} \tag{8.167}$$

Since, in this case, we assume that the eigenvalues of the covariance matrices are bounded with probability one, multiplications by these matrices do not change these two previous rates. The remaining convergence rates that need to be determined are those of the Moreau envelopes. By assumption, the function g is separable, and pseudo-Lipschitz of order two. Moreover, the argument $\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}$ is an i.i.d. Gaussian random vector with finite variance. The Moreau envelope $\frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}\right)\right)$ is therefore a sum of pseudo-Lipschitz functions of order 2 of scalar Gaussian random variables. Using the concentration Lemma 31, we can find constants $C, c, \gamma > 0$ such that, for any $\epsilon > 0$, the following holds:

$$\mathbb{P}\left(\left|\alpha\frac{1}{n}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right) - \mathbb{E}\left[\alpha\frac{1}{n}\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right]\right| \geqslant \gamma\epsilon\right) \leqslant Ce^{-cn\epsilon^2} \tag{8.168}$$

For the second Moreau envelope, the argument $\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}\right)$ is not separable. If the regularization is a square, it is the concentration will reduce to that of the terms $\frac{1}{d}\mathbf{g}^T\mathbf{v}$ and $\frac{1}{d}\mathbf{g}^T\mathbf{g}$. If the regularization is a Lipschitz function, then the Moreau envelope is also Lipschitz from Lemma 26. Furthermore, since the eigenvalues of the covariance matrix Ω are bounded with probability one, the composition with the deterministic term $\nu\Omega^{1/2}\mathbf{v}$ does not change the Lipschitz property. Gaussian concentration of Lipschitz functions then gives an exponential decay independent of the magnitude of the deviation. Taking the loosest bound, which is the one obtained with the square penalty, we obtain that, for any $\epsilon > 0$, there exist constants $C, c, \gamma > 0$ such that the event

$$\left\{\left|\frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}\right)\right) - \mathbb{E}\left[\frac{1}{d}\mathcal{M}_{\frac{\eta}{\tau_2}\mathbf{F}(\Omega^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Omega^{-1/2}\tilde{\mathbf{v}} + \kappa\mathbf{g}\right)\right)\right]\right| \geqslant \gamma\epsilon\right\} \tag{8.169}$$

has probability at most $Ce^{-cn\epsilon^2}$. Combining these bounds gives the exponential rate for the convergence of $\mathcal{E}_n$ to $\mathcal{E}$ for any fixed $(\tau_1, \tau_2, \kappa, \nu, \eta, m)$. An ε-net argument can then be used to obtain the bound on the minmax values. □

The next lemma shows that the function E_n evaluated at $\tilde{\mathbf{w}}^*, \mathbf{z}^*$ is close to the optimal value $\mathcal{E}^*$.

Lemma 40. *For any $\epsilon > 0$, there exist constants C, c, γ such that:*

$$\mathbb{P}\left(|E_n(\tilde{\mathbf{w}}^*, \mathbf{z}^*) - \mathcal{E}^*| \geqslant \gamma\epsilon\right) \leqslant Ce^{-cn\epsilon^2} \tag{8.170}$$

Proof of Lemma 40: this Lemma can be proved in similar fashion to [204] Theorem B.1. using the strong convexity in $\tilde{\mathbf{w}}$ and $\mathbf{z}$ of $E_n(\tilde{\mathbf{w}}, \mathbf{z})$ along with Gordon's Lemma. We leave the detail of this part to a longer version of this paper.

Lemma 41. *For any $\epsilon > 0$, there exists constants $\gamma, c, C > 0$ such that the event*

$$\exists(\tilde{\mathbf{w}}, \mathbf{z}) \in \mathbb{R}^{n+d},\ \frac{1}{d}\min(\frac{\kappa_n^*}{2\tau_{1,n}^*}, \frac{\tau_{2,n}^*}{2\eta_n^*})\|(\tilde{\mathbf{w}}, \mathbf{z}) - (\tilde{\mathbf{w}}^*, \mathbf{z}^*)\|_2^2 > \epsilon \text{ and } \min_{\tilde{\mathbf{w}}, \mathbf{z}} E_n(\tilde{\mathbf{w}}, \mathbf{z}) \leqslant E_n(\tilde{\mathbf{w}}^*, \mathbf{z}^*) + \gamma\epsilon \tag{8.171}$$

has probability at most $\frac{C}{\epsilon}e^{-cn\epsilon^2}$.

This lemma can be proven using the same arguments as in [204] Appendix B, Theorem B.1. Intuitively, if two values of a strongly convex function are arbitrarily close, then the corresponding points are arbitrarily close. Note that we are normalizing the norm of a vector of size $(n+d)$ with d, which are proportional. This shows that any solution outside the ball centered around $\tilde{\mathbf{w}}^*, \mathbf{z}^*$ is sub-optimal. Now define the set:

$$D_{\tilde{\mathbf{w}}, \mathbf{z}, \epsilon} = \left\{\tilde{\mathbf{w}} \in \mathbb{R}^d, \mathbf{z} \in \mathbb{R}^n : \left|\phi_1(\frac{\tilde{\mathbf{w}}}{\sqrt{d}}) - \mathbb{E}\left[\phi_1\left(\frac{\tilde{\mathbf{w}}^*}{\sqrt{d}}\right)\right]\right| > \epsilon,\ \left|\phi_2(\frac{\mathbf{z}}{\sqrt{n}}) - \mathbb{E}\left[\phi_2\left(\frac{\mathbf{z}^*}{\sqrt{n}}\right)\right]\right| > \epsilon\right\} \tag{8.172}$$

where ϕ_1 is either a square or a Lipschitz function, and ϕ_2 is a separable, pseudo-Lipschitz function of order 2. Using the same arguments as in the proof of Lemma 40 and the assumptions on ϕ_1, ϕ_2, Gaussian concentration will give sub-exponential rates for the event $(\tilde{\mathbf{w}}^*, \mathbf{z}^*) \in D_{\tilde{\mathbf{w}}, \mathbf{z}, \epsilon}$. A similar argument to the proof of Lemma B.3 from [57] then shows that a distance of ϵ in $D_{\tilde{\mathbf{w}}, \mathbf{z}, \epsilon}$ results in a distance of ϵ^2 in the event (8.171), leading to the following result:

Lemma 42. *For any $\epsilon > 0$, there exists constants $\gamma, c, C > 0$ such that the event*

$$\exists(\tilde{\mathbf{w}}, \mathbf{z}) \in \mathbb{R}^{n+d}, (\tilde{\mathbf{w}}^*, \mathbf{z}^*) \in D_{\tilde{\mathbf{w}}, \mathbf{z}, \epsilon} \text{ and } \min_{\tilde{\mathbf{w}}, \mathbf{z}} E_n(\tilde{\mathbf{w}}, \mathbf{z}) \leqslant E_n(\tilde{\mathbf{w}}^*, \mathbf{z}^*) + \gamma\epsilon^2 \tag{8.173}$$

has probability at most $\frac{C}{\epsilon^2}e^{-cn\epsilon^4}$.

which proves Theorem 15 using the fact that $\hat{\mathbf{w}}, \hat{\mathbf{z}}$ are minimizers of the initial cost function. Theorem 14 is a consequence of Theorem 15.

If the restriction on f, g, ϕ_1, ϕ_2 are relaxed to any pseudo-Lipschitz functions of finite orders, the exponential rates involving them are lost and become linear following Lemma 29.

8.2.5 Relaxing the deterministic teacher assumption

The entirety of the previous proof has been done with a deterministic vector $\boldsymbol{\theta}_0$. Now, if $\boldsymbol{\theta}_0$ is assumed to be a random vector independent of all other quantities, as prescribed in the set of assumptions 8.1, we can "freeze" the variable $\boldsymbol{\theta}_0$ by conditioning on it. The whole proof can then be understood as studying the value of the cost conditioned on the value of $\boldsymbol{\theta}_0$. Note that, in the Gaussian case, correlations between the teacher and student are expressed through the covariance

matrices, thus leaving the possibility to parametrise the teacher with a vector $\boldsymbol{\theta}_0$ indeed independent of all the rest. To lift the conditioning in the end, one only needs to average out on the distribution of $\boldsymbol{\theta}_0$, the summability conditions of which are prescribed in the set of assumptions 8.1. Thus, random teacher vectors can be treated simply by taking an additional expectation in the expressions of Theorem 15, provided $\boldsymbol{\theta}_0$ is independent of the matrices A, B and the randomness in f_0.

As mentioned at the end of the previous section, the finite size rates will be determined by the assumptions made on the teacher vector and decay of the eigenvalues of the covariance matrices. We do not investigate in detail the limiting assumptions under which exponential rates still hold regarding the randomness of the teacher or tails of the eigenvalue distributions of covariance matrices.

8.2.6 The 'vanilla' teacher-student scenario

In this section, we give the explicit forms of the fixed points equations and optimal asymptotic estimators in the case where the teacher and the student are sampled from the same distribution, i.e. $\Omega = \Phi = \Psi = \Sigma$ where Σ is a positive definite matrix with sub-Gaussian eigenvalue decay. This setup was rigorously studied in [57] for the LASSO and heuristically in [129] for the ridge regularized logistic regression. In this case, the fixed point equations become

$$\tau_1 = \frac{1}{d}\mathbb{E}\left[\mathbf{g}^\top \text{prox}_{\frac{\eta}{\tau_2} f(\Sigma^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Sigma^{1/2}\boldsymbol{\theta}_0 + \kappa\mathbf{g}\right)\right)\right] \tag{8.174}$$

$$m\sqrt{\gamma} = \frac{1}{d}\mathbb{E}\left[\mathbf{v}^\top\Sigma^{-1/2}\text{prox}_{\frac{\eta}{\tau_2} f(\Sigma^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Sigma^{1/2}\boldsymbol{\theta}_0 + \kappa\mathbf{g}\right)\right)\right] \tag{8.175}$$

$$\tau_2 = \alpha\frac{\kappa}{\tau_1}\eta - \frac{\kappa\alpha}{\tau_1 n}\mathbb{E}\left[\mathbf{h}^\top \text{prox}_{\frac{\tau_1}{\kappa} g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right] \tag{8.176}$$

$$\eta^2 + \frac{m^2}{\rho} = \frac{1}{d}\mathbb{E}\left[\left\|\text{prox}_{\frac{\eta}{\tau_2} f(\Sigma^{-1/2}.)}\left(\frac{\eta}{\tau_2}\left(\nu\Sigma^{1/2}\boldsymbol{\theta}_0 + \kappa\mathbf{g}\right)\right)\right\|_2^2\right] \tag{8.177}$$

$$\nu\sqrt{\gamma} = \alpha\frac{\kappa}{n\tau_1}\mathbb{E}\left[(\frac{m}{\eta\rho}\mathbf{h} - \frac{\mathbf{s}}{\sqrt{\rho}})^\top \text{prox}_{\frac{\tau_1}{\kappa} g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right] \tag{8.178}$$

$$\tau_1^2 = \frac{\alpha}{n}\mathbb{E}\left[\left\|\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h} - \text{prox}_{\frac{\tau_1}{\kappa} g(.,y)}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right\|_2^2\right] \tag{8.179}$$

and the asymptotic optimal estimators read:

$$\mathbf{w}^* = \Sigma^{-1/2}\text{prox}_{\frac{\eta^*}{\tau_2^*} f(\Sigma^{-1/2}.)}(\frac{\eta^*}{\tau_2^*}(\nu^*\Sigma^{1/2}\boldsymbol{\theta}_0 + \kappa^*\mathbf{g})), \qquad \mathbf{z}^* = \text{prox}_{\frac{\tau_1^*}{\kappa^*} g(.,\mathbf{y})}\left(\frac{m^*}{\sqrt{\rho}}\mathbf{s} + \eta^*\mathbf{h}\right) \tag{8.180}$$

8.3 Equivalence with the replica prediction

In this Appendix, we show that the rigorous result of Theorem 15 can be used to prove the replica prediction in the case of a separable loss, a ridge penalty. For simplicity, we restrict ourselves to the case of random teacher weights with $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, \mathrm{I}_p)$. We provide an exact analytical matching between the replica prediction and the one obtained with Gordon's theorem. We start by an explicit derivation of the form presented in Corollary 11 from the main result (13.15).

8.3.1 Solution for separable loss and ridge regularization

Replacing $\mathbf{F}$ with a ridge penalty, we can go back to step (8.98) of the main proof and finish the calculation without inverting the matrix Ω. The assumption on the invertibility of Ω can thus be dropped in the case of ℓ_2 regularization. Letting $G = \left(\frac{\tau_2}{\eta}\Omega + \lambda_2 \mathbf{I}_d\right)^{-1}$, we get

$$\mathcal{E}(\tau_1, \tau_2, \kappa, \eta, \nu, m) = \frac{\kappa\tau_1}{2} - \frac{\eta\tau_2}{2} + m\nu\sqrt{\gamma} - \frac{\tau_2}{2\eta}\frac{m^2}{\rho} + \alpha\frac{1}{n}\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right] - \frac{1}{2d}\nu^2\boldsymbol{\theta}_0^\top\Phi G\Phi^\top\boldsymbol{\theta}_0 - \frac{1}{2d}\kappa^2\mathrm{Tr}\left(\Omega^{1/2}\mathrm{G}\Omega^{1/2}\right) \tag{8.181}$$

using Lemma 29 with a separable function, the expectation over the Moreau envelope converges to:

$$\frac{1}{n}\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right] = \mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa}g(.,y)}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right)\right] \tag{8.182}$$

where s and h are standard normal random variables and $y = f_0(\sqrt{\rho}s)$. The corresponding optimality conditions then reads:

$$\frac{\partial}{\partial\kappa} : \frac{\tau_1}{2} + \frac{1}{2\tau_1}\alpha\mathbb{E}\left[\left(\frac{m}{\sqrt{\rho}}s + \eta h - \mathrm{prox}_{\frac{\tau_1}{\kappa}g(.,y)}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right)\right)^2\right] - \kappa\frac{1}{d}\mathrm{Tr}\left(\Omega^{1/2}\mathrm{G}\Omega^{1/2}\right) = 0 \tag{8.183}$$

$$\frac{\partial}{\partial\nu} : m\sqrt{\gamma} - \frac{1}{d}\nu\boldsymbol{\theta}_0\Phi^\top G\Phi^\top\boldsymbol{\theta}_0 = 0 \tag{8.184}$$

$$\frac{\partial}{\partial\tau_2} : -\frac{\eta}{2} - \frac{m^2}{2\rho\eta} + \frac{1}{2}\frac{\nu^2}{\eta}\left(\Omega^{1/2}\Phi^\top\boldsymbol{\theta}_0\right)^\top G^2\Omega^{1/2}\Phi^\top\boldsymbol{\theta}_0 + \frac{\kappa^2}{2\eta}Tr\left(G^2\Omega^2\right) = 0 \tag{8.185}$$

$$\frac{\partial}{\partial m} : \nu\sqrt{\gamma} - \frac{\tau_2}{\rho\eta}m + \alpha\mathbb{E}\left[\frac{\kappa}{\tau_1}\frac{s}{\sqrt{\rho}}(\frac{m}{\sqrt{\rho}}s + \eta h - \mathrm{prox}_{\frac{\tau_1}{\kappa}g(.,y)}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right))\right] = 0 \tag{8.186}$$

$$\frac{\partial}{\partial\eta} : -\frac{\tau_2}{2} + \frac{\tau_2 m^2}{2\rho\eta^2} + \alpha\mathbb{E}\left[\frac{\kappa}{\tau_1}h\left(\frac{m}{\sqrt{\rho}}s + \eta h - \mathrm{prox}_{\frac{\tau_1}{\kappa}g(.,y)}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right)\right)\right] - \frac{1}{2}\frac{\tau_2\nu^2}{\eta^2}\left(\Omega^{1/2}\Phi^\top\boldsymbol{\theta}_0\right)^\top G^2\Omega^{1/2}\Phi^\top\boldsymbol{\theta}_0 - \frac{\tau_2\kappa^2}{2\eta^2}\mathrm{Tr}(\mathrm{G}^2\Omega^2) = 0 \tag{8.187}$$

$$\frac{\partial}{\partial\tau_1} : \frac{\kappa}{2} - \frac{\kappa}{2\tau_1^2}\alpha\mathbb{E}\left[\left(\frac{m}{\sqrt{\rho}}s + \eta h - \mathrm{prox}_{\frac{\tau_1}{\kappa}g(.,y)}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right)\right)^2\right] = 0 \tag{8.188}$$

simplifying these equations using Stein's lemma, we get:

$$\frac{\partial}{\partial \kappa} : \frac{\tau_1}{\kappa} = \frac{1}{d}\mathrm{Tr}\left(\Omega^{1/2}\left(\frac{\tau_2}{\eta}\Omega + \lambda_2 \mathbf{I}_{\mathrm{d}}\right)^{-1}\Omega^{1/2}\right) \tag{8.189}$$

$$\frac{\partial}{\partial \nu} : m\sqrt{\gamma} = \frac{1}{d}\nu\boldsymbol{\theta}_0\Phi\left(\frac{\tau_2}{\eta}\Omega + \lambda_2 \mathbf{I}_d\right)^{-1}\Phi^\top\boldsymbol{\theta}_0 \tag{8.190}$$

$$\frac{\partial}{\partial \tau_2} : \eta^2 + \frac{m^2}{\rho} = \frac{1}{d}\nu^2\left(\Omega^{1/2}\Phi^\top\boldsymbol{\theta}_0\right)^\top\left(\frac{\tau_2}{\eta}\Omega + \lambda_2 \mathbf{I}_d\right)^{-2}\left(\Omega^{1/2}\Phi^\top\boldsymbol{\theta}_0\right) + \frac{1}{d}\kappa^2\mathrm{Tr}(\left(\frac{\tau_2}{\eta}\Omega + \lambda_2 \mathbf{I}_{\mathrm{d}}\right)^{-2}\Omega^2) \tag{8.191}$$

$$\frac{\partial}{\partial m} : \nu\sqrt{\gamma} = \alpha\frac{\kappa}{\sqrt{\rho}\tau_1}\left(\mathbb{E}\left[s\mathrm{prox}_{\frac{\tau_1}{\kappa}g(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}} + \eta h\right)\right] - \frac{m}{\sqrt{\rho}}\mathbb{E}\left[\mathrm{prox}'_{\frac{\kappa}{\tau_1}g(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right)\right]\right) \tag{8.192}$$

$$\frac{\partial}{\partial \eta} : \frac{\tau_2}{\eta} = \alpha\frac{\kappa}{\tau_1}\left(1 - \mathbb{E}\left[\mathrm{prox}'_{\frac{\tau_1}{\kappa}g(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right)\right]\right) \tag{8.193}$$

$$\frac{\partial}{\partial \tau_1} : \kappa^2 = \left(\frac{\kappa}{\tau_1}\right)^2\alpha\mathbb{E}\left[\left(\frac{m}{\sqrt{\rho}}s + \eta h - \mathrm{prox}_{\frac{\tau_1}{\kappa}g(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}}s + \eta h\right)\right)^2\right] \tag{8.194}$$

8.3.2 Matching with Replica equations

In this section, we show that the fixed point equations obtained from the asymptotic optimality condition of the scalar minimization problem 11 match the ones obtained using the replica method. In what follows we will use the same notations as in [106], and an explicit, clear match with the notations from the proof of the main theorem will be shown. The replica computation, similar to the one from [106], leads to the following fixed point equations, in the replica notations:

$$V = \frac{1}{p}\mathrm{Tr}\left(\lambda\hat{V}I_p + \Omega\right)^{-1}\Omega \tag{8.195}$$

$$q = \frac{1}{p}\mathrm{Tr}\left[(\hat{q}\Omega + \hat{m}^2\Phi^\top\Phi)\Omega\left(\lambda\hat{V}I_p + \Omega\right)^{-2}\right] \tag{8.196}$$

$$m = \frac{1}{\sqrt{\gamma}}\frac{\hat{m}}{p}\mathrm{Tr}\left[\Phi^\top\Phi\left(\lambda\hat{V}I_p + \Omega\right)^{-1}\right] \tag{8.197}$$

$$\hat{V} = \alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\ \mathcal{Z}_y^0\left(y, \frac{m}{\sqrt{q}}, \rho - \frac{m^2}{q}\right)\partial_\omega f_g(y, \sqrt{q}\xi, V)\right] \tag{8.198}$$

$$\hat{q} = \alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\ \mathcal{Z}_y^0\left(y, \frac{m}{\sqrt{q}}, \rho - \frac{m^2}{q}\right)f_g(y, \sqrt{q}\xi, V)^2\right] \tag{8.199}$$

$$\hat{m} = \frac{\alpha}{\sqrt{\gamma}}\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\ \partial_\omega\mathcal{Z}_y^0\left(y, \frac{m}{\sqrt{q}}, \rho - \frac{m^2}{q}\right)f_g(y, \sqrt{q}\xi, V)\right] \tag{8.200}$$

where $f_g(y, \omega, V) = -\partial_\omega\mathcal{M}_{Vg(y,.)}(\omega)$ and $\mathcal{Z}_0$ is given by:

$$\mathcal{Z}_0\left(y, \omega, V\right) = \int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{1}{2V}(x-\omega)^2}\delta(y - f^0(x)). \tag{8.201}$$

In particular we have:

$$\partial_\omega\mathcal{Z}_0\left(y, \omega, V\right) = \int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{1}{2V}(x-\omega)^2}\left(\frac{x-\omega}{V}\right)\delta(y - f^0(x)) \tag{8.202}$$

To be explicit with the notation, let's open the equations up. Take for instance the one for $\hat{m}$. Opening all the integrals:

$$\begin{aligned}\hat{m} &= \int \frac{\mathrm{d}\xi}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} \int \mathrm{d}y \int \frac{\mathrm{d}x}{\sqrt{2\pi\left(\rho - m^2/q\right)}} e^{-\frac{1}{2}\frac{\left(x-\frac{m}{\sqrt{q}}\xi\right)^2}{\rho - m^2/q}} \left(\frac{x-\frac{m}{\sqrt{q}}\xi}{\rho - m^2/q}\right) f_g(y, \sqrt{q}\xi, V) \\ &\overset{(a)}{=} \int \frac{\mathrm{d}\xi}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} \int \frac{\mathrm{d}x}{\sqrt{2\pi\left(\rho - m^2/q\right)}} e^{-\frac{1}{2}\frac{\left(x-\frac{m}{\sqrt{q}}\xi\right)^2}{\rho - m^2/q}} \left(\frac{x-\frac{m}{\sqrt{q}}\xi}{\rho - m^2/q}\right) f_g(f_0(x), \sqrt{q}\xi, V)\end{aligned} \tag{8.203}$$

where in (a) we integrated over y explicitly. A direct comparison between the two sets of equations suggests the following mapping to navigate between the replica derivation and the proof using Gaussian comparison theorems. We denote replica quantities with *Rep* indices:

$$\begin{aligned} V_{Rep} &\iff \frac{\tau_1}{\kappa}, & \hat{V}_{Rep} &\iff \frac{\tau_2}{\eta}, & q_{Rep} &\iff \eta^2 + \frac{m^2}{\rho} \\ \hat{q}_{Rep} &\iff \kappa^2, & m_{Rep} &\iff m, & \hat{m}_{Rep} &\iff \nu \end{aligned} \tag{8.204}$$

with these notations, we get :

$$\frac{\partial}{\partial \kappa} : V = \frac{1}{d}\mathrm{Tr}((\hat{\mathrm{V}}\Omega + \lambda_2 \mathbf{I}_\mathrm{d})^{-1}\Omega) \tag{8.205}$$

$$\frac{\partial}{\partial \nu} : m = \frac{1}{\sqrt{\gamma}}\frac{\hat{m}}{d}\mathrm{Tr}((\hat{\mathrm{V}}\Omega + \lambda_2 \mathbf{I}_\mathrm{d})^{-1}\Phi^\top\Phi) \tag{8.206}$$

$$\frac{\partial}{\partial \tau_2} : q = \frac{1}{d}\mathrm{Tr}((\hat{\mathrm{q}}\Omega + \hat{\mathrm{m}}^2\Phi^\top\Phi)\Omega(\hat{\mathrm{V}}\Omega + \lambda_2 \mathbf{I}_\mathrm{d})^{-2}) \tag{8.207}$$

$$\begin{aligned}\frac{\partial}{\partial m} : \hat{m} = \frac{\alpha}{\sqrt{\gamma}}\frac{1}{V}\Bigg(&\mathbb{E}\left[\frac{s}{\sqrt{\rho}}\mathrm{prox}_{Vg(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}h\right)\right] \\ &- \frac{m}{\rho}\mathbb{E}\left[\mathrm{prox}'_{Vg(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}\ h\right)\right]\Bigg)\end{aligned} \tag{8.208}$$

$$\frac{\partial}{\partial \eta} : \hat{V} = \frac{\alpha}{V}\left(1 - \mathbb{E}\left[\mathrm{prox}'_{Vg(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}h\right)\right]\right) \tag{8.209}$$

$$\frac{\partial}{\partial \tau_1} : \hat{q} = \left(\frac{\alpha}{V^2}\right)\mathbb{E}\left[\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}\ h - \mathrm{prox}_{Vg(.,f_0(\sqrt{\rho}s))}\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}h\right)\right)^2\right] \tag{8.210}$$

The first three equations match the replica prediction, the last three can be exactly matched using the following change of variable and Gaussian integration:

$$\tilde{x} = \frac{x}{\sqrt{\rho}} \quad \tilde{\xi} = \left(\frac{\rho}{\rho - \frac{m^2}{q}}\right)^{1/2}\left(\frac{m}{\sqrt{q\rho}}\tilde{x} - \xi\right) \tag{8.211}$$

Chapter 9

Learning Gaussian mixtures with convex generalized linear models

The results in this chapter are based oon the publication [178]. Generalised linear models for multi-class classification problems are one of the fundamental building blocks of modern machine learning tasks. In this manuscript, we characterise the learning of a mixture of K Gaussians with generic means and covariances via empirical risk minimisation (ERM) with any convex loss and regularisation. In particular, we prove exact asymptotics characterising the ERM estimator in high-dimensions, extending several previous results about Gaussian mixture classification in the literature. We exemplify our result in two tasks of interest in statistical learning: a) classification for a mixture with sparse means, where we study the efficiency of ℓ_1 penalty with respect to ℓ_2; b) max-margin multi-class classification, where we characterise the phase transition on the existence of the multi-class logistic maximum likelihood estimator for $K > 2$. Finally, we discuss how our theory can be applied beyond the scope of synthetic data, showing that in different cases Gaussian mixtures capture closely the learning curve of classification tasks in real data sets.

9.1 Introduction

A recurring observation in modern deep learning practice is that neural networks often defy the standard wisdom of classical statistical theory. For instance, deep neural networks typically achieve good generalisation performances at a regime in which it interpolates the data, a fact at odds with the intuitive bias-variance trade-off picture stemming from classical theory [105, 126, 31]. Surprisingly, many of the "exotic" behaviours encountered in deep neural networks have recently been shown to be shared by models as simple as overparametrised linear classifiers [125, 34], e.g., the aforementioned benign over-fitting [21]. Therefore, understanding the generalisation properties of simple models in high-dimensions has proven to be a fertile ground for elucidating some of the challenging statistical questions posed by modern machine learning practice [190, 106, 112, 116, 118, 176, 169, 197, 53].

In this manuscript, we pursue this enterprise in the context of a commonly used model for high-dimensional classification problems: the Gaussian mixture. Indeed, it has been recently argued that the features learned by deep neural networks trained on the cross-entropy loss "collapse" in a mixture of well-separated clusters, with the last layer acting as a simple linear classifier [223]. Another observation put forward in [262] is that data obtained using generative adversarial networks behave as Gaussian mixtures. Here, we derive an exact asymptotic formula characterising the performance

of generalised linear classifiers trained on K Gaussian clusters with generic covariances and means. Our formula is valid for any convex loss and penalty, encompassing popular tasks in the machine learning literature such as ridge regression, basis pursuit, cross-entropy minimisation and max-margin estimation. This allow us to answer relevant questions for statistical learning, such as: what is the separability threshold for K-clustered data? How does regularisation affects estimation? Can different penalties help when the means are sparse? We also extend the observation of [262] showing that the learning curves of binary classification tasks on *real data* are indeed well captured by our asymptotic analysis.

Model definition We consider learning from a d-dimensional mixture of K Gaussian clusters $\mathcal{C}_{k\in[K]}$. The data set is obtained by sampling n pairs $(\mathbf{x}^\nu, \mathbf{y}^\nu)_{\nu\in[n]} \in \mathbb{R}^{d+K}$ identically and independently. We adopt the one-hot encoded representation of the labels, i.e., if $\mathbf{x}^\nu \in \mathcal{C}_k$, then $\mathbf{y}^\nu = \mathbf{e}_k$, kth basis vector of $\mathbb{R}^K$. We will denote the matrix of concatenated samples $\mathbf{X} \in \mathbb{R}^{d\times n}$. The mixture density then reads:

$$P(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} y_k \rho_k \mathcal{N}\left(\mathbf{x} \,|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{9.1}$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The matrix of concatenated means is denoted $\mathbf{M} \in \mathbb{R}^{d\times K}$. In Eq. (9.1), $\forall k$, $\rho_k = P(\mathbf{y} = \mathbf{e}_k) \geqslant 0$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ is positive-definite. We will consider the estimator obtained by minimising the following empirical risk:

$$\mathcal{R}(\mathbf{W}, \mathbf{b}) \equiv \sum_{\nu=1}^{n} \ell\left(\mathbf{y}^\nu, \frac{\mathbf{W}\mathbf{x}^\nu}{\sqrt{d}} + \mathbf{b}\right) + \lambda r(\mathbf{W}), \tag{9.2}$$

$$(\mathbf{W}^\star, \mathbf{b}^\star) \equiv \underset{\mathbf{W}\in\mathbb{R}^{K\times d},\, \mathbf{b}\in\mathbb{R}^K}{\arg\min} \mathcal{R}(\mathbf{W}, \mathbf{b}), \tag{9.3}$$

where $\mathbf{W} \in \mathbb{R}^{K\times d}$ and $\mathbf{b} \in \mathbb{R}^K$ are the weights and bias to be learned, ℓ is a convex loss function, and r is a regularisation function whose strength is tuned by the parameter $\lambda \geqslant 0$. For example the loss function ℓ can represent the composition of a cross-entropy loss with a softmax thresholding on the linear part of Eq. (9.2). We will characterise the distribution of the estimator $(\mathbf{W}^\star, \mathbf{b}^\star)$, and we will evaluate the average training loss defined as

$$\epsilon_\ell = \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(\mathbf{y}^\nu, \frac{\mathbf{W}^\star\mathbf{x}^\nu}{\sqrt{d}} + \mathbf{b}^\star\right), \tag{9.4}$$

as well as the average training error ϵ_t and generalisation error ϵ_g, defined as the misclassification rates:

$$\epsilon_t = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{I}\left[\mathbf{y}^\nu \neq \hat{\mathbf{y}}\left(\frac{\mathbf{W}^\star\mathbf{x}^\nu}{\sqrt{d}} + \mathbf{b}^\star\right)\right], \quad \epsilon_g = \mathbb{E}_{(\mathbf{x}^{\rm new},\mathbf{y}^{\rm new})}\left[\mathbb{I}\left[\mathbf{y}^{\rm new} \neq \hat{\mathbf{y}}\left(\frac{\mathbf{W}^\star\mathbf{x}^{\rm new}}{\sqrt{d}} + \mathbf{b}^\star\right)\right]\right],$$

where $(\mathbf{x}^{\rm new}, \mathbf{y}^{\rm new})$ is a new unseen data point sampled from the distribution in Eq. (9.1). In the previous equations, we have used the function $\hat{\mathbf{y}}\colon \mathbb{R}^K \to \mathbb{R}^K$, so that $\hat{y}_k(\mathbf{x}) \coloneqq \mathbb{I}(\max_\kappa x_\kappa = x_k)$.

The **main contributions** in this manuscript are the following:

(C1) In Sec. 9.2 and Chapter 10 we prove closed-form equations characterizing the asymptotic distribution of the matrix of weights $\mathbf{W}^\star \in \mathbb{R}^{K\times d}$, enabling the exact computation of key quantities

such as the training and generalisation error. Our proof method solves shortcomings of previous approaches by introducing a novel approximate message-passing sequence, building on recent advances in this framework, that is of independent interest.
(C2) In Sec. 9.3.1 we study the problem of classifying an anisotropic mixture with sparse means, where the strong or weak directions in the data are correlated with the non-zero components of the mean as in [81]. We study how learning the sparsity with an ℓ_1 penalty improves the classification performance.
(C3) In Sec. 9.3.2 we study the performance of the cross-entropy estimator in the limit of vanishing regularisation $\lambda \to 0^+$ for K Gaussian clusters as a function of the sample complexity $\alpha = n/d$; we show that a phase transition takes place at a certain value $\alpha_K^\star$ between a regime of complete separability of the data and a regime in which the correct classification of almost all points in the data set is not possible. We also investigate the effect of $\lambda > 0$ regularisation on the generalisation error, comparing the $K > 2$ case with the results given in the literature for $K = 2$ [197, 283].
(C4) In Sec. 9.3.3 we investigate the applicability of our formula beyond the Gaussian assumption by applying it to classification tasks on *real data*. We show that for different tasks and losses, it closely captures the real learning curves, even when data is mapped through a non-linear feature map. This further shows that Gaussian mixtures are a good surrogate model for investigating real classification tasks, as put forward in [262].

Relation to previous work The analysis of Gaussian mixture models in the high-dimensional regime has been the subject of many recent works. Exact asymptotics has been derived for the binary classification case with diagonal covariances in [73, 182] for the logistic loss and in [78, 145] for the square loss, both with ℓ_2 penalty. A similar analysis has been performed in [267] for the hard-margin SVM. These works were generalised to generic convex losses and ℓ_2 penalty in [197], where it has been also shown that the regularisation term can play an important role in reaching Bayes-optimal performances. Hinge regression with ℓ_1 penalty and diagonal covariance was treated in [169]. Recently, these asymptotic results were generalised to the case in which both clusters share the same covariance in [291], and finite rate bounds were given in [59, 54] in the case of sub-Gaussian mixtures. Asymptotic results for the multiclass problem with diagonal covariance were derived in [283] for the restricted case of the square loss with ℓ_2 penalty. Our result unifies all the aforementioned asymptotic formulas, and extends them to the general case of a multiclass problem with generic covariances and arbitrary convex losses and penalties.
From a technical standpoint, in [73, 252, 145, 283, 197, 169, 291] the authors use convex Gaussian comparison inequalities, see e.g. [281, 273], to prove their result. In particular, the proof given in [283] for the multiclass problem harnesses the geometry of least-squares, and it is then stressed that this method breaks down for multiclass problems in which the risk does not factorise over the K clusters (as for the cross-entropy, for example). We solve this problem using an innovative proof technique which has an interest in its own. Our approach is to capture the effect of non-linearity and generic covariances via the rigorous study of an approximate message-passing (AMP) sequence, a family of iterations that admit closed-form asymptotics at each step called *state evolution equations* [28]. Our proof relies on several refinements of AMP methods to handle the full complexity of the problem, notably spatial coupling with matrix valued variables [153, 80, 135] and non-separable update functions [37], via a multi-layer approach to AMP [188].
The sparse Gaussian mixture model analysed in Section 9.3.1 is closely related to the rare/weak features model introduced in [81] and widely studied in the context of sparse linear discriminant analysis [136, 264, 181, 167]. It was recently revisited in [54, 59] in the context of ERM with max-

margin classifiers. Here, we consider a correlated variation of the model and study the benefit of using a sparsity inducing ℓ_1 penalty.
The separability transition is a classical topic [66, 102] that has recently witnessed a renewal of interest thanks to its connection to overparametrization. It was studied in [53] in the context of uncorrelated Gaussian data, in [106] in the random features model and in [73, 197] for binary Gaussian mixtures.
Recently, [133, 45, 176] showed that the performance of different regression tasks on real data are well-captured by a teacher-student Gaussian model in high-dimensions for ridge regression, but this turned not to be true for non-linear problems such as logistic classification [176]. Authors of [262] showed instead that data from generative adversarial networks behave like Gaussian mixtures, motivating the modeling of such mixture for real-data in the present paper.

9.2 Technical results

Our main technical result is an exact asymptotic characterization of the distribution of the estimator $\mathbf{W}^\star$. Informally, the estimator $\mathbf{W}^\star$ and the quantity $\mathbf{W}^\star\mathbf{X}/\sqrt{d}$ behave asymptotically as non-linear transforms of multivariate Gaussian distributions. These transforms are directly linked to the proximal operators [224, 25] associated to the loss and regularisation functions, summarizing the effect of the cost function landscape on the estimator. The parameters of these Gaussian distributions and proximals can then be computed from the fixed point of a self-contained set of equations. We start by presenting the most generic form of our result in a concentration of measure-like statement in Theorem 17, and discuss an intuitive interpretation of the different quantities involved. Theorem 18 then states how the training and generalisation errors can be computed. All results presented in the experiments section can be obtained from Theorem 17. In Corollary 3 we discuss a particular case where explicit simplifications can be obtained. But first, let's summarise the required assumptions for our result to hold.

(A1) The functions ℓ (as a function of its second argument) and r are proper, closed, lower semi-continuous convex functions. We assume additionally that either the cost function $\ell(\mathbf{y}, \bullet\mathbf{X}) + r(\bullet)$ is strictly convex, or that $\ell(\mathbf{y}, \bullet)$ is strictly convex in its second argument and r is the ℓ_1 norm. We also assume that the cost function $\ell(\mathbf{y}, \bullet\mathbf{X}) + r(\bullet)$ is coercive.
(A2) The covariance matrices are positive definite and their spectral norms are bounded (with probability one).
(A3) The mean vectors $\boldsymbol{\mu}_k$ are distributed according to some density $P_{\boldsymbol{\mu}}(\mathbf{M})$ such that the following quantity is finite

$$\forall d \qquad \mathbb{E}\left[\left\|\mathbf{M}^\top\mathbf{M}\right\|_{\mathrm{F}}\right] < +\infty, \tag{9.5}$$

where $\|\bullet\|_{\mathrm{F}}$ denotes the Frobenius norm.
(A4) The number of samples n and dimension d both go to infinity with fixed ratio $\alpha = {}^{n}/_{d}$, called hereafter the sample complexity. The number of clusters K is finite.
(A5) The fixed point of the set of self-consistent equations Eq.(9.8) exists and is unique.

As specified by assumption **(A1)**, our proof does not apply to any convex problem. We discuss this assumption further in Appendix 10.4. We also comment on the existence and uniqueness of the solution to the set of self consistent equations Eq.(9.8) in Appendix 10.4. Before proceeding further, let us specify a useful notation. Suppose that the matrix $\mathbf{G} = (G_{ki})_{ki} \in \mathbb{R}^{K\times d}$ is given,

Theorem 17 (Concentration properties of the estimator)**.** *Let $\boldsymbol{\xi}_{k\in[K]} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ be collection of K-dimensional standard normal vectors independent of other quantities. Let also be $\{\boldsymbol{\Xi}_k\}$ a set of K matrices, $\boldsymbol{\Xi}_k \in \mathbb{R}^{K\times d}$, with i.i.d. standard normal entries, independent of other quantities. Under the set of assumptions (A1–A5), for any pseudo-Lispchitz functions of finite order $\phi_1 : \mathbb{R}^{K\times d} \to \mathbb{R}, \phi_2 : \mathbb{R}^{K\times n} \to \mathbb{R}$, the estimator $\mathbf{W}^\star$ and the matrix $\boldsymbol{Z}^\star = \frac{1}{\sqrt{d}}\mathbf{W}^\star\mathbf{X}$ verify:*

$$\phi_1(\mathbf{W}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\Xi}}\left[\phi_1(\mathbf{G})\right], \qquad \phi_2(\boldsymbol{Z}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\xi}}\left[\phi_2(\mathbf{H})\right], \tag{9.6}$$

where we have introduced the proximal for the loss:

$$\mathbf{h}_k = \mathbf{V}_k^{1/2} \underset{\ell(\mathbf{e}_k, \mathbf{V}_k^{1/2}\bullet)}{\mathrm{Prox}} (\mathbf{V}_k^{-1/2}\boldsymbol{\omega}_k) \in \mathbb{R}^K, \qquad \boldsymbol{\omega}_k \equiv \mathbf{M}_k + \mathbf{b} + \mathbf{Q}_k^{1/2}\boldsymbol{\xi}_k, \tag{9.7}$$

and $\boldsymbol{H} \in \mathbb{R}^{K\times n}$ is obtained by concatenating each $\mathbf{h}_k$, $\rho_k n$ times. We have also introduced the matrix proximal $\mathbf{G} \in \mathbb{R}^{K\times d}$:

$$\mathbf{G} = \mathbf{A}^{\frac{1}{2}} \odot \underset{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}{\mathrm{Prox}} (\mathbf{A}^{\frac{1}{2}} \odot \mathbf{B}), \qquad \mathbf{A}^{-1} \equiv \sum_k \hat{\mathbf{V}}_k \otimes \boldsymbol{\Sigma}_k, \quad \mathbf{B} \equiv \sum_k \left(\boldsymbol{\mu}_k \hat{\mathbf{m}}_k^\top + \boldsymbol{\Xi}_k \odot \sqrt{\hat{\mathbf{Q}}_k \otimes \boldsymbol{\Sigma}_k}\right).$$

The collection of parameters $(\mathbf{Q}_k, \mathbf{M}_k, \mathbf{V}_k, \hat{\mathbf{Q}}_k, \hat{\mathbf{m}}_k, \hat{\mathbf{V}}_k)_{k\in[K]}$ is given by the fixed point of the following self-consistent equations:

$$\begin{cases} \mathbf{Q}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}[\mathbf{G}\boldsymbol{\Sigma}_k\mathbf{G}^\top] \\ \mathbf{M}_k = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{\Xi}}[\mathbf{G}\boldsymbol{\mu}_k] \\ \mathbf{V}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}\left[\left(\mathbf{G} \odot \left(\hat{\mathbf{Q}}_k^{-\frac{1}{2}} \otimes \boldsymbol{\Sigma}_k^{\frac{1}{2}}\right)\right)\boldsymbol{\Xi}_k^\top\right] \end{cases} \qquad \begin{cases} \hat{\mathbf{Q}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\mathbf{f}_k\mathbf{f}_k^\top\right] \\ \hat{\mathbf{V}}_k = -\alpha\rho_k\mathbf{Q}_k^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{\xi}}\left[\mathbf{f}_k\boldsymbol{\xi}^\top\right] \\ \hat{\mathbf{m}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\mathbf{f}_k\right] \end{cases} \tag{9.8}$$

where $\mathbf{f}_k \equiv \mathbf{V}_k^{-1}(\mathbf{h}_k - \boldsymbol{\omega}_k)$, and the vector $\mathbf{b}^\star$ is such that $\sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\mathbf{V}_k\mathbf{f}_k\right] = \mathbf{0}$ holds.

The purpose of this statement is to have an asymptotically exact description of the distribution of the estimator, where the dimensions going to infinity are effectively summarized as averages over simple, independent distributions. Those distributions are parametrised by the set of finite-size parameters $(\mathbf{Q}_k, \mathbf{M}_k, \mathbf{V}_k, \hat{\mathbf{Q}}_k, \hat{\mathbf{m}}_k, \hat{\mathbf{V}}_k)_{k\in[K]}$ that can be exactly evaluated and have a clear interpretation. Indeed, the parameters $(\mathbf{M}_k, \hat{\mathbf{m}}_k)$ and $(\mathbf{Q}_k, \hat{\mathbf{Q}}_k)$ respectively represent means and covariances of multivariate Gaussians (combined with the original $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$), and the $(\mathbf{V}_k, \hat{\mathbf{V}}_k)$ parametrise the deformations that should be applied to these Gaussians to obtain the distribution of $\mathbf{W}^\star, \boldsymbol{Z}^\star$. The distribution is characterized in a weak sense with concentration of pseudo-Lipschitz (i.e., sufficiently regular) functions, whose definition is reminded in the Chapter 10. From this result one can work out a number of properties of the weights $\mathbf{W}^\star$, e.g., training and generalisation error, but also hypothesis tests as done in [57] for the LASSO. Due to the generality of the statement, no direct simplification is possible. However, we will see that in certain specific cases all quantities can be greatly simplified. This is notably the case for diagonal covariance matrices and separable estimators and observables ϕ_1, ϕ_2, where the sums over high-dimensional Gaussians concentrate explicitly to one-dimensional expectations. For instance the results of [283, 197] can be recovered as special cases of the present work. Theorem 17 then allows to obtain the asymptotic values of the generalisation error, of the training loss and of the training error. Their explicit expression is given in the following Theorem.

Theorem 18 (generalisation error and training loss)**.** *In the hypotheses of Theorem 17, the training loss, the training error and the generalisation error are given by*

$$\epsilon_\ell = \sum_{k=1}^{K} \rho_k \mathbb{E}_{\boldsymbol{\xi}}[\ell(\mathbf{e}_k, \mathbf{h}_k)], \qquad \epsilon_t = 1 - \sum_{k=1}^{K} \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\mathbf{h}_k)\right], \qquad \epsilon_g = 1 - \sum_{k=1}^{K} \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\boldsymbol{\omega}_k)\right]. \tag{9.9}$$

The case of ridge regularisation and diagonal $\boldsymbol{\Sigma}_k$ The general formulas given above can be remarkably simplified under some assumptions about the choice of the regularisation and about the structure of the covariance matrices $\boldsymbol{\Sigma}_k$. This is the case for instance for the ridge regularisation $r(\mathbf{W}) = \|\mathbf{W}\|_{\mathrm{F}}^2/2$ and jointly diagonalizable covariances. In this case, Theorem 17 simplifies as follows.

Corollary 3. *Under the hypotheses of Theorem 17, let us further assume that a ridge regularisation is adopted, $r(\mathbf{W}) = \|\mathbf{W}\|_{\mathrm{F}}^2/2$, and that the covariance matrices $\boldsymbol{\Sigma}_k$ have a common set of orthonormal eigenvectors $\{\mathbf{v}_i\}_{i=1}^d$, so that, for each $\boldsymbol{\Sigma}_k = \sum_{i=1}^d \sigma_i^k \mathbf{v}_i \mathbf{v}_i^\top$. Let us also introduce, in the $d \to +\infty$ limit, the joint distribution for the K-dimensional vectors $\boldsymbol{\sigma} = (\sigma^1, \dots, \sigma^K)$ and $\boldsymbol{\mu} = (\mu^1, \dots, \mu^K)$,*

$$\frac{1}{d}\sum_{i=1}^{d}\prod_{k=1}^{K} \delta(\sigma^k - \sigma_i^k)\delta(\mu^k - \sqrt{d}\boldsymbol{\mu}_k^\top \mathbf{v}_i) \xrightarrow{d\to+\infty} p(\boldsymbol{\sigma}, \boldsymbol{\mu}), \tag{9.10}$$

Then, the first three saddle point equations in eqs. (9.8) take the form

$$\begin{cases} \mathbf{Q}_k = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}\left[\sigma^k \left(\lambda \mathbf{I}_K + \sum_{\kappa=1}^K \sigma^\kappa \hat{\mathbf{V}}_k\right)^{-2} \left(\sum_{\kappa\kappa'} \mu^\kappa \mu^{\kappa'} \hat{\mathbf{m}}_\kappa \hat{\mathbf{m}}_{\kappa'}^\top + \sum_{\kappa=1}^K \sigma^\kappa \hat{\mathbf{Q}}_k\right)\right], \\ \mathbf{M}_k = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}\left[\mu^k \left(\lambda \mathbf{I}_K + \sum_{\kappa=1}^K \sigma^\kappa \hat{\mathbf{V}}_k\right)^{-1} \sum_{\kappa=1}^K \mu^\kappa \hat{\mathbf{m}}_\kappa\right], \\ \mathbf{V}_k = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}\left[\sigma^k \left(\lambda \mathbf{I}_K + \sum_{\kappa=1}^K \sigma^\kappa \hat{\mathbf{V}}_k\right)^{-1}\right]. \end{cases} \tag{9.11}$$

Narrative of the proof The proof is detailed in Chapter 10. It overcomes problems that existing methods, notably convex Gaussian comparison inequalities [283], have yet to be adapted to. The first main technical difficulty resides in the estimator of interest being a matrix learned with non-linear functions. This makes it impossible to decompose the problem on each row of the estimator, which must be characterized in a probabilistic sense directly as a matrix. The second main difficulty is brought by the mixture of arbitrary covariances. Intuitively, the covariances correlate the estimator with the individual clusters, and therefore the correlation function cannot be represented by a single quantity. In our proof, these points are handled using the AMP and related state-evolution techniques [42, 28, 29, 109]. The main idea of the proof is to express the estimator $\mathbf{W}^\star$ as the limit of a convergent sequence whose structure enables the decomposition of all correlations and distributions in closed form. AMP iterations can handle matrix valued variables [15, 135], correlations in block-structure [135], non-separable functions [188, 37] and compositions of the previous three, leaving a large choice of possibilities in their design. We thus reformulate the problem in a way that makes the interaction between the estimator and each cluster explicit, effectively introducing a block structure to the problem, and isolate the overlaps with the means $\{\boldsymbol{\mu}_k\}$. We then design a matrix-valued sequence that obeys the update rule of an AMP sequence, in order to benefit from its exact asymptotics, and whose fixed point condition matches the optimality condition of the ERM problem, Eq. (9.2). Our proof builds on the spatial coupling framework in the AMP literature [154, 135], which shows that the effect of random matrices defined with non-identically distributed

blocks can be embedded in an AMP iteration while explicitly keeping the effect of each block. The non-linearities are then obtained by a block decomposition of the proximal operators defined on sets of matrices, acting on different variables of the AMP sequence and representing the effect of each cluster. The convergence analysis is made possible by the convexity of the problem: the sequence is defined with proximal operators of convex functions which are roughly contractions, and results in converging sequences when combined with the high-dimensional properties of the iteration. It is also interesting to note that the replica method, although heuristic, yet again gives the correct prediction without any hindering from the aforementioned main difficulties, as detailed in the Appendix of the original paper.

Universality AMP-type proofs are amenable to both finite sample size analysis and universality proofs. For instance, in [251] it is shown that simpler instances of AMP for the LASSO exhibit exponential concentration in the system size, and the i.i.d. Gaussian assumption can be relaxed to independently sampled sub-Gaussian distributions, as shown in [27, 62]. Although these results do not formally encompass our case, their proof method contains most of the required technicalities, and it should be possible to prove similar results in the present setting. Indeed, recent results in [262] suggest that the formula of Theorem 17 and 18 should be universal for all mixtures of concentrated distribution in high-dimension, not only Gaussian ones. As we discuss Sec. 9.3.3, even real data learning curves are empirically found to follow the behavior of the mixture of Gaussians.

9.3 Results on synthetic and real datasets

In this section we exemplify how Theorem 17 can be employed to compute quantities of interest in different empirical risk minimisation tasks in high-dimensions. In all cases discussed below, eqs. (9.8) have been solved numerically. A repository with a polished version of the code we used to solve the equations is available on GitHub.

9.3.1 Correlated sparse mixtures

As a first example, consider a binary classification problem in which the most relevant features live in a subspace of $\mathbb{R}^d$, and can be either weaker or stronger with respect to the irrelevant features. This problem can be modelled with a Gaussian mixture model with sparse means, and where the strong/weak directions of the covariance matrix are correlated with the non-zero components of the means. Mathematically, we consider a data set with n independent samples $(\mathbf{x}^\nu, y^\nu) \in \mathbb{R}^d \times \{-1, 1\}$ drawn from a Gaussian mixture $\mathbf{x}^\nu \sim \mathcal{N}(y^\nu \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with diagonal covariance $\Sigma_{ij} = \sigma_i \delta_{ij}$ which is correlated with the sparse means:

$$P(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{i=1}^{d} \{\rho \mathcal{N}(\mu_i|0, 1)\delta_{\sigma_i,\Delta_1} + (1-\rho)\delta_{\mu_i,0}\delta_{\sigma_i,\Delta_2}\} \tag{9.12}$$

where $\rho > 0$ is the fraction of non-zero entries in $\boldsymbol{\mu}$. This model is closely related to the rare/weak features model introduced by Donoho and Jin in [81]. Indeed, in the case $\Delta_1 = \Delta_2 \equiv \Delta$ the signal-to-noise ratio of the model is proportional to $\rho/\sqrt{\Delta}$, with ρ and $\Delta^{-1/2}$ playing the roles of the parameters ϵ and μ_0 setting the "rareness" and "strength" of the features in [81].

The formulas given in Theorem 17 simplify considerably for this model (see Appendix of the original paper), and therefore can be readily used to characterise the learning performance of different losses and penalties. For instance, one fundamental question we can address is when learning

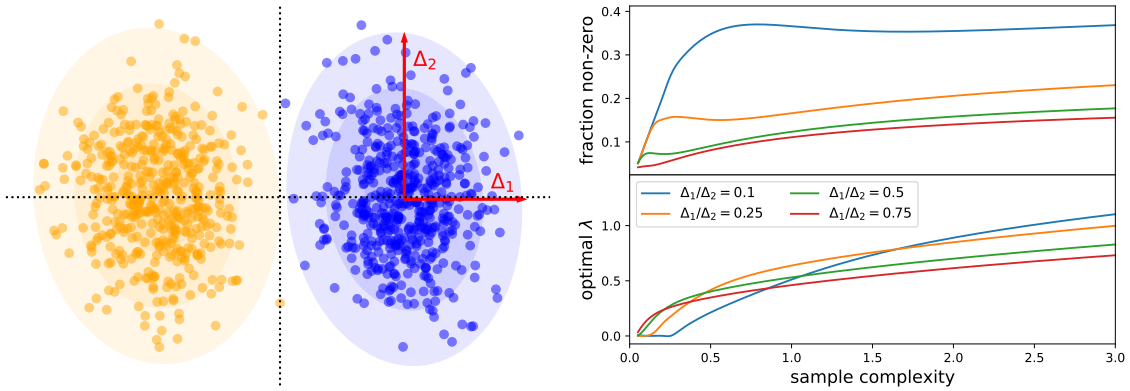


Figure 9.1: (**Left**) Two-dimensional projection of the Gaussian mixture introduced via Eq. (9.12) in which the sparse directions of the means are correlated with the weak/strong directions in the data. (**Right**) Fraction of non-zero elements of the lasso estimator (*top*) and optimal regularisation strength (*bottom*) as a function of the sample complexity $\alpha = n/d$ for different anisotropy ratios and fixed sparsity $\rho = 0.1$. Note that for $\Delta_1/\Delta_2 \lesssim 1$ and for low α the optimal error is achieved for vanishing regularisation, which corresponds to the *basis pursuit* algorithm [61].

a sparse solution with the ℓ_1 regularization is advantageous over the usual ℓ_2. Figure 9.2 compares the learning curves computed from Theorem 17 for the lasso and ridge estimators, with optimal regularisation strength $\lambda^\star(\alpha) = \text{argmin}\ \epsilon_g(\alpha, \lambda)$ at fixed sparsity $\rho = 0.1$. We can see that lasso performs considerably better than ridge in the regime where $\Delta_1/\Delta_2 \lesssim 1$, while it achieves a similar performance when $\Delta_1/\Delta_2 \gtrsim 1$. This is quite intuitive: the sparse directions are uninformative, and therefore learning the relevant features is better when they are stronger. Figure 9.1 (right) shows how the sparsity of the learned estimator $\mathbf{W}^\star$ and the optimal regularisation $\lambda^\star$ depends on the sample complexity $\alpha = n/d$. Interestingly, for $\Delta_1/\Delta_2 = 0.1$ or lower there is a region of small α in which basis pursuit ($\lambda = 0^+$) [61] is optimal, and the sparsity of the estimator has a curious non-monotonic behaviour with α.

9.3.2 Separability transition for the cross-entropy loss

We now consider the problem of classifying points of K Gaussian clusters using a cross-entropy loss

$$\ell(\mathbf{y}, \mathbf{x}) = -\sum_{k=1}^{K} y_k \ln \frac{e^{x_k}}{\sum_{\kappa=1}^{K} e^{x_\kappa}}. \tag{9.13}$$

Using the results of Theorem 18, we estimate the dependence of the generalisation error ϵ_g on the sample complexity α and on the regularisation λ. We assume Gaussian means $\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d/d)$ and diagonal covariances $\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma} = \Delta \mathbf{I}_d$. Finally, we adopt a ridge penalty, $r(\mathbf{W}) \equiv \|\mathbf{W}\|_{\text{F}}^2/2$, and we focus on the case of balanced clusters, i.e., $\rho_k = 1/K$ for the sake of simplicity.

Separability transition In Fig. 9.3 (left top) we plot the generalisation error ϵ_g as function of α for $2 \leqslant K \leqslant 5$ and $\lambda = 10^{-4}$. The smooth curve is obtained solving the fixed point equations in Theorem 17 and plugging the results in the formulas in Theorem 18. The results of numerical experiments are obtained averaging over 10^2 instances of the problem solved using the `LogisticRegression` module in the `Scikit-learn` package [229]. An excellent agreement is

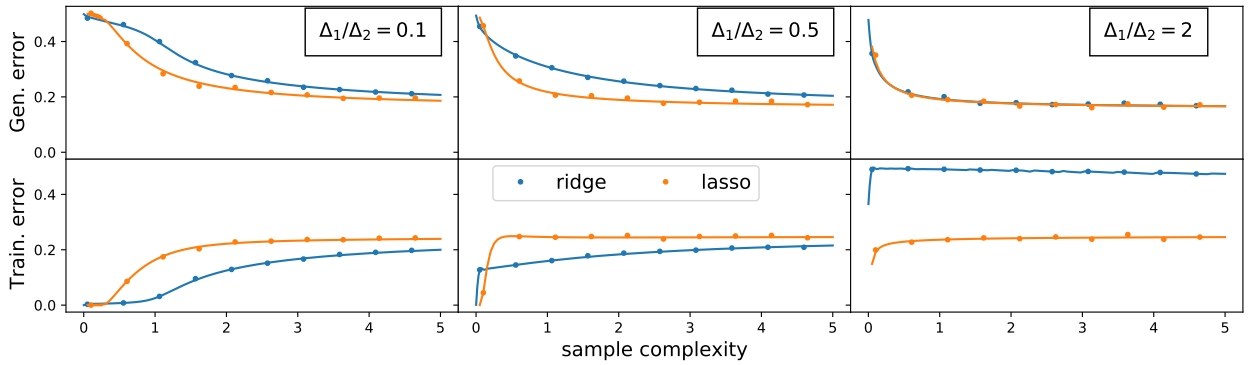


Figure 9.2: Learning curves for the sparse mixture model defined via Eq. (9.12) at fixed sparsity $\rho = 0.1$, comparing the performance of the ridge (blue) and the lasso (orange) estimators at optimal regularisation strength λ^* and for different anisotropy ratio Δ_1/Δ_2 (here $\Delta_1 = 0.1$ and we vary Δ_2). Full lines denote the theoretical prediction, and dots denote finite instance simulations with $d = 1000$ using the `ElasticNet` module in the `Scikit-learn` package [229]. Above a certain sample complexity α, we can identify two regimes: a) a $\Delta_1/\Delta_2 \lesssim 1$ regime in which the ℓ_1 penalty improves significantly over ℓ_2; b) a $\Delta_1/\Delta_2 \gtrsim 1$ regime in which the performance is similar. Interestingly, even though the generalisation error of lasso is considerably better in a), the training loss (i.e. the mse on the labels) is higher, & vice-versa in b).

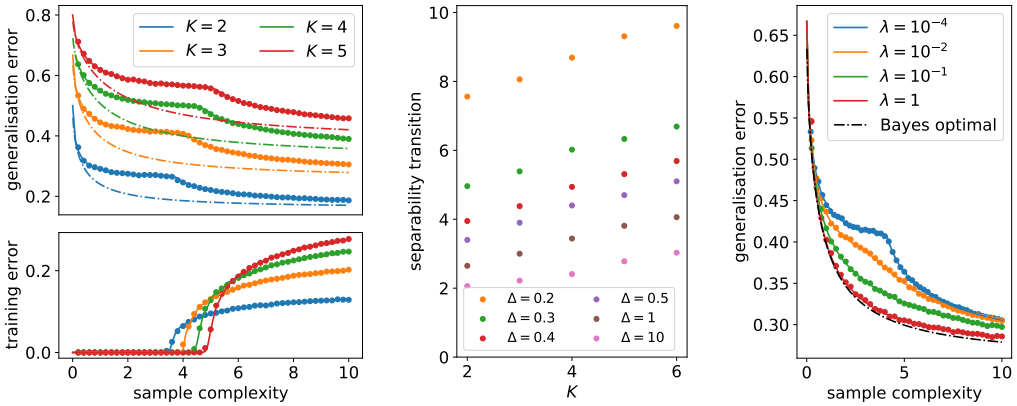


Figure 9.3: Classification of K Gaussian clusters in d dimensions, having Gaussian means and $\mathbf{\Sigma}_k \equiv \mathbf{\Sigma} = \Delta \mathbf{I}_d$ with $\Delta = 1/2$. In all presented cases, a quadratic regularisation has been adopted. Numerical experiments have been performed using $d = 10^3$. (**Left**) Generalisation error ϵ_g (*top*) and training error ϵ_t (*bottom*) as function of α at $\lambda = 10^{-4}$. Theoretical predictions (full lines) are compared with the results of numerical experiments (dots). Dash-dotted lines of the corresponding color represent, for comparison, the Bayes-optimal error. The results of numerical experiments are in agreement with the theoretical predictions in all cases. (**Center**) Separability transition $\alpha_K^\star$ as a function of K in the same setting for different values of Δ. (**Right**) Dependence of the generalisation error on the regularization λ for $K = 3$ and $\Delta = 1/2$ in the balanced case, $\rho_k = 1/K$.

observed. For each pair (K, Δ) and for vanishing regularisation $\lambda \to 0^+$ we observe a double-descent-like behaviour in the generalisation error. Indeed, the cusp $\alpha_K^\star(\Delta)$ in the generalisation error corresponds to the point in which the cross-entropy estimator ceases to perfectly interpolate the data, revealing the existence of a separability transition of the type discussed in [53] for Gaussian i.i.d. data. As stressed therein, a phase of perfect separability of the data points corresponds to a regime in which the maximum-likelihood estimate does not exist with probability one. This is visible, in the same figure (left bottom), from the training error ϵ_t that is identically zero for $\alpha < \alpha_K^\star$, and strictly positive otherwise. Our result extends the observations in [73, 197], where an analytic expression for $\alpha_2^\star$ has been given in the case of for $K = 2$, $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ Gaussian vector, generalising the classical result of Cover [66]. The separability transition point $\alpha_K^\star$ decreases with Δ and increases with K, showing that for larger K it is easier to separate the different clusters: this intuitively follows from the fact that, at fixed α and Δ, each cluster is given by $\alpha d/K$ points, i.e., fewer for increasing K and therefore easier to classify, see Fig. 9.3 (center).

The role of regularisation In Fig. 9.3 (right) we compare the performances of the cross-entropy loss with respect to the Bayes-optimal error (detailed in the appendix of the original paper) for different strength λ of the regularisation, assuming all identical diagonal covariances $\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma} = \Delta \mathbf{I}_d$. In the case of balanced clusters (i.e., $\rho_k = 1/K$ for all k) it is observed that the generalisation error approaches the Bayes-optimal error for $\lambda \to +\infty$. The same phenomenology has been observed in [78, 197] in the $K = 2$ case with opposite means and generic loss, and in [283] for $K > 2$ for the square loss. Using the concentration results of Section 9.2, we investigated the robustness of this result in the case of balanced clusters but with different covariances and various losses. First, we considered two opposite *balanced* clusters with $\boldsymbol{\Sigma}_1 = \Delta_1 \mathbf{I}_d$ and $\boldsymbol{\Sigma}_2 = \Delta_2 \mathbf{I}_2$, $\Delta_1 \neq \Delta_2$, and we estimated the generalisation error at fixed sample complexity as function of $\lambda \in [10^{-4}, 10^2]$ using ridge regression. As shown in Fig. 9.4 (left), the regularisation strength optimising the error is finite, and in particular depends on the sample complexity. This situation is closer to what is observed in real problems with balanced data analysed using logistic regression. Indeed, using the covariances from real data sets such as MNIST or Fashion-MNIST yields a similar behaviour, see Fig. 9.4 (right), with an optimal λ that is found to be finite.

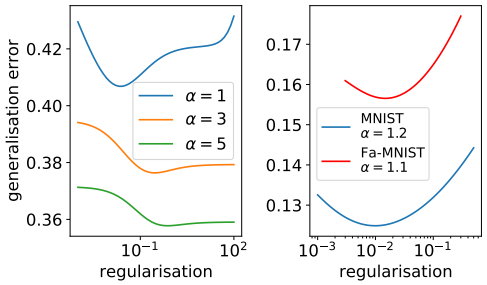


Figure 9.4: (**Left.**) Generalisation error obtained using ridge regression in the case of two balanced Gaussian clusters having $\boldsymbol{\Sigma}_1 = \frac{1}{10}\mathbf{I}_d$ and $\boldsymbol{\Sigma}_2 = \frac{1}{100}\mathbf{I}_d$ as function of λ for different values of the sample complexity α. (**Right**) Generalisation error ϵ_g as a function of λ at fixed α in the binary classification of MNIST and in the FashionMNIST via logistic regression (see Sec. 9.3.3 for details).

9.3.3 Binary classification with real data

A recent line of works has reported that the asymptotic learning curves of simple regression tasks on real data sets can be well approximated by a surrogate Gaussian model matching the first two moments of the data [45, 133, 176]. However, this analysis was fundamentally restricted to least-squares regression, and considerable deviation from the Gaussian model was observed for

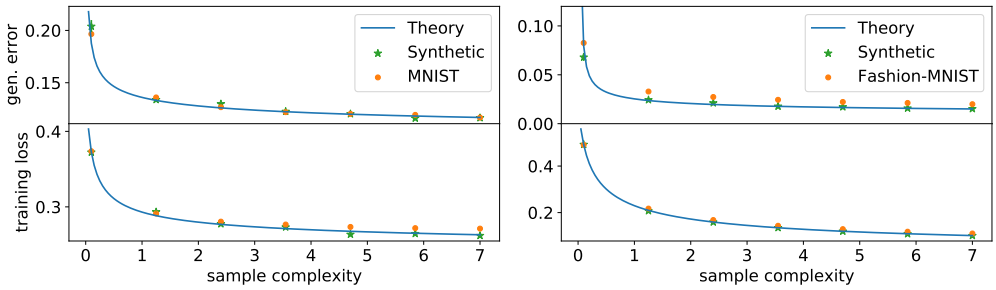


Figure 9.5: Generalisation error and training loss for the binary classification using the logistic loss on MNIST with $\lambda = 0.05$ (**left**) and on Fashion-MNIST with $\lambda = 1$ (**right**). The results are compared with synthetic data produced from the corresponding Gaussian mixture, and the theoretical prediction.

classification tasks [176]. Authors of [262] have shown that realistic-looking data from trained generative adversarial networks behave like Gaussian mixtures. Here, we pursue these observations and investigate whether Theorem 18 can be used to capture the learning curves of classification tasks on two popular data sets: MNIST [158] and Fashion-MNIST [298]. Our goal is to compare the performances of some classification tasks on them with the predictions provided by the theory for the Gaussian mixture model.

Both data sets consist of $n_{\text{tot}} = 7 \times 10^4$ images $\hat{\mathbf{x}}^\mu \in \mathbb{R}^d$, $d = 784$. Each image $\hat{\mathbf{x}}^\mu$ is associated to a label $\hat{y}^\mu = \{0, 1, \dots, 9\}$ specifying the type of represented digit (in the case of MNIST) or item (in the case of Fashion-MNIST). In both cases, we divided the database into two balanced classes (even vs odd digits for MNIST, clothes vs accessories for Fashion-MNIST), relabelling the elements $\hat{\mathbf{x}}^\mu$ with $y^\mu \in \{-1, 1\}$ depending on their class, and we selected $n < n_{\text{tot}}$ elements to perform the training, leaving the others for the test of the performances. We adopted a logistic loss with ℓ_2 regularisation. First, we performed logistic regression on the training real data set, then we tested the learned estimators on the remaining $n_{\text{tot}} - n$ images. At the same time, for each class k of the training set, we empirically estimated the corresponding mean $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$. We then assumed that the classification problem on the real database corresponds to a Gaussian mixture model of $K = 2$ clusters with means $\{\boldsymbol{\mu}_k\}_{k \in [2]}$ and covariances $\{\boldsymbol{\Sigma}_k\}_{k \in [2]}$. Under this assumption, we computed the generalisation error and the training loss predicted by the theory inserting the empirical means and covariances in our general formulas. The results are given in Fig. 9.5, showing a good agreement between the theoretical prediction and the results obtained on MNIST and Fashion-MNIST. In Fig. 9.5 we also plot, as reference, the results of a classification task performed on synthetic data, obtained generating a genuine Gaussian mixture with the means and covariances of the real data set.

Interestingly, this construction can also be used to analyse the learning curves of classification problems with non-linear feature maps [176], e.g. random features [238]. In this case, we first apply to our data set a feature map $\mathbf{x}^\mu = \text{erf}(\mathbf{F}\hat{\mathbf{x}}^\mu)$, where $\mathbf{F} \in \mathbb{R}^{p \times d}$ has i.i.d. Gaussian entries and the erf function is applied component wise. The classification task is then performed on the new data set $\{(\mathbf{x}^\nu, y^\nu)\}_{\nu \in [n]}$, the new data points $\mathbf{x}^\nu$ living in a p-dimensional space. We denote $\gamma = p/d$. We repeat the analysis described above in this new setting. Our results are in Fig. 9.6 for different values of γ. Once again, the generalisation error and the training loss are shown to be in a good agreement with both the theoretical prediction and the synthetic data sets obtained plugging in our formulas the real data means and the real data covariance matrices.

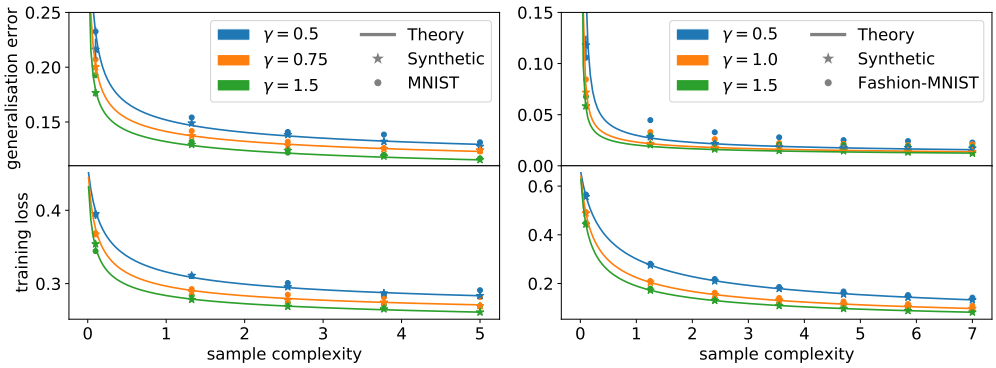


Figure 9.6: Generalisation error and training loss for the binary classification using the logistic on MNIST at $\lambda = 0.05$ (**left**) and on Fashion-MNIST at $\lambda = 1$ (**right**) in the random feature setting, for different values of γ, ratio between the number of parameters and the dimensionality of the data. The results are compared with synthetic data produced with the same γ, and the theoretical prediction.

Chapter 10

Proofs for the Gaussian mixture

This appendix presents the proof of the main technical result, Theorem 17. Throughout the whole proof, we assume that the set of conditions from Sec. 9.2 is verified.

10.1 Required background

In this Section, we give an overview of the main concepts and tools on approximate message passing algorithms which will be required for the proof.

We start with some definitions that commonly appear in the approximate message-passing literature, see e.g. [28, 135, 37]. The main regularity class of functions we will use is that of pseudo-Lipschitz functions, which roughly amounts to functions with polynomially bounded first derivatives. We include the required scaling w.r.t. the dimensions in the definition for convenience.

Definition 15 (Pseudo-Lipschitz function)**.** *For $k, K \in \mathbb{N}^*$ and any $n, m \in \mathbb{N}^*$, a function $\boldsymbol{\phi}\colon \mathbb{R}^{n\times K} \to \mathbb{R}^{m\times K}$ is called a* pseudo-Lipschitz of order k *if there exists a constant $L(k, K)$ such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n\times K}$,*

$$\frac{\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y})\|_{\mathrm{F}}}{\sqrt{m}} \leqslant L(k, K)\left(1 + \left(\frac{\|\mathbf{x}\|_{\mathrm{F}}}{\sqrt{n}}\right)^{k-1} + \left(\frac{\|\mathbf{y}\|_{\mathrm{F}}}{\sqrt{n}}\right)^{k-1}\right)\frac{\|\mathbf{x} - \mathbf{y}\|_{\mathrm{F}}}{\sqrt{n}} \tag{10.1}$$

where $\|\bullet\|_{\mathrm{F}}$ denotes the Frobenius norm. Since K will be kept finite, it can be absorbed in any of the constants.

For example, the function $f : \mathbb{R}^n \to \mathbb{R}, \mathbf{x} \mapsto \frac{1}{n}\|\mathbf{x}\|_2^2$ is pseudo-Lipshitz of order 2.

Moreau envelopes and Bregman proximal operators — In our proof, we will also frequently use the notions of Moreau envelopes and proximal operators, see e.g. [224, 25]. These elements of convex analysis are often encountered in recent works on high-dimensional asymptotics of convex problems, and more detailed analysis of their properties can be found for example in [281, 176]. For the sake of brevity, we will only sketch the main properties of such mathematical objects, referring to the cited literature for further details. In this proof, we will mainly use proximal operators acting on sets of real matrices endowed with their canonical scalar product. Furthermore, proximals will be defined with matrix valued parameters in the following way: for a given convex function $f\colon \mathbb{R}^{d\times K} \to \mathbb{R}$, a given matrix $\mathbf{X} \in \mathbb{R}^{d\times K}$ and a given symmetric positive definite matrix

$\mathbf{V} \in \mathbb{R}^{K\times K}$ with bounded spectral norm, we will consider operators of the type

$$\arg\min_{\mathbf{T}\in\mathbb{R}^{d\times K}} \left\{ f(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\mathbf{X})\mathbf{V}^{-1}(\mathbf{T}-\mathbf{X})^\top\right)\right\} \tag{10.2}$$

This operator can either be written as a standard proximal operator by factoring the matrix $\mathbf{V}^{-1}$ in the arguments of the trace:

$$\operatorname*{Prox}_{f(\bullet\mathbf{V}^{1/2})}(\mathbf{X}\mathbf{V}^{-1/2})\mathbf{V}^{1/2} \in \mathbb{R}^{d\times K} \tag{10.3}$$

or as a Bregman proximal operator [24] defined with the Bregman distance induced by the strictly convex, coercive function (for positive definite $\mathbf{V}$)

$$\mathbf{X} \mapsto \frac{1}{2}\mathrm{tr}(\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^\top) \tag{10.4}$$

which justifies the use of the Bregman resolvent

$$\arg\min_{\mathbf{T}\in\mathbb{R}^{d\times K}} \left\{ f(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\mathbf{X})\mathbf{V}^{-1}(\mathbf{T}-\mathbf{X})^\top\right)\right\} = (\mathrm{Id} + \partial f(\bullet)\mathbf{V})^{-1}(\mathbf{X}) \tag{10.5}$$

Many of the usual or similar properties to that of standard proximal operators (i.e. firm non-expansiveness, link with Moreau/Bregman envelopes,...) hold for Bregman proximal operators defined with the function (10.4), see e.g. [24, 26]. In particular, we will be using the equivalent notion to firmly nonexpansive operators for Bregman proximity operators, called *D-firm* operators. Consider the Bregman proximal defined with a differentiable, strictly convex, coercive function $g : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a given input Hilbert space. Let T be the associated Bregman proximal of a given convex function $f : \mathcal{X} \to \mathbb{R}$, i.e., for any $\mathbf{x} \in \mathcal{X}$

$$T(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathcal{X}} \{f(\mathbf{x}) + D_g(\mathbf{x},\mathbf{y})\} \tag{10.6}$$

Then T is *D-firm*, meaning it verifies

$$\langle T\mathbf{x} - T\mathbf{y}, \nabla g(T\mathbf{x}) - \nabla g(T\mathbf{y})\rangle \leqslant \langle T\mathbf{x} - T\mathbf{y}, \nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\rangle \tag{10.7}$$

for any $\mathbf{x}, \mathbf{y}$ in $\mathcal{X}$.

Gaussian concentration — Gaussian concentration properties are at the root of this proof. Such properties are reviewed in more detail, for example, in [37, 176].

Notations — For any set of matrices $\{\mathbf{A}_k \in \mathbb{R}^{n_k\times d_k}\}_{k\in[K]}$ we will use the following notation:

$$\begin{bmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & (*) & \\ & (*) & \ddots & \\ & & & \mathbf{A}_K \end{bmatrix} \equiv [\mathbf{A}_k]_{k=1}^K \in \mathbb{R}^{(\sum_{k=1}^K n_k)\times(\sum_{k=1}^K d_k)} \tag{10.8}$$

where the terms denoted by $(*)$ will be zero most of the time.
For a given function $\boldsymbol{\phi}\colon \mathbb{R}^{d\times K} \to \mathbb{R}^{d\times K}$, we write :

$$\boldsymbol{\phi}(\mathbf{X}) = \begin{bmatrix} \boldsymbol{\phi}^1(\mathbf{X}) \\ \vdots \\ \boldsymbol{\phi}^d(\mathbf{X}) \end{bmatrix} \in \mathbb{R}^{d\times K} \tag{10.9}$$

where each $\boldsymbol{\phi}^i\colon \mathbb{R}^{d\times K} \to \mathbb{R}^K$. We then write the $K \times K$ Jacobian

$$\frac{\partial \boldsymbol{\phi}^i}{\partial \mathbf{X}_j}(\mathbf{X}) = \begin{bmatrix} \frac{\partial \phi^i_1(\mathbf{X})}{\partial X_{j1}} & \cdots & \frac{\partial \phi^i_1(\mathbf{X})}{\partial X_{jK}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi^i_K(\mathbf{X})}{\partial X_{j1}} & \cdots & \frac{\partial \phi^i_K(\mathbf{X})}{\partial X_{jK}} \end{bmatrix} \in \mathbb{R}^{K\times K} \tag{10.10}$$

For a given matrix $\mathbf{Q} \in \mathbb{R}^{K\times K}$, we write $\mathbf{Z} \in \mathbb{R}^{n\times K} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q} \otimes \mathbf{I}_n)$ to denote that the lines of $\mathbf{Z}$ are sampled i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{Q})$. Note that this is equivalent to saying that $\mathbf{Z} = \tilde{\mathbf{Z}}\mathbf{Q}^{1/2}$ where $\tilde{\mathbf{Z}} \in \mathbb{R}^{n\times K}$ is an i.i.d. standard normal random matrix. The notation $\overset{\mathrm{P}}{\simeq}$ denotes convergence in probability.

Approximate message-passing — Approximate message-passing algorithms are a statistical physics inspired family of iterations which can be used to solve high dimensional inference problems [300]. One of the central objects in such algorithms are the so called *state evolution equations*, a low-dimensional recursion equations which allow to exactly compute the high dimensional distribution of the iterates of the sequence. In this proof we will use a specific form of matrix-valued approximate message-passing iteration with non-separable non-linearities. In its full generality, the validity of the state evolution equations in this case is an extension of the works of [135, 37] included in [110]. Consider a sequence Gaussian matrices $\mathbf{A}(n) \in \mathbb{R}^{n\times d}$ with i.i.d. Gaussian entries, $A_{ij}(n) \sim \mathcal{N}(0, 1/d)$. For each $n, d \in \mathbb{N}$, consider two sequences of pseudo-Lipschitz functions

$$\{\mathbf{h}_t : \mathbb{R}^{n\times K} \to \mathbb{R}^{n\times K}\}_{t\in\mathbb{N}} \qquad \{\mathbf{e}_t : \mathbb{R}^{d\times K} \to \mathbb{R}^{d\times K}\}_{t\in\mathbb{N}} \tag{10.11}$$

initialized on $\mathbf{u}^0 \in \mathbb{R}^{d\times K}$ in such a way that the limit

$$\lim_{d\to\infty} \frac{1}{d}\left\|\mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0)\right\|_{\mathrm{F}} \tag{10.12}$$

exists and it is finite, and recursively define:

$$\mathbf{u}^{t+1} = \mathbf{A}^\top \mathbf{h}_t(\mathbf{v}^t) - \mathbf{e}_t(\mathbf{u}^t)\langle \mathbf{h}_t' \rangle^\top \tag{10.13}$$

$$\mathbf{v}^t = \mathbf{A}\mathbf{e}_t(\mathbf{u}^t) - \mathbf{h}_{t-1}(\mathbf{v}^{t-1})\langle \mathbf{e}_t' \rangle^\top \tag{10.14}$$

where the dimension of the iterates are $\mathbf{u}^t \in \mathbb{R}^{d\times K}$ and $\mathbf{v}^t \in \mathbb{R}^{n\times K}$. The terms in brackets are defined as:

$$\langle \mathbf{h}_t' \rangle = \frac{1}{d}\sum_{i=1}^{n} \frac{\partial \mathbf{h}_t^i}{\partial \mathbf{v}_i}(\mathbf{v}^t) \in \mathbb{R}^{K\times K} \quad \langle \mathbf{e}_t' \rangle = \frac{1}{d}\sum_{i=1}^{d} \frac{\partial \mathbf{e}_t^i}{\partial \mathbf{u}_i}(\mathbf{u}^t) \in \mathbb{R}^{K\times K} \tag{10.15}$$

We define now the *state evolution recursion* on two sequences of matrices $\{\mathbf{Q}_{r,s}\}_{s,r\geqslant 0}$ and $\{\hat{\mathbf{Q}}_{r,s}\}_{s,r\geqslant 1}$ initialized with $\mathbf{Q}_{0,0} = \lim_{d\to\infty} \frac{1}{d}\mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0)$:

$$\mathbf{Q}_{t+1,s} = \mathbf{Q}_{s,t+1} = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{e}_s(\hat{\mathbf{Z}}^s)^\top \mathbf{e}_{t+1}(\hat{\mathbf{Z}}^{t+1})\right] \in \mathbb{R}^{K\times K} \tag{10.16}$$

$$\hat{\mathbf{Q}}_{t+1,s+1} = \hat{\mathbf{Q}}_{s+1,t+1} = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{h}_s(\mathbf{Z}^s)^\top \mathbf{h}_t(\mathbf{Z}^t)\right] \in \mathbb{R}^{K\times K} \tag{10.17}$$

where $(\mathbf{Z}^0, \dots, \mathbf{Z}^{t-1}) \sim \mathcal{N}(\mathbf{0}, \{\mathbf{Q}_{r,s}\}_{0 \leqslant r,s \leqslant t-1} \otimes \mathbf{I}_n), (\hat{\mathbf{Z}}^1, \dots, \hat{\mathbf{Z}}^t) \sim \mathcal{N}(\mathbf{0}, \{\hat{\mathbf{Q}}_{r,s}\}_{1 \leqslant r,s \leqslant t} \otimes \mathbf{I}_d)$. Then the following holds

Theorem 19. *In the setting of the previous paragraph, for any sequence of pseudo-Lipschitz functions $\phi_n : (\mathbb{R}^{n \times K} \times \mathbb{R}^{d \times K})^t \to \mathbb{R}$, for $n, d \to +\infty$:*

$$\phi_n(\mathbf{u}^0, \mathbf{v}^0, \mathbf{u}^1, \mathbf{v}^1, \dots, \mathbf{v}^{t-1}, \mathbf{u}^t) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\mathbf{u}^0, \mathbf{Z}^0, \hat{\mathbf{Z}}^1, \mathbf{Z}^1, \dots, \mathbf{Z}^{t-1}, \hat{\mathbf{Z}}^t\right)\right] \tag{10.18}$$

where $(\mathbf{Z}^0, \dots, \mathbf{Z}^{t-1}) \sim \mathcal{N}(\mathbf{0}, \{\mathbf{Q}_{r,s}\}_{0 \leqslant r,s \leqslant t-1} \otimes \mathbf{I}_n), (\hat{\mathbf{Z}}^1, \dots, \hat{\mathbf{Z}}^t) \sim \mathcal{N}(\mathbf{0}, \{\hat{\mathbf{Q}}_{r,s}\}_{1 \leqslant r,s \leqslant t} \otimes \mathbf{I}_n)$.

Spatial coupling As a final premise to our proof, we give the intuition on how to handle a specific form of block random matrix in an AMP sequence. Consider the iteration (10.13), but this time with a Gaussian matrix defined as:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & (0) & \\ & (0) & \ddots & \\ & & & \mathbf{A}_K \end{bmatrix} \in \mathbb{R}^{n \times Kd} \tag{10.19}$$

where $\mathbf{A}_k \in \mathbb{R}^{n_k \times d}$ and $\sum_{k=1}^K n_k = n$, which leads to the following form for the products between matrices and non-linearities:

$$\mathbf{A}^\top \mathbf{h}_t(\mathbf{v}^t) = \begin{bmatrix} \mathbf{A}_1^\top \mathbf{h}_{1,t}(\mathbf{v}^t) \\ \mathbf{A}_2^\top \mathbf{h}_{2,t}(\mathbf{v}^t) \\ \vdots \\ \mathbf{A}_K^\top \mathbf{h}_{K,t}(\mathbf{v}^t) \end{bmatrix} \in \mathbb{R}^{Kd \times K} \quad \mathbf{A}\mathbf{e}_t(\mathbf{u}_t) = \begin{bmatrix} \mathbf{A}_1 \mathbf{e}_{1,t}(\mathbf{u}^t) \\ \mathbf{A}_2 \mathbf{e}_{2,t}(\mathbf{u}^t) \\ \vdots \\ \mathbf{A}_K \mathbf{e}_{K,t}(\mathbf{u}^t) \end{bmatrix} \in \mathbb{R}^{n \times K} \tag{10.20}$$

where the blocks $\mathbf{h}_{k,t}(\mathbf{v}^t) \in \mathbb{R}^{n_k \times K}, \mathbf{e}_{k,t}(\mathbf{u}_t) \in \mathbb{R}^{d \times K}$ may depend on their full arguments or only the corresponding blocks depending on their separability. This iteration can be embedded as a subset of the iterates of a larger sequence defined with the full version of the matrix $\mathbf{A}$ and non-linearities defined as:

$$\begin{aligned} &\mathbf{e}_t : \mathbb{R}^{Kd \times K^2} \to \mathbb{R}^{Kd \times K^2} \\ &\text{generates} \quad \begin{bmatrix} \mathbf{e}_{1,t}(\bullet) & & & \\ & \mathbf{e}_{2,t}(\bullet) & (0) & \\ & (0) & \ddots & \\ & & & \mathbf{e}_{K,t}(\bullet) \end{bmatrix} \in \mathbb{R}^{Kd \times K^2} \end{aligned} \tag{10.21}$$

$$\begin{aligned} &\mathbf{h}_t : \mathbb{R}^{n \times K^2} \to \mathbb{R}^{n \times K^2} \\ &\text{generates} \quad \begin{bmatrix} \mathbf{h}_{1,t}(\bullet) & & & \\ & \mathbf{h}_{2,t}(\bullet) & (0) & \\ & (0) & \ddots & \\ & & & \mathbf{h}_{K,t}(\bullet) \end{bmatrix} \in \mathbb{R}^{n \times K^2} \end{aligned} \tag{10.22}$$

The original iteration is recovered on the block diagonal of the variables of the iteration. This new setting, however, introduces a richer correlation structure, since each block will be described by a different $K \times K$ covariance according to the state evolution equations. Formally, the new covariance will be a $K^2 \times K^2$ block diagonal matrix. Also, the shape of the Onsager term changes from a matrix of size $K \times K$ to one of size $K^2 \times K^2$ with a $K \times (K \times K)$ block diagonal structure.

10.2 Reformulation of the problem

We start by reformulating problem (9.2) in a way that can be treated efficiently using an AMP iteration. With respect to the main part of this paper, we will consider the estimator $\mathbf{W} \in \mathbb{R}^{d\times K}$ instead of $\mathbb{R}^{K\times d}$. The normalized (so that the cost does not diverge with the dimension) problem (9.2) then reads:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times K},\mathbf{b}\in\mathbb{R}^K} \frac{1}{d}\left(L\left(\mathbf{Y}, \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}+\mathbf{b}\right)+r(\mathbf{W})\right) \tag{10.23}$$

where we have introduced the function $L : \mathbb{R}^{n\times K}\times\mathbb{R}^{n\times K}\to\mathbb{R}$ acting as

$$\left(\mathbf{Y}, \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}+\mathbf{b}\right) \mapsto \sum_{\nu=1}^{n} \ell\left(\mathbf{y}^\nu, \frac{\mathbf{W}\mathbf{x}^\nu}{\sqrt{d}}+\mathbf{b}\right), \tag{10.24}$$

the matrix $\mathbf{Y} \in \mathbb{R}^{n\times K}$ of concatenated one-hot encoded labels, and the matrix of concatenated means $\mathbf{M} \in \mathbb{R}^{K\times d}$ (in the main we took the transpose $\mathbf{M} \in \mathbb{R}^{d\times K}$). Until further notice, we will drop the scaling $\frac{1}{d}$ for convenience and study the problem

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times K},\mathbf{b}\in\mathbb{R}^K} L\left(\mathbf{Y}, \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}+\mathbf{b}\right)+r(\mathbf{W}) \tag{10.25}$$

We will write L_k the application of ℓ on each row of a sub-block in $\mathbb{R}^{n_k\times K}$. Without loss of generality, we can assume that the samples are grouped by clusters in the data matrix, giving the following form for $\mathbf{X} \in \mathbb{R}^{n\times d}$, separating the mean part $\mathbf{Y}\mathbf{M}$ and centered Gaussian part :

$$\mathbf{X} = \mathbf{Y}\mathbf{M} + \tilde{\mathbf{Z}}\mathbf{\Sigma} \in \mathbb{R}^{n\times d} \tag{10.26}$$

where we have introduced the block-diagonal matrix $\tilde{\mathbf{Z}}$ and the $Kd\times d$ full-column-rank matrix $\mathbf{\Sigma}$

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{Z}_1 & & & \\ & \mathbf{Z}_2 & (0) & \\ & (0) & \ddots & \\ & & & \mathbf{Z}_K \end{bmatrix} \in \mathbb{R}^{n\times Kd} \qquad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1^{1/2} \\ \mathbf{\Sigma}_2^{1/2} \\ \vdots \\ \mathbf{\Sigma}_K^{1/2} \end{bmatrix} \in \mathbb{R}^{Kd\times d}. \tag{10.27}$$

Here $(\mathbf{Z}_1,\dots,\mathbf{Z}_K) \in \mathbb{R}^{n_1\times d}\times\cdots\times\mathbb{R}^{n_K\times d}$ are independent, i.i.d. standard normal matrices.

The product between the data matrix and the weights $\mathbf{W} \in \mathbb{R}^{d\times K}$ then reads:

$$\mathbf{X}\mathbf{W} = \mathbf{Y}\mathbf{M}\mathbf{W} + \tilde{\mathbf{Z}}\mathbf{\Sigma}\mathbf{W} = \begin{bmatrix} \mathbf{Y}_1\mathbf{M}\mathbf{W} + \mathbf{Z}_1\mathbf{\Sigma}_1^{1/2}\mathbf{W} \\ \vdots \\ \mathbf{Y}_K\mathbf{M}\mathbf{W} + \mathbf{Z}_K\mathbf{\Sigma}_K^{1/2}\mathbf{W} \end{bmatrix} \in \mathbb{R}^{n\times K} \tag{10.28}$$

where each $\mathbf{Y}_k \in \mathbb{R}^{n_k\times d}$ is a n_k copy of the same label vector. Defining now $\tilde{\mathbf{W}} = \mathbf{\Sigma}\mathbf{W}$, observe that

$$\tilde{\mathbf{W}} = \mathbf{\Sigma}\mathbf{W} \quad\Longrightarrow\quad \mathbf{W} = \mathbf{\Sigma}^+\tilde{\mathbf{W}}, \tag{10.29}$$

where

$$\mathbf{\Sigma}^+ \equiv \left(\sum_{k=1}^{K}\mathbf{\Sigma}_k\right)^{-1}\mathbf{\Sigma}^\top \tag{10.30}$$

is the pseudo-inverse of the matrix $\mathbf{\Sigma}$. The optimization problem (9.2) is thus equivalent to

$$\inf_{\substack{\tilde{\mathbf{W}}\in\mathbb{R}^{Kd\times K}\\ \mathbf{b}\in\mathbb{R}^K}} \sum_{k=1}^{K} L_k\left(\frac{1}{\sqrt{d}}\mathbf{Y}_k\mathbf{M}\mathbf{W}+\frac{1}{\sqrt{d}}\mathbf{Z}_k\tilde{\mathbf{W}}_k,\mathbf{b}\right)+r\left(\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right) \tag{10.31}$$

Introducing the order parameter $\mathbf{M}=\frac{1}{\sqrt{d}}\mathbf{M}\mathbf{W}\in\mathbb{R}^{K\times K}$, we reformulate Eq.(10.31) as a constrained optimization problem :

$$\begin{aligned}\inf_{\mathbf{M},\tilde{\mathbf{W}},\mathbf{b}} &\sum_{k=1}^{K} L_k\left(\mathbf{Y}_k\mathbf{M}+\frac{1}{\sqrt{d}}\mathbf{Z}_k\tilde{\mathbf{W}}_k\right)+r\left(\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right) \\ \text{s.t.}\quad &\frac{1}{\sqrt{d}}\mathbf{M}\mathbf{\Sigma}^+\tilde{\mathbf{W}}=\mathbf{M}\end{aligned} \tag{10.32}$$

whose Lagrangian form, with dual parameters $\hat{\mathbf{M}}\in\mathbb{R}^{K\times K}$, reads

$$\inf_{\mathbf{M},\tilde{\mathbf{W}},\mathbf{b}}\sup_{\hat{\mathbf{M}}}\sum_{k=1}^{K} L_k\left(\mathbf{Y}_k\mathbf{M}+\frac{1}{\sqrt{d}}\mathbf{Z}_k\tilde{\mathbf{W}}_k\right)+r\left(\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right)+\operatorname{tr}\left(\hat{\mathbf{M}}^\top\left(\mathbf{M}-\frac{1}{\sqrt{d}}\mathbf{M}\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right)\right). \tag{10.33}$$

This is a proper, closed, convex, strictly feasible optimization problem, thus strong duality holds and we can invert the order of the inf-sup to focus on the minimization problem in $\tilde{\mathbf{W}}$ for fixed $\mathbf{M},\hat{\mathbf{M}},\mathbf{b}$:

$$\inf_{\tilde{\mathbf{W}}\in\mathbb{R}^{Kd\times K}}\tilde{L}\left(\frac{1}{\sqrt{d}}\tilde{\mathbf{Z}}\tilde{\mathbf{W}}\right)+\tilde{r}(\tilde{\mathbf{W}}) \tag{10.34}$$

where we defined the loss term

$$\begin{aligned}&\tilde{L}:\mathbb{R}^{n\times K}\to\mathbb{R}\\ &\frac{1}{\sqrt{d}}\tilde{\mathbf{Z}}\tilde{\mathbf{W}}\mapsto\sum_{k=1}^{K}L_k\left(\mathbf{Y}_k\mathbf{M}+\frac{1}{\sqrt{d}}\mathbf{Z}_k\tilde{\mathbf{W}}_k\right)=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\ell\left(\left[\mathbf{Y}_k\mathbf{M}+\frac{1}{\sqrt{d}}\mathbf{Z}_k\tilde{\mathbf{W}}_k\right]_i\right)\end{aligned} \tag{10.35a}$$

and the regularisation term

$$\begin{aligned}&\tilde{r}:\mathbb{R}^{Kd\times K}\to\mathbb{R}\\ &\tilde{\mathbf{W}}\mapsto r\left(\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right)+\operatorname{tr}\left(\hat{\mathbf{M}}^\top\left(\mathbf{M}-\frac{1}{\sqrt{d}}\mathbf{M}\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right)\right)\end{aligned} \tag{10.35b}$$

where $\mathbf{\Sigma}^\top\tilde{\mathbf{W}}=\sum_{k=1}^{K}\mathbf{\Sigma}_k^{1/2}\mathbf{W}_k$ and $\tilde{\mathbf{Z}}=[\mathbf{Z}_k]_{k=1}^{K}\in\mathbb{R}^{n\times Kd}$ is an i.i.d. standard normal block diagonal matrix as in Eq. (10.27).

10.3 Finding the AMP sequence

We now need to find an AMP iteration relating to $\tilde{\mathbf{W}}$ that solve the optimization problem in Eq. (10.34). Although this section is not written as a formal proof, all steps are rigorous. The aim is to give the reader the core intuition on how the AMP iteration is found, otherwise the solution may feel "parachuted". The reader uninterested in the underlying intuition may directly skip to the next section. In order to find the appropriate sequence two key points must be considered :

- the fixed point of the sequence has to match the optimality condition of Eq. (10.34);

- the update rule of the sequence should have the form Eq. (10.13) for the state evolution equations to hold.

These two points completely determine the form of the iteration. In the subsequent derivation, we absorb the scaling $\frac{1}{\sqrt{d}}$ in the matrix $\tilde{\mathbf{Z}}$, such that the $\mathbf{Z}_k \in \mathbb{R}^{n_k \times d}$ have i.i.d. $\mathcal{N}(0, 1/d)$ elements.

Resolvent of the loss term — Going back to problem Eq. (10.34), its optimality condition will look like :

$$\tilde{\mathbf{Z}}^\top \partial \tilde{L}(\mathbf{Z}\tilde{\mathbf{W}}) + \partial \tilde{r}(\tilde{\mathbf{W}}) = 0 \iff \begin{bmatrix} \mathbf{Z}_1^\top & & & \\ & \mathbf{Z}_2^\top & (0) & \\ & (0) & \ddots & \\ & & & \mathbf{Z}_K^\top \end{bmatrix} \begin{bmatrix} \partial \tilde{L}_1(\mathbf{Z}_1\tilde{\mathbf{W}}_1) \\ \partial \tilde{L}_2(\mathbf{Z}_2\tilde{\mathbf{W}}_2)) \\ \vdots \\ \partial \tilde{L}_K(\mathbf{Z}_K\tilde{\mathbf{W}}_K)) \end{bmatrix} + \partial \tilde{r}(\tilde{\mathbf{W}}) = 0 \quad (10.36)$$

where each $\mathbf{Z}_k \in \mathbb{R}^{n_k \times d}$, and the subdifferential of $\tilde{L}$ is separable across blocks of size $n_k \times d$, and $\partial \tilde{r}(\tilde{\mathbf{W}}) \in \mathbb{R}^{Kd \times K}$. Following the intuition of spatial coupling, we introduce the *full* matrix $\mathbf{Z} \in \mathbb{R}^{n \times Kd}$, with i.i.d. $\mathcal{N}(0, 1/d)$ entries. The optimality condition can then be written on the diagonal of a $Kd \times K^2$ matrix:

$$\mathbf{Z}^\top \begin{bmatrix} \partial \tilde{L}_1(\mathbf{Z}_1\tilde{\mathbf{W}}_1) & & & \\ & \partial \tilde{L}_2(\mathbf{Z}_2\tilde{\mathbf{W}}_2) & (0) & \\ & (0) & \ddots & \\ & & & \partial \tilde{L}_K(\mathbf{Z}_K\tilde{\mathbf{W}}_K) \end{bmatrix} + \begin{bmatrix} \partial \tilde{r}(\tilde{\mathbf{W}})_1 & & & \\ & \partial \tilde{r}(\tilde{\mathbf{W}})_2 & (0) & \\ & (0) & \ddots & \\ & & & \partial \tilde{r}(\tilde{\mathbf{W}})_K \end{bmatrix} = \mathbf{0} \quad (10.37)$$

where $\partial \tilde{r}(\tilde{\mathbf{W}})_k$ represents the k-th block of the subdifferential of $\tilde{r}$ which is non-separable across the blocks of $\tilde{\mathbf{W}}$. To make the resolvents/proximals appear, we add the argument of the subdifferentials on both sides weighted by a (symmetric) positive definite matrix $\mathbf{S}_k \in \mathbb{R}^{K \times K}$ which will be used to allow for Onsager correction while respecting the fixed point condition. Using the notation defined in section 10.1

$$\begin{aligned} &\left[\mathbf{Z}_k^\top \partial \tilde{L}_k(\mathbf{Z}_k\tilde{\mathbf{W}}_k)\right]_{k=1}^K + \left[\partial \tilde{r}(\tilde{\mathbf{W}})\right]_{k=1}^K = 0 \\ &\iff \left[\mathbf{Z}_k^\top \partial \tilde{L}_k(\mathbf{Z}_k\tilde{\mathbf{W}}_k) + \mathbf{Z}_k^\top \mathbf{Z}_k \tilde{\mathbf{W}}_k \mathbf{S}_k^{-1}\right]_{k=1}^K + \left[\partial \tilde{r}(\tilde{\mathbf{W}})\right]_{k=1}^K = \left[\mathbf{Z}_k^\top \mathbf{Z}_k \tilde{\mathbf{W}}_k \mathbf{S}_k^{-1}\right]_{k=1}^K \end{aligned} \quad (10.38)$$

for a given set of positive definite matrices $\{\mathbf{S}_k\}_{k \in [K]}$. Again, the reason for introducing different $\mathbf{S}_k$ on each block is to match the expected structure of the Onsager term. We can introduce the resolvent, formally Bregman resolvent/proximal operator:

$$\mathbf{U}_k \equiv \partial \tilde{L}_k(\mathbf{Z}_k\tilde{\mathbf{W}}_k)\mathbf{S}_k + \mathbf{Z}_k\tilde{\mathbf{W}}_k \iff \mathbf{Z}_k\tilde{\mathbf{W}}_k = \mathbf{R}_{\tilde{L}_k, \mathbf{S}_k}(\mathbf{U}_k) \quad (10.39)$$

where

$$\begin{aligned}\mathbf{R}_{\tilde{L}_k,\mathbf{S}_k}(\mathbf{U}_k) &= (\mathrm{Id} + \partial\tilde{L}_k(\bullet)\mathbf{S}_k)^{-1}(\mathbf{U}_k)\\ &= \underset{\mathbf{T}\in\mathbb{R}^{n_k\times K}}{\arg\min}\left\{\tilde{L}_k(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\mathbf{U}_k)\mathbf{S}_k^{-1}(\mathbf{T}-\mathbf{U}_k)^\top\right)\right\}\\ &= \underset{\mathbf{T}\in\mathbb{R}^{n_k\times K}}{\arg\min}\left\{L_k(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-(\mathbf{Y}_k\mathbf{M}+\mathbf{U}_k))\mathbf{S}_k^{-1}(\mathbf{T}-(\mathbf{Y}_k\mathbf{M}+\mathbf{U}_k))^\top\right)\right\} - \mathbf{Y}_k\mathbf{M}.\end{aligned} \tag{10.40}$$

In the previous expressions $\partial\tilde{L}_k \in \mathbb{R}^{n_k\times K}$ and $\mathbf{V}_k \in \mathbb{R}^{K\times K}$. The following formulation of the optimality condition is reached:

$$\begin{aligned}&\left[\mathbf{Z}_k^\top\mathbf{U}_k\mathbf{S}_k^{-1}\right]_{k=1}^K + \left[\partial\tilde{r}(\tilde{\mathbf{W}})_k\right]_{k=1}^K = \left[\mathbf{Z}_k^\top\mathbf{R}_{\tilde{L}_k,\mathbf{S}_k}(\mathbf{U}_k)\mathbf{S}_k^{-1}\right]_{k=1}^K\\ \iff &\left[\mathbf{Z}_k^\top\left(\mathbf{U}_k - \mathbf{R}_{\tilde{L}_k,\mathbf{S}_k}(\mathbf{U}_k)\right)\mathbf{S}_k^{-1}\right]_{k=1}^K + \left[\partial\tilde{r}(\tilde{\mathbf{W}})_k\right]_{k=1}^K = 0\end{aligned} \tag{10.41}$$

Resolvent of the regularization term Determining the block decomposition of the subdifferential of the regularization term is less simple. We would like a block expression in the flavour of:

$$\left[\partial\tilde{r}(\tilde{\mathbf{W}})_k\right]_{k=1}^K + \left[\tilde{\mathbf{W}}_k\hat{\mathbf{S}}_k^{-1}\right]_{k=1}^K = \left[\tilde{\mathbf{W}}_k\hat{\mathbf{S}}_k^{-1}\right]_{k=1}^K \tag{10.42}$$

At this point it becomes clear that we cannot consider the resolvent as acting on $\tilde{\mathbf{W}} \in \mathbb{R}^{Kd\times K}$ otherwise there could be only one $\hat{\mathbf{S}} \in \mathbb{R}^{K\times K}$ and there would be a mismatch with the expected form of the Onsager terms. As specified by the definitions Eq.(10.35), the subdifferential of $\tilde{r}$ is acting on the whole block diagonal matrix $[\tilde{\mathbf{W}}_k]_{k=1}^K$, by way of summation due to the action of the pseudo-inverse $\mathbf{\Sigma}^+$. We can thus consider its proximal acting on $\mathbb{R}^{d\times K^2}$ as $[\tilde{\mathbf{W}}_1\tilde{\mathbf{W}}_2...\tilde{\mathbf{W}}_K]$ (note that we could have also worked directly with a block diagonal matrix in $\mathbb{R}^{Kd\times K^2}$). Proceeding in this way, we can directly write our expression as an application parametrized by another set of positive definite matrices $\{\hat{\mathbf{S}}_k\}_{k\in[K]}$.

$$\hat{\mathbf{U}} = \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}}\right)(\tilde{\mathbf{W}}) \qquad \tilde{\mathbf{W}} = \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{U}}) \tag{10.43}$$

where

$$\begin{aligned}\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{U}}) &= \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}}\right)^{-1}(\hat{\mathbf{U}})\\ &= \underset{\mathbf{T}\in\mathbb{R}^{d\times K^2}}{\arg\min}\left\{\tilde{r}(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\hat{\mathbf{U}})\hat{\mathbf{S}}^{-1}(\mathbf{T}-\hat{\mathbf{U}})^\top\right)\right\}\end{aligned} \tag{10.44}$$

where $\hat{\mathbf{S}} \in \mathbb{R}^{K^2\times K^2}$ block diagonal, and $\hat{\mathbf{U}} \in \mathbb{R}^{d\times K^2}$. This would lead to the equivalent optimality condition for the regularization part:

$$\hat{\mathbf{U}}\hat{\mathbf{S}}^{-1} = \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{U}})\hat{\mathbf{S}}^{-1} \iff \left[\hat{\mathbf{U}}_k\hat{\mathbf{S}}_k^{-1}\right]_{k=1}^K = \left[\mathbf{R}_{\tilde{r},\hat{\mathbf{S}},k}(\hat{\mathbf{U}})\hat{\mathbf{S}}_k^{-1}\right]_{k=1}^K \tag{10.45}$$

We now need to figure out the block structure of this resolvent since we want to spread it across a block diagonal matrix. Let $\mathbf{C} = \sum_{k=1}^K \mathbf{\Sigma}_k$, so that $\mathbf{\Sigma}^+ = \mathbf{C}^{-1}\mathbf{\Sigma}^\top$, and the blocks $\mathbf{T}_k \in \mathbb{R}^{d\times K}$ are

the solution to the minimization problem

$$\min_{\{\mathbf{T}_k\}_{k\in[K]}\in(\mathbb{R}^{d\times K})^K} r(\mathbf{C}^{-1}\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\mathbf{T}_k) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\hat{\mathbf{U}})\hat{\mathbf{S}}^{-1}(\mathbf{T}-\hat{\mathbf{U}}^\top)\right) + \mathrm{tr}\left(\hat{\mathbf{M}}^\top\left(\mathbf{M}-\frac{1}{\sqrt{d}}\mathbf{M}\mathbf{\Sigma}^+\mathbf{T}\right)\right) \tag{10.46}$$

Let $\tilde{\mathbf{T}} = \mathbf{C}^{-1}\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\mathbf{T}_k \in \mathbb{R}^{d\times K}$, and the equivalent reformulation as a constraint optimization problem:

$$\min_{\substack{\mathbf{T}_{k\in[K]}\in\mathbb{R}^{d\times K}\\ \tilde{\mathbf{T}}\in\mathbb{R}^{d\times K}}} r(\tilde{\mathbf{T}}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\hat{\mathbf{U}})\hat{\mathbf{S}}^{-1}(\mathbf{T}-\hat{\mathbf{U}}^\top)\right) + \mathrm{tr}\left(\hat{\mathbf{M}}^\top\left(\mathbf{M}-\frac{1}{\sqrt{d}}\mathbf{M}\tilde{\mathbf{T}}\right)\right) \tag{10.47}$$

$$\text{s.t.} \quad \tilde{\mathbf{T}} = \mathbf{C}^{-1}\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\mathbf{T}_k$$

This is a feasible convex problem under convex constraint with a strongly convex term, it thus has a unique solution and strong duality holds. Introducing the Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^{d\times K}$, we get the equivalent representation:

$$\min_{\substack{\mathbf{T}_{k\in[K]}\in\mathbb{R}^{d\times K}\\ \tilde{\mathbf{T}}\in\mathbb{R}^{d\times K}}} \max_{\boldsymbol{\lambda}\in\mathbb{R}^{d\times K}} r(\tilde{\mathbf{T}}) + \sum_{k=1}^K \mathrm{tr}\left((\mathbf{T}_k-\hat{\mathbf{U}}_k)\hat{\mathbf{S}}_k^{-1}(\mathbf{T}_k-\hat{\mathbf{U}}_k)^\top\right) + \mathrm{tr}\left(\boldsymbol{\lambda}^\top\left(\tilde{\mathbf{T}}-\mathbf{C}^{-1}\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\mathbf{T}_k\right)\right) + \mathrm{tr}\left(\hat{\mathbf{M}}^\top\left(\mathbf{M}-\frac{1}{\sqrt{d}}\mathbf{M}\tilde{\mathbf{T}}\right)\right). \tag{10.48}$$

The optimality condition for this problem reads:

$$\partial_{\tilde{\mathbf{T}}}: \quad \partial r(\tilde{\mathbf{T}}) + \boldsymbol{\lambda} - \frac{1}{\sqrt{d}}\mathbf{M}^\top\hat{\mathbf{m}} = 0 \tag{10.49}$$

$$\partial_{\mathbf{T}}: \quad (\mathbf{T}_k-\mathbf{U}_k)\hat{\mathbf{S}}_k^{-1} = \mathbf{\Sigma}_k^{1/2}\mathbf{C}^{-1}\boldsymbol{\lambda} \qquad \forall k\in[K] \tag{10.50}$$

$$\partial_{\boldsymbol{\lambda}}: \quad \tilde{\mathbf{T}} = \mathbf{C}^{-1}\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\mathbf{T}_k \tag{10.51}$$

Using the gradient condition on $\mathbf{T}$, we get

$$\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}(\mathbf{T}_k-\hat{\mathbf{U}}_k)\hat{\mathbf{S}}_k^{-1} = \boldsymbol{\lambda} \tag{10.52}$$

The constraint $\tilde{\mathbf{T}} = \mathbf{C}^{-1}\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\mathbf{T}_k$ is solved by $\mathbf{T}_k = \mathbf{\Sigma}_k^{1/2}\tilde{\mathbf{T}}$ which gives the solution for $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda} = \sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}(\mathbf{\Sigma}_k^{1/2}\tilde{\mathbf{T}}-\hat{\mathbf{U}}_k)\hat{\mathbf{S}}_k^{-1} = \sum_{k=1}^K \mathbf{\Sigma}_k\tilde{\mathbf{T}}\hat{\mathbf{S}}_k^{-1} - \sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\hat{\mathbf{U}}_k\hat{\mathbf{S}}_k^{-1} \tag{10.53}$$

and prescribes the following form for $\tilde{\mathbf{T}}$, as solution to the problem

$$\begin{aligned} &\partial r(\tilde{\mathbf{T}}) + \sum_{k=1}^K \mathbf{\Sigma}_k\tilde{\mathbf{T}}\hat{\mathbf{S}}_k^{-1} - \sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\hat{\mathbf{U}}_k\hat{\mathbf{S}}_k^{-1} - \frac{1}{\sqrt{d}}\mathbf{M}^\top\hat{\mathbf{M}} = 0 \\ \iff &\arg\min_{\tilde{\mathbf{T}}} r(\tilde{\mathbf{T}}) + \frac{1}{2}\sum_{k=1}^K \mathbf{\Sigma}_k\tilde{\mathbf{T}}\hat{\mathbf{S}}_k^{-1}\tilde{\mathbf{T}} - \left(\sum_{k=1}^K \mathbf{\Sigma}_k^{1/2}\hat{\mathbf{U}}_k\hat{\mathbf{S}}_k^{-1} + \frac{1}{\sqrt{d}}\mathbf{M}^\top\hat{\mathbf{M}}\right)\tilde{\mathbf{T}} \end{aligned} \tag{10.54}$$

We then recover $\mathbf{T}$ from $\mathbf{T} = \boldsymbol{\Sigma}\tilde{\mathbf{T}}$. Thus, defining the function

$$\begin{aligned}
&\boldsymbol{\eta} : \mathbb{R}^{d\times K^2} \to \mathbb{R}^{d\times K} \\
&\hat{\mathbf{U}} \mapsto \arg\min_{\tilde{\mathbf{T}}} r(\tilde{\mathbf{T}}) + \frac{1}{2}\sum_{k=1}^{K} \boldsymbol{\Sigma}_k \tilde{\mathbf{T}} \hat{\mathbf{S}}_k^{-1} \tilde{\mathbf{T}} - \left(\sum_{k=1}^{K} \boldsymbol{\Sigma}_k^{1/2} \hat{\mathbf{U}}_k \hat{\mathbf{S}}_k^{-1} + \frac{1}{\sqrt{d}} \mathbf{M}^\top \hat{\mathbf{M}} \right) \tilde{\mathbf{T}}
\end{aligned} \tag{10.55}$$

the block decomposition of the resolvent for the regularizer reads:

$$\mathbf{R}_{\tilde{r},\hat{\mathbf{S}},k}(\hat{\mathbf{U}}) = \boldsymbol{\Sigma}_k^{1/2} \boldsymbol{\eta}(\hat{\mathbf{U}}) \tag{10.56}$$

Matching the optimality condition with the AMP fixed point The global optimality condition then reads:

$$\left[\mathbf{Z}_k^\top \left(\mathbf{R}_{\tilde{L}_k,\mathbf{S}_k}(\mathbf{U}_k) - \mathbf{U}_k \right) \mathbf{S}_k^{-1} \right]_{k=1}^{K} = \left[(\hat{\mathbf{U}}_k - \mathbf{R}_{\tilde{r},\hat{\mathbf{S}},k}(\hat{\mathbf{U}})) \hat{\mathbf{S}}_k^{-1} \right]_{k=1}^{K} \tag{10.57}$$

$$\left[\mathbf{Z}_k \mathbf{R}_{\tilde{r},\hat{\mathbf{S}},k}(\hat{\mathbf{U}}) \right]_{k=1}^{K} = \left[\mathbf{R}_{\tilde{L}_k,\mathbf{S}_k}(\mathbf{U}_k) \right]_{k=1}^{K} \tag{10.58}$$

where both equations should be satisfied. We can now define update functions based on the previously obtained block decomposition. The fixed point of the matrix-valued AMP Eq.(10.13) reads:

$$\mathrm{Id} + \mathbf{e}(\mathbf{u}) \langle \mathbf{h}' \rangle^\top = \mathbf{Z}^\top \mathbf{h}(\mathbf{v}) \tag{10.59}$$

$$\mathrm{Id} + \mathbf{h}(\mathbf{v}) \langle \mathbf{e}' \rangle^\top = \mathbf{Z} \mathbf{e}(\mathbf{u}) \tag{10.60}$$

Matching this fixed point with the optimality condition Eq.(10.57) suggests the following mapping:

$$\begin{aligned}
\mathbf{h}_k(\mathbf{U}_k) &= \left(\mathbf{R}_{\tilde{L}_k,\mathbf{S}_k}(\mathbf{U}_k) - \mathbf{U}_k \right) \mathbf{S}_k^{-1}, & \mathbf{S}_k &= \langle \mathbf{e}_k' \rangle, \\
\mathbf{e}_k(\hat{\mathbf{U}}) &= \mathbf{R}_{\tilde{r},\hat{\mathbf{S}},k}(\hat{\mathbf{U}}\hat{\mathbf{S}}), & \hat{\mathbf{S}}_k &= -\langle \mathbf{h}_k' \rangle^{-1},
\end{aligned} \tag{10.61}$$

where we redefined $\hat{\mathbf{U}} \equiv \hat{\mathbf{U}}\hat{\mathbf{S}}$ in (10.43), and the subscripts on the non-linearities are block indexes.

10.4 Proof of Theorem 17 using the AMP sequence

Following the analysis carried out in the previous section, define the following two sequences of non-linearities, for fixed values of the parameters $\hat{\mathbf{M}}, \mathbf{M}, \mathbf{b}$ and any $\mathbf{u} \in \mathbb{R}^{d\times K^2}, \mathbf{v} \in \mathbb{R}^{n\times K}$:

$$\begin{aligned}
&\mathbf{e}_t : \mathbb{R}^{Kd\times K^2} \to \mathbb{R}^{Kd\times K^2} \\
&\mathbf{u} \mapsto \begin{bmatrix} \mathbf{e}_{1,t}(\mathbf{u}) & & & \\ & \mathbf{e}_{2,t}(\mathbf{u}) & (0) & \\ & (0) & \ddots & \\ & & & \mathbf{e}_{K,t}(\mathbf{u}) \end{bmatrix} \in \mathbb{R}^{Kd\times K^2}
\end{aligned} \tag{10.62}$$

$$\begin{aligned}
&\mathbf{h}_t : \mathbb{R}^{n\times K^2} \to \mathbb{R}^{n\times K^2} \\
&\mathbf{v} \mapsto \begin{bmatrix} \mathbf{h}_{1,t}(\mathbf{v}_1) & & & \\ & \mathbf{h}_{2,t}(\mathbf{v}_2) & (0) & \\ & (0) & \ddots & \\ & & & \mathbf{h}_{K,t}t(\mathbf{v}_K) \end{bmatrix} \in \mathbb{R}^{n\times K^2}
\end{aligned} \tag{10.63}$$

where $\mathbf{Y}_k \in \mathbb{R}^{n_k \times K}$ and

$$
\begin{aligned}
&\mathbf{h}_{k,t} : \mathbb{R}^{n_k \times K} \to \mathbb{R}^{n_k \times K} \\
&\quad \mathbf{v}_k \mapsto \left(\mathbf{R}_{\tilde{L}_k, \mathbf{V}^{k,t}}(\mathbf{v}_k) - \mathbf{v}_k\right)(\mathbf{V}^{k,t})^{-1} \\
&\quad = \left(\underset{\mathbf{T} \in \mathbb{R}^{n_k \times K}}{\arg\min} \left\{ \tilde{L}_k(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T} - \mathbf{v}_k)(\mathbf{V}_{k,t})^{-1}(\mathbf{T} - \mathbf{v}_k)^\top\right)\right\} - \mathbf{v}_k\right)(\mathbf{V}_{k,t})^{-1} \\
&\quad = \left(\mathrm{Prox}_{L_k(\bullet(\mathbf{V}_{k,t})^{1/2})}((\mathbf{Y}_k\mathbf{M} + \mathbf{v}_k)(\mathbf{V}_{k,t})^{-1/2})(\mathbf{V}_{k,t})^{1/2} - (\mathbf{Y}_k\mathbf{M} + \mathbf{v}_k)\right)(\mathbf{V}_{k,t})^{-1} \qquad (10.64) \\
&\mathbf{e}_{k,t} : \mathbb{R}^{d \times K^2} \to \mathbb{R}^{d \times K} \\
&\quad \mathbf{u} \mapsto \boldsymbol{\Sigma}_k^{1/2} \underset{\tilde{\mathbf{T}} \in \mathbb{R}^{d \times K}}{\arg\min}\, r(\tilde{\mathbf{T}}) + \frac{1}{2}\sum_{k=1}^K \boldsymbol{\Sigma}_k \tilde{\mathbf{T}} \hat{\mathbf{V}}_{k,t} \tilde{\mathbf{T}} - \left(\sum_{k=1}^K \boldsymbol{\Sigma}_k^{1/2}\mathbf{u}_k + \frac{1}{\sqrt{d}}\mathbf{M}^\top \hat{\mathbf{M}}\right)\tilde{\mathbf{T}} \\
&\quad = \boldsymbol{\Sigma}_k^{1/2} \boldsymbol{\eta}(\mathbf{u}(\hat{\mathbf{V}}^t)^{-1}) \qquad (10.65)
\end{aligned}
$$

where $(\mathbf{V}_t, \hat{\mathbf{V}}_t) \in \mathbb{R}^{K^2 \times K^2}$, are defined as the block diagonal matrices $[\mathbf{V}_{k,t}]_{k\in[K]}, \left[\hat{\mathbf{V}}_{k,t}\right]_{k\in[K]}$ such that

$$
\mathbf{V}_{k,t} = \langle (\mathbf{e}_{k,t-1})' \rangle \quad \hat{\mathbf{V}}_{k,t} = -\langle (\mathbf{h}_{k,t})' \rangle \qquad (10.66)
$$

using the notation from Eq. (10.15). Now define the following sequence, initialized with

$$
\begin{aligned}
&\mathbf{u}^0, \mathbf{h}^{-1} \equiv 0, \hat{\mathbf{V}}_0 \qquad (10.67) \\
&\text{such that } \lim_{d\to\infty} \frac{1}{d}\left\|\mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0)\right\|_{\mathrm{F}} < +\infty \text{ and } \hat{\mathbf{V}}_0 \in \mathbb{S}_K^{++} \\
&\text{and recursively define} \\
&\quad \mathbf{u}^{t+1} = \boldsymbol{Z}^\top \mathbf{h}_t(\mathbf{v}^t) - \mathbf{e}_t(\mathbf{u}^t)\langle \mathbf{h}_t' \rangle^\top \qquad (10.68) \\
&\quad \mathbf{v}^t = \boldsymbol{Z}\mathbf{e}_t(\mathbf{u}^t) - \mathbf{h}_{t-1}(\mathbf{v}^{t-1})\langle \mathbf{e}_t' \rangle^\top \qquad (10.69)
\end{aligned}
$$

where $\boldsymbol{Z} \in \mathbb{R}^{n \times Kd}$ has i.i.d. $\mathcal{N}(0, 1/d)$ elements, and in the Jacobians defining $\hat{\mathbf{V}}, \mathbf{V}$, we used the notation from Eq. (10.10).

State evolution equations The results from section 10.3 show that the functions $\mathbf{e}^t, \mathbf{h}^t$ are proximals operators, and thus are Lipschitz continuous for all $t \in \mathbb{N}$, along with their block restrictions. Therefore the conditions of Theorem 19 are verified and we have the following lemma:

Lemma 43. *Consider the sequence defined by Eq.(10.68), for any fixed* $\mathbf{M}, \hat{\mathbf{M}}, \mathbf{b}$. *For any sequences of pseudo-Lipschitz functions* $\phi_{1,n} : \mathbb{R}^{d \times K^2} \to \mathbb{R}, \phi_{2,n} : \mathbb{R}^{n \times K^2} \to \mathbb{R}$, *for any* $t \in \mathbb{N}^*$:

$$
\begin{aligned}
\phi_{1,n}(\mathbf{u}_1^t, \dots, \mathbf{u}_K^t) &\overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{1,n}(\mathbf{H}_1(\hat{\mathbf{Q}}_{1,t})^{1/2}, \dots, \mathbf{H}_K(\hat{\mathbf{Q}}_{K,t})^{1/2})\right] \qquad (10.70) \\
\phi_{2,n}(\mathbf{v}_1^t, \dots, \mathbf{v}_K^t) &\overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{1,n}(\mathbf{G}_1(\mathbf{Q}_{1,t})^{1/2}, \dots, \mathbf{G}_K(\mathbf{Q}_{K,t})^{1/2})\right] \qquad (10.71)
\end{aligned}
$$

where the matrices $\mathbf{H}_k \in \mathbb{R}^{d \times K}, \mathbf{G}_k \in \mathbb{R}^{n_k \times K}$ *are independent matrices with i.i.d. standard normal*

elements, and at each time step $t \geqslant 1$

$$\mathbf{Q}_{k,t} = \lim_{d\to+\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{e}_{k,t}(\{\mathbf{H}_k(\hat{\mathbf{Q}}_{k,t})^{1/2}(\hat{\mathbf{V}}_{k,t})^{-1}\}_{k\in[K]})^\top \mathbf{e}_{k,t}(\{\mathbf{H}_k(\hat{\mathbf{Q}}_{k,t})^{1/2}(\hat{\mathbf{V}}_{k,t})^{-1}\}_{k\in[K]})\right] \in \mathbb{R}^{K\times K} \tag{10.72}$$

$$\hat{\mathbf{Q}}_{k,t} = \lim_{d\to+\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{h}_{k,t-1}(\mathbf{G}_k(\mathbf{Q}_{k,t-1})^{1/2})^\top \mathbf{h}_{k,t-1}(\mathbf{G}_k(\mathbf{Q}_{k,t-1})^{1/2})\right] \in \mathbb{R}^{K\times K} \tag{10.73}$$

$$\mathbf{V}_{k,t} = \lim_{d\to+\infty} \frac{1}{d}\sum_{i=1}^{d} \frac{\partial \mathbf{e}_{k,t-1}(\{\mathbf{H}_k(\hat{\mathbf{Q}}_{k,t-1})^{1/2}\}_{k\in[K]})}{\partial(\mathbf{H}_k(\hat{\mathbf{Q}}_{k,t-1})^{1/2})_i} \in \mathbb{R}^{K\times K} \tag{10.74}$$

$$\hat{\mathbf{V}}_{k,t} = -\lim_{d\to+\infty} \frac{1}{d}\sum_{i=1}^{n_k} \frac{\partial \mathbf{h}_{k,t}(\mathbf{G}_k(\mathbf{Q}_{k,t})^{1/2})}{\partial(\mathbf{G}_k(\mathbf{Q}_{k,t})^{1/2})_i} \in \mathbb{R}^{K\times K} \tag{10.75}$$

where the sequence is initialized with $\hat{\mathbf{V}}_0, \mathbf{e}_0, \mathbf{Q}_{0,0} = \lim_{d\to\infty} \frac{1}{d}\left\|\mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0)\right\|_{\mathrm{F}}$.

Proof. Lemma 43 is a consequence of Theorem 19 whose assumptions have been verified in the paragraph. □

Note that in Lemma 43, we have directly written the block decomposition of the state evolution corresponding to the iteration Eq. (10.68), which involves the block diagonal matrices $\mathbf{Q}_t, \hat{\mathbf{Q}}_t, \mathbf{V}_t, \hat{\mathbf{V}}_t$ which are all in $\mathbb{R}^{K^2\times K^2}$. Using the notations introduced in section 10.1

$$\mathbf{V} = [\mathbf{V}_k]_{k=1}^K \quad \hat{\mathbf{V}} = \left[\hat{\mathbf{V}}_k\right]_{k=1}^K \quad \mathbf{Q} = [\mathbf{Q}_k]_{k=1}^K \quad \hat{\mathbf{Q}} = \left[\hat{\mathbf{Q}}_k\right]_{k=1}^K \tag{10.76}$$

Also note that we do not use the full state evolution giving the correlations across all time steps, but only use those at equal times t.

Trajectories and fixed point of the AMP sequence Now that we have a sequence with state evolution equations, the following two lemmas link the fixed points of this iteration to any optimal solution of problem Eq.(10.34).

Lemma 44. *Consider any fixed point* $\mathbf{V}, \hat{\mathbf{V}}, \mathbf{Q}, \hat{\mathbf{Q}}$ *of the state evolution equations from Lemma 43. For any fixed point* $\mathbf{u}^*, \mathbf{v}^*$ *of iteration Eq.(10.68), the quantity*

$$\mathbf{R}_{\tilde{r},\hat{\mathbf{V}}^{-1}}(\mathbf{u}^*\hat{\mathbf{V}}^{-1}) = \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\mathbf{V}}^{-1}\right)(\mathbf{u}^*\hat{\mathbf{V}}^{-1}) \tag{10.77}$$

is an optimal solution $\tilde{\mathbf{W}}^\star$ *of problem Eq.(10.34). Furthermore*

$$\mathbf{R}_{\tilde{L},\mathbf{V}}(\mathbf{v}^*) = (\mathrm{Id} + \partial\tilde{L}(\bullet)\mathbf{V})(\mathbf{v}^*) = \mathbf{Z}\tilde{\mathbf{W}}^\star \tag{10.78}$$

where the block decompositions of each resolvents have been explicitly calculated in section 10.3.

Proof. Lemma 44 is a direct consequence of the analysis carried out in section 10.3. □

At this point we know the fixed points of the AMP iteration correspond to the optimal solutions of problem Eq.(10.34). Note that the resolvents/proximals linking the fixed point of the AMP iteration with the solutions of Eq.(10.34) are Lipschitz continuous, making them acceptable transforms for state evolution observables. However this does not guarantee that the optimal solution is characterized by the fixed point of the state evolution equations. Indeed, we need to show that a converging trajectory can be systematically found for any instance of the problem Eq.(10.34). This is the purpose of the following lemma.

Lemma 45. *Consider iteration Eq.(10.68), where the parameters* $\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{V}, \hat{\mathbf{V}}$ *are initialized at any fixed point of the state evolution equations of Lemma 43. For any sequence initialized with* $\hat{\mathbf{V}}_0 = \hat{\mathbf{V}}$ *and* $\mathbf{u}^0$ *such that*

$$\lim_{d\to\infty} \frac{1}{d}\mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0) = \mathbf{Q} \tag{10.79}$$

the following holds

$$\lim_{t\to\infty}\lim_{d\to\infty} \frac{1}{\sqrt{d}}\left\|\mathbf{u}^t - \mathbf{u}^\star\right\|_{\mathrm{F}} = 0 \quad \lim_{t\to\infty}\lim_{d\to\infty} \frac{1}{\sqrt{d}}\left\|\mathbf{v}^t - \mathbf{v}^\star\right\|_{\mathrm{F}} = 0 \tag{10.80}$$

Proof. The proof of Lemma 45 is deferred to subsection 10.5. □

Note that the $\mathbf{G}$ defined here is not the same as the $\mathbf{G}$ in the replica computation. Combining the lemmas 43, 44 and 45 with the pseudo-Lipschitz property, we have reached the following lemma

Lemma 46. *For any fixed* $\mathbf{M}, \hat{\mathbf{M}}, \mathbf{b}$*, consider the fixed point* $(\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{V}, \hat{\mathbf{V}})$ *of the state evolution equations from Lemma. 43. Then, for any sequences of pseudo-Lipschitz functions* $\phi_{1,n} : \mathbb{R}^{d\times K^2} \to \mathbb{R}, \phi_{2,n} : \mathbb{R}^{n\times K} \to \mathbb{R}$*, for* $n, d \to \infty$

$$\phi_{1,n}(\tilde{\mathbf{W}}^\star) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{1,n}\left(R_{\tilde{r},\hat{\mathbf{V}}^{-1}}(\mathbf{H}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1})\right)\right] \tag{10.81}$$

$$\phi_{2,n}(\mathbf{Z}\tilde{\mathbf{W}}^\star) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{2,n}\left(R_{\tilde{L},\mathbf{V}}(\mathbf{G}\mathbf{Q}^{1/2})\right)\right] \tag{10.82}$$

where we remind that $\mathbf{G} = [\mathbf{G}_k]_{k=1}^K, \mathbf{H} = [\mathbf{H}_k]_{k=1}^K$ *are block diagonal i.i.d. standard normal matrices as in Lemma 43, and* $\mathbf{Q} = [\mathbf{Q}_k]_{k=1}^K$ $\hat{\mathbf{Q}} = \left[\hat{\mathbf{Q}}_k\right]_{k=1}^K$ *are the* $K^2 \times K^2$ *block diagonal covariances.*

Proof. Lemma 46 is a consequence of Lemmas 43,44,45 and applying the pseudo-Lipschitz property along with the fact that the iterates of the AMP have bounded norm using the state evolution and that the estimator also has bounded norm (feasibility assumption). Note that, for a generically non-strictly convex problem, being close to the zero gradient condition does not guarantee being close to the estimator. This is further discussed in Appendix 10.4. □

Note that the resolvents are implicitly acting on the block diagonals of their arguments. At this point we are quite close to Theorem 17(details for the exact matching will be given later), but we are missing the equations on $\mathbf{M}, \hat{\mathbf{M}}, \mathbf{b}$.

Fixed point equations for $\mathbf{M}, \hat{\mathbf{M}}, \mathbf{b}$ We drop the dependence on the bias term $\mathbf{b}$ as its solution is very similar to the one for $\mathbf{M}, \hat{\mathbf{M}}$. To obtain the equations for $\mathbf{M}, \hat{\mathbf{M}}$, we go back to the complete optimization problem

$$\inf_{\mathbf{M},\tilde{\mathbf{W}},\mathbf{b}} \sup_{\hat{\mathbf{M}}} L(\mathbf{Y}_k\mathbf{M} + \mathbf{Z}_k\tilde{\mathbf{W}}_k) + r\left(\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right) + \mathrm{tr}\left(\hat{\mathbf{M}}^\top\left(\mathbf{M} - \frac{1}{\sqrt{d}}\mathbf{M}\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right)\right) \tag{10.83}$$

where we can use strong duality to write the equivalent form

$$\inf_{\mathbf{M},\mathbf{b}} \sup_{\hat{\mathbf{M}}} L(\mathbf{Y}_k\mathbf{M} + \mathbf{Z}_k\tilde{\mathbf{W}}_k^\star) + r\left(\mathbf{\Sigma}^+\tilde{\mathbf{W}}\right) + \mathrm{tr}\left(\hat{\mathbf{M}}^\top\left(\mathbf{M} - \frac{1}{\sqrt{d}}\mathbf{M}\mathbf{\Sigma}^+\tilde{\mathbf{W}}^\star\right)\right) \tag{10.84}$$

The gradients w.r.t. $\mathbf{M}, \hat{\mathbf{M}}$ then read:

$$\partial\hat{\mathbf{M}} = \mathbf{M} - \frac{1}{\sqrt{d}}\mathbf{M}\boldsymbol{\Sigma}^{+}\tilde{\mathbf{W}}^{\star} \tag{10.85}$$

$$\partial\mathbf{M} = \hat{\mathbf{M}} + \partial_{\mathbf{M}} L(\mathbf{Y}\mathbf{M} + \mathbf{Z}\tilde{\mathbf{W}}^{\star}) \tag{10.86}$$

Uniform convergence of derivatives and conditions for the dominated convergence theorem are verified using similar arguments as in [176, Lemma 12]. We can thus invert limits and derivatives, and expectations and derivatives. To facilitate taking the derivative $\partial_{\mathbf{M}}$, we use Lemma 46 (assuming the normalized loss function is pseudo-Lipschitz, which is a very loose assumption verified by most machine learning losses) to obtain, reintroducing the scaling $1/d$

$$\frac{1}{d}L(\mathbf{Y}\mathbf{M} + \mathbf{Z}\tilde{\mathbf{W}}^{\star}) \xrightarrow[d\to\infty]{P} \frac{1}{d}\mathbb{E}\left[L(\mathbf{Y}\mathbf{M} + \mathbf{R}_{\tilde{L},\mathbf{V}}(\mathbf{G}\mathbf{Q}^{1/2}))\right] \tag{10.87}$$

Using the block decomposition from Eq.(10.40), the blocks $(\mathbf{R}_{\tilde{L},\mathbf{V}}(\mathbf{G}\mathbf{Q}^{1/2}))_k \in \mathbb{R}^{n_k \times K}$ are given by:

$$\underset{\mathbf{T}\in\mathbb{R}^{n_k\times K}}{\arg\min}\left\{L_k(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T} - (\mathbf{Y}_k\mathbf{M} + \mathbf{G}_k\mathbf{Q}_k^{1/2}))\mathbf{V}_k^{-1}(\mathbf{T} - (\mathbf{Y}_k\mathbf{M} + \mathbf{G}_k\mathbf{Q}_k^{1/2}))^{\top}\right)\right\} - \mathbf{Y}_k\mathbf{M} \tag{10.88}$$

Using a block diagonal representation, we can write:

$$\begin{aligned}\frac{1}{d}L(\mathbf{Y}\mathbf{M} + R_{\tilde{L},\mathbf{V}}(\mathbf{G}\mathbf{Q}^{1/2})) &= \frac{1}{d}L(R_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2})) \\ &= \frac{1}{d}\mathcal{M}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}) - \\ \frac{1}{2d}\mathrm{tr}\left((\mathbf{R}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}) - (\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}))\mathbf{V}^{-1}(\mathbf{R}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}) - (\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}))^{\top}\right)\end{aligned} \tag{10.89}$$

where we have introduced the Bregman-envelope [26] with respect to the distance Eq. (10.4)

$$\begin{aligned}&\mathcal{M}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}) = \\ &\min_{\mathbf{T}}\left\{L(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T} - (\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}))\mathbf{V}^{-1}(\mathbf{T} - (\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}))^{\top}\right)\right\}\end{aligned} \tag{10.90}$$

Then, using the state evolution equations from Lemma 43 and Stein's lemma, we can write:

$$\frac{1}{d}L(\mathbf{Y}\mathbf{M} + \mathbf{R}_{\tilde{L},\mathbf{V}}(\mathbf{G}\mathbf{Q}^{1/2})) = \frac{1}{d}\mathcal{M}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}) - \frac{1}{2}\mathrm{tr}(\mathbf{V}^{\top}\mathbf{Q}) \tag{10.91}$$

Taking the gradient w.r.t. $\mathbf{M}$ using the expression for the derivative of a Bregman envelope [26], we get:

$$\partial_{\mathbf{M}}L(\mathbf{Y}\mathbf{M} + \mathbf{R}_{\tilde{L},\mathbf{V}}(\mathbf{G}\mathbf{Q}^{1/2})) = \frac{1}{d}\mathbf{Y}^{\top}\left(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2} - \mathbf{R}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2})\right)\mathbf{V}^{-1} \tag{10.92}$$

which prescribes, using Lemma 46

$$\hat{\mathbf{M}} \overset{\mathrm{P}}{\simeq} \frac{1}{d}\mathbf{Y}^{\top}\left(\mathbf{R}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}) - \mathbf{Y}\mathbf{M} + \mathbf{G}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1} \tag{10.93}$$

For $\mathbf{M}$, we use the block decomposition from Eq.(10.54), which simplifies the pseudo-inverse $\mathbf{\Sigma}^+$ in Eq. (10.85) to give, using Lemma 46 again

$$\mathbf{M} \overset{\text{P}}{\simeq} \frac{1}{\sqrt{d}} \mathbf{M}\boldsymbol{\eta}(\mathbf{H}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}) \tag{10.94}$$

where the function $\boldsymbol{\eta}$ acts on the block diagonal and is defined by Eq.(10.55). Using those results and the definition of $\tilde{\mathbf{W}}$, the solution $\mathbf{W}^\star$ and the quantity $\mathbf{X}\mathbf{W}^\star$ are characterized, in the pseudo-Lipschitz sense of Theorem 17, by the fixed point of the system of equations (the first four equations are meant for all $1 \leqslant k \leqslant K$):

$$\mathbf{Q}_k = \lim_{d\to+\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{e}_k(\{\mathbf{H}_k(\hat{\mathbf{Q}}_k)^{1/2}\hat{\mathbf{V}}_k^{-1}\}_{k\in[K]})^\top \mathbf{e}_k(\{\mathbf{H}_k(\hat{\mathbf{Q}}_k)^{1/2}\hat{\mathbf{V}}_k^{-1}\}_{k\in[K]})\right] \in \mathbb{R}^{K\times K} \tag{10.95}$$

$$\hat{\mathbf{Q}}_k = \lim_{d\to+\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{h}_k(\mathbf{G}_k\mathbf{Q}_k^{1/2})^\top \mathbf{h}_k(\mathbf{G}_k\mathbf{Q}_k^{1/2})\right] \in \mathbb{R}^{K\times K} \tag{10.96}$$

$$\mathbf{V}_k = \lim_{d\to+\infty} \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\partial \mathbf{e}_k(\{\mathbf{H}_k(\hat{\mathbf{Q}}_k)^{1/2}\}_{k\in[K]})}{\partial(\mathbf{H}_k(\hat{\mathbf{Q}}_k)^{1/2})_i}\right] \in \mathbb{R}^{K\times K} \tag{10.97}$$

$$\hat{\mathbf{V}}_k = -\lim_{d\to+\infty} \frac{1}{d}\sum_{i=1}^{n_k}\mathbb{E}\left[\frac{\partial \mathbf{h}_{k,t}(\mathbf{G}_k(\mathbf{Q}_{k,t})^{1/2})}{\partial(\mathbf{G}_k(\mathbf{Q}_k)^{1/2})_i}\right] \in \mathbb{R}^{K\times K} \tag{10.98}$$

$$\mathbf{M} = \frac{1}{\sqrt{d}}\mathbb{E}\left[\mathbf{M}\boldsymbol{\eta}(\mathbf{H}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1})\right] \in \mathbb{R}^{K\times K} \tag{10.99}$$

$$\hat{\mathbf{M}} = \frac{1}{d}\mathbf{Y}^\top\left(\mathbf{R}_{L,\mathbf{V}}(\mathbf{Y}\mathbf{M}+\mathbf{G}\mathbf{Q}^{1/2}) - \mathbf{Y}\mathbf{M}+\mathbf{G}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1} \in \mathbb{R}^{K\times K} \tag{10.100}$$

Using the explicit form of the different functions given in section 10.3 and Stein's lemma for the derivatives, these equations match those of Theorem 17. This completes the proof.

On the strict convexity assumption If the optimization problem defining $\mathbf{W}^\star$ is strictly convex, there is only one minimizer and the provided proof is enough. Additionally it is shown in [284] that for any loss function that is strictly convex in its argument and penalized with the ℓ_1 norm, provided the data is sampled from a continuous distribution, the solution is unique with probability one regardless of the rank of the design matrix. Thus finding a point verifying the optimality condition of (10.34) is also enough to complete the proof. For generic convex (non-strictly) problems a more careful analysis could be performed in the same spirit as the one of [29]. Empirically the result still holds.

On the uniqueness of the solution to the fixed point equations (10.95) It is possible to reconstruct Bregman envelopes on problem (10.34) for the loss and regularization as we have done for the loss in the previous section. We can then show that the fixed point equations (10.95) are the optimality condition of a convex-concave problem involving both Bregman envelopes and linear combinations of the order parameters. In the same spirit as [57, 176], this problem should be asymptotically strictly convex. This is supported by the simulations presented in the experiments sections but left as an assumption in the main paper.

10.5 Proof of Lemma 45

This proof follows a similar argument to the one used to control the trajectory of the AMP studied in [42]. Note that, because of the way the AMP is initialized using the fixed point of the state

evolution equations, for any $t \geqslant 1$ the following holds:

$$\lim_{d\to+\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{e}(\mathbf{u}^t)^\top \mathbf{e}(\mathbf{u}^t)\right] \overset{w.h.p}{=} \mathbf{Q} \in \mathbb{R}^{K^2\times K^2} \tag{10.101}$$

$$\lim_{d\to+\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{h}(\mathbf{v}^t)^\top \mathbf{h}(\mathbf{v}^t)\right] \overset{w.h.p}{=} \hat{\mathbf{Q}} \in \mathbb{R}^{K^2\times K^2} \tag{10.102}$$

where

$$\mathbf{e}(\mathbf{u}^t) = (Id + \partial\tilde{r}(\bullet)\hat{\mathbf{V}}^{-1})^{-1}(\mathbf{u}^t\hat{\mathbf{V}}^{-1}) \quad \mathbf{h}(\mathbf{v}^t) = \left(\left(Id + \partial\tilde{L}(\bullet)\mathbf{V}\right)^{-1}(\mathbf{v}^t) - \mathbf{v}^t\right)\mathbf{V}^{-1} \tag{10.103}$$

then the limit we are looking for reads:

$$\begin{aligned}\lim_{d\to\infty} \frac{1}{d}\left\|\mathbf{u}^t - \mathbf{u}^{t-1}\right\|_F^2 &= \lim_{d\to\infty} 2(\hat{\mathbf{Q}} - \frac{1}{d}\mathrm{tr}((\mathbf{u}^t)^\top\mathbf{u}^{t-1}) \\ \lim_{d\to\infty} \frac{1}{d}\left\|\mathbf{v}^t - \mathbf{v}^{t-1}\right\|_F^2 &= 2(\mathbf{Q} - \frac{1}{d}\mathrm{tr}((\mathbf{v}^t)^\top\mathbf{v}^{t-1})\end{aligned} \tag{10.104}$$

We thus need to study the correlation between successive iterates. At each time step, denote $(\hat{\mathbf{C}}_t, \mathbf{C}_t)$ in $\mathbb{R}^{K^2\times K^2}$ the correlation matrices between iterates at times $t, t-1$ describing the Gaussian fields respectively associated to $\mathbf{u}^t, \mathbf{v}^t$ i.e.,

$$\lim_{d\to\infty} \frac{1}{d}\mathrm{tr}((\mathbf{u}^t)^\top\mathbf{u}^{t-1} = \hat{\mathbf{C}}_t \quad \lim_{d\to\infty} \frac{1}{d}\mathrm{tr}((\mathbf{v}^t)^\top\mathbf{v}^{t-1} = \mathbf{C}_t \tag{10.105}$$

we can then write the block diagonal Gaussian fields $\hat{\mathbf{Z}}^t, \hat{\mathbf{Z}}^{t-1}, \mathbf{Z}^t, \mathbf{Z}^{t-1}$ in $\mathbb{R}^{Kd\times K^2}$ and in the following way

$$\hat{\mathbf{Z}}^t \sim \mathbf{H}(\hat{\mathbf{C}}_t)^{1/2} + \mathbf{H}^{'}(\hat{\mathbf{Q}} - \hat{\mathbf{C}}_t)^{1/2} \tag{10.106}$$

$$\hat{\mathbf{Z}}^{t-1} \sim \mathbf{H}(\hat{\mathbf{C}}_t)^{1/2} + \mathbf{H}^{''}(\hat{\mathbf{Q}} - \hat{\mathbf{C}}_t)^{1/2} \tag{10.107}$$

$$\mathbf{Z}^t \sim \mathbf{G}(\mathbf{C}_t)^{1/2} + \mathbf{G}^{'}(\mathbf{Q} - \mathbf{C}_t)^{1/2} \tag{10.108}$$

$$\mathbf{Z}^{t-1} \sim \mathbf{G}(\mathbf{C}_t)^{1/2} + \mathbf{G}^{''}(\mathbf{Q} - \mathbf{C}_t)^{1/2} \tag{10.109}$$

where the matrices $\mathbf{H}, \mathbf{H}', \mathbf{H}''$ are in $\mathbb{R}^{Kd\times K^2}$, $\mathbf{G}, \mathbf{G}', \mathbf{G}''$ are in $\mathbb{R}^{n\times K^2}$ and all have i.i.d. standard normal elements. The recursion describing the evolution of these correlations then reads :

$$\mathbf{C}_{t+1} = \frac{1}{d}\mathbb{E}\left[\mathbf{e}(\mathbf{H}\hat{\mathbf{C}}_t^{1/2} + \mathbf{H}^{'}(\hat{\mathbf{Q}} - \hat{\mathbf{C}}_t)^{1/2})^\top \mathbf{e}(\mathbf{H}\hat{\mathbf{C}}_t^{1/2} + \mathbf{H}^{''}(\hat{\mathbf{Q}} - \hat{\mathbf{C}}_t)^{1/2})\right] \tag{10.110}$$

$$\hat{\mathbf{C}}_t = \frac{1}{d}\mathbb{E}\left[\mathbf{h}(\mathbf{G}\mathbf{C}_t^{1/2} + \mathbf{G}^{'}(\mathbf{Q} - \mathbf{C}_t)^{1/2})^\top \mathbf{h}(\mathbf{G}\mathbf{C}_t^{1/2} + \mathbf{G}^{''}(\mathbf{Q} - \mathbf{C}_t)^{1/2})\right] \tag{10.111}$$

Integrating out the independent $\mathbf{H}^{'}, \mathbf{H}^{''}$ first, we get

$$\mathbf{C}_{t+1} = \int_{\mathbb{R}^{Kd\times K^2}} d\mu(\mathbf{H})\mathbf{I}(H)^\top\mathbf{I}(\mathbf{H}) \tag{10.112}$$

where $\mathbf{I}(\mathbf{H}) = \int_{\mathbb{R}^{Kd\times K^2}} d\mu(\mathbf{H}^{'})\mathbf{e}(\mathbf{H}\hat{\mathbf{C}}_t^{1/2} + \mathbf{H}^{'}(\hat{\mathbf{Q}} - \hat{\mathbf{C}}_t)^{1/2})$. So $\mathbf{C}^t$ is symmetric positive definite, assuming the resolvents aren't trivial. The same argument applied to $\hat{\mathbf{C}}^t$ shows it is also symmetric positive definite. From [24], the operators

$$(Id + \partial\tilde{r}(\bullet)\hat{\mathbf{V}}^{-1})^{-1}(\bullet) \quad \left(Id + \partial\tilde{L}(\bullet)\mathbf{V}\right)^{-1}(\bullet) \tag{10.113}$$

are *D-firm* w.r.t. the Bregman distances induced by the differentiable, strictly convex functions $\frac{1}{2}\text{tr}(X\hat{\mathbf{V}}X^\top)$ and $\frac{1}{2}\text{tr}(\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^\top)$ respectively. Recall

$$\mathbf{e}(\mathbf{u}^t) = (Id + \partial\tilde{r}(\bullet)\hat{\mathbf{V}}^{-1})^{-1}(\mathbf{u}^t\hat{\mathbf{V}}^{-1}) \quad \mathbf{h}(\mathbf{v}^t) = \left(\left(Id + \partial\tilde{L}(\bullet)\mathbf{V}\right)^{-1}(\mathbf{v}^t) - \mathbf{v}^t\right)\mathbf{V}^{-1} \tag{10.114}$$

Then, using the definition of *D-firm*

$$\langle \mathbf{e}(\hat{\mathbf{Z}}^t) - \mathbf{e}(\hat{\mathbf{Z}}^{t-1}), \left(\mathbf{e}(\hat{\mathbf{Z}}^t) - \mathbf{e}(\hat{\mathbf{Z}}^{t-1})\right)\hat{\mathbf{V}}\rangle \leqslant \langle \mathbf{e}(\hat{\mathbf{Z}}^t) - \mathbf{e}(\hat{\mathbf{Z}}^{t-1}), (\hat{\mathbf{Z}}^t - \hat{\mathbf{Z}}^{t-1})\hat{\mathbf{V}}^{-1}\hat{\mathbf{V}}\rangle \tag{10.115}$$

Adding the normalization by $\frac{1}{d}$, using the representation Eq.(10.106-10.109), taking expectations and applying the matrix form of Stein's lemma, see for example [110] Lemma 12, we get:

$$\text{tr}((\mathbf{Q} - \mathbf{C}_{t+1})\hat{\mathbf{V}}) \leqslant \text{tr}((\hat{\mathbf{Q}} - \hat{\mathbf{C}}_t)\mathbf{V}) \tag{10.116}$$

Using a similar argument on $\mathbf{h}$, we get

$$\text{tr}((\hat{\mathbf{Q}} - \hat{\mathbf{C}}_t)\mathbf{V}) \leqslant \text{tr}((\mathbf{Q} - \mathbf{C}_t)\hat{\mathbf{V}}) \tag{10.117}$$

and

$$\text{tr}(\mathbf{C}_{t+1}\hat{\mathbf{V}}) \geqslant \text{tr}(\mathbf{C}_t\hat{\mathbf{V}}) \tag{10.118}$$

thus the sequence $\text{tr}(\mathbf{C}_{t+1}\hat{\mathbf{V}})$ is a bounded (above) monotone (increasing) sequence, and therefore converges. Since $\hat{\mathbf{V}}$ is positive definite and given the iteration defining $\mathbf{C}_{t+1}$ from $\mathbf{C}_t$, any fixed point of this iteration is a fixed point of $\text{tr}(\mathbf{C}_t\hat{\mathbf{V}})$. Assuming there is only one fixed point to the set of self-consistent equations Eq.(9.8) (see previous section), the proof is complete. (A similar argument can be carried out on $\hat{\mathbf{C}}_t$).

Chapter 11

Fluctuations, Bias, Variance & Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension

The results in this chatper are based on the publication [177]. From the sampling of data to the initialisation of parameters, randomness is ubiquitous in modern Machine Learning practice. Understanding the statistical fluctuations engendered by the different sources of randomness in prediction is therefore key to understanding robust generalisation. In this manuscript we develop a quantitative and rigorous theory for the study of fluctuations in an ensemble of generalised linear models trained on different, but correlated, features in high-dimensions. In particular, we provide a complete description of the asymptotic joint distribution of the empirical risk minimiser for generic convex loss and regularisation in the high-dimensional limit. Our result encompasses a rich set of classification and regression tasks, such as the lazy regime of overparametrised neural networks, or equivalently the random features approximation of kernels. While allowing to study directly the mitigating effect of ensembling (or bagging) on the bias-variance decomposition of the test error, our analysis also helps disentangle the contribution of statistical fluctuations, and the singular role played by the interpolation threshold that are at the roots of the "double-descent" phenomenon.

11.1 Introduction

Randomness is ubiquitous in Machine Learning. It is present in the data (e.g., noise in acquisition and annotation), in commonly used statistical models (e.g., random features [238]), or in the algorithms used to train them (e.g., in the choice of initialisation of weights of neural networks [210], or when sampling a mini-batch in Stochastic Gradient Descent [46]). Strikingly, fluctuations associated to different sources of randomness can have a major impact in the generalisation performance of a model. For instance, this is the case in least-squares regression with random features, where it has been shown [104, 71, 132] that the variance associated with the random projections matrix is responsible for poor generalisation near the interpolation peak [4, 270, 32]. As a consequence, this *double-descent* behaviour can be mitigated by averaging over a large *ensemble* of learners, effectively suppressing this variance. Indeed, considering an ensemble (sometimes also refereed to as

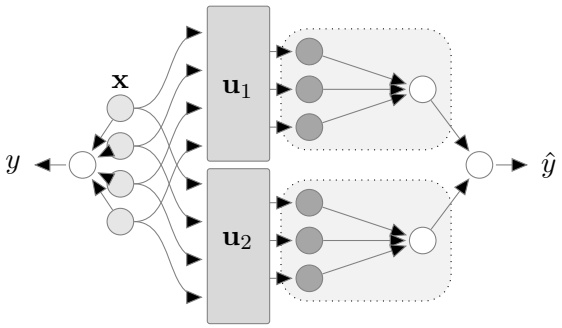


Figure 11.1: Pictorial representation of the model considered in the paper for $K = 2$. Two learners with the same architecture (in gray) receive a correlated input generated from the same vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. The output $\hat{y}$ is an average of their outputs. While the study of an ensemble of learners is already interesting *per se*, it is also pivotal to study the fluctuation between learners, and the error steaming from the difference in the weights in random features and lazy training.

a committee [86]) of independent learners provide a natural framework to study the contribution of the variance of prediction in the estimation accuracy. In this manuscript we leverage this idea to provide an exact asymptotic characterisation of the statistics of fluctuations in empirical risk minimisation with generic convex losses and penalties in high-dimensional models. We focus on the case of synthetic datasets, and we apply our results to random feature learning in particular.

11.1.1 Setting

Let $(\mathbf{x}^\mu, y^\mu) \in \mathbb{R}^d \times \mathcal{Y}$, $\mu \in [n] \coloneqq \{1, \dots, n\}$, denote a labelled data set composed of n independent samples from a joint density $p(\mathbf{x}, y)$ (e.g., $\mathcal{Y} = \{-1, 1\}$ for a binary classification problem). In this manuscript we are interested in studying an ensemble of K parametric predictors, each of them depending on a vector of parameters $\mathbf{w}_k \in \mathbb{R}^p$, $k \in [K]$, and independently trained on the dataset $\{(\mathbf{x}^\mu, y^\mu)\}_{\mu \in [n]}$. Note that even if the vectors of parameters $\{\mathbf{w}_k\}_{k \in [K]}$ are trained independently, they correlate through the training data. Statistical fluctuations in the learnt parameters can then arise for different reasons. For instance, a common practice is to initialise the parameters randomly during optimisation, which will induce statistical variability between the different predictors. Alternatively, each predictor could be trained on a subsample of the data, as it is commonly done in bagging [51]. The statistical model can also be inherently stochastic, e.g., the random features approximation for kernel methods [238]. Finally, the predictors could also be jointly trained, e.g., coupling them through the loss or penalty as it is done in boosting [254].

Our goal in this work is to provide a sharp characterisation of the statistical fluctuations of the ensemble of parameters $\{\mathbf{w}_k\}_{k \in [K]}$ in a particular, mathematically tractable, class of predictors: *generalised linear models*,

$$\hat{y}(\mathbf{x}) = \hat{f}\left(\frac{\hat{\mathbf{w}}_1^\top \mathbf{u}_1(\mathbf{x})}{\sqrt{p}}, \dots, \frac{\hat{\mathbf{w}}_K^\top \mathbf{u}_K(\mathbf{x})}{\sqrt{p}}\right) \tag{11.1}$$

where $\mathbf{u}_k : \mathbb{R}^d \to \mathbb{R}^p$, $k \in [K]$ is an ensemble of possibly correlated features and $\hat{f} \colon \mathbb{R}^K \to \mathcal{Y}$ is an activation function. For most of this work, we discuss the case in which the predictors are *independently trained* through regularised empirical risk minimisation:

$$\hat{\mathbf{w}}_k = \underset{\mathbf{w} \in \mathbb{R}^p}{\arg\min} \left[\frac{1}{n} \sum_{\mu=1}^{n} \ell\left(y^\mu, \frac{\mathbf{w}^\top \mathbf{u}_k(\mathbf{x}^\mu)}{\sqrt{p}}\right) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2\right] \tag{11.2}$$

with a convex loss function $\ell: \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ (e.g., the logistic loss) and ridge penalty whose strength is given by $\lambda \in \mathbb{R}^+$. However, our analysis also includes the case in which the learners are jointly trained with a generic convex penalty. This case will be further discussed in Sec. 11.4.

In what follows we will also concentrate in the random features case where $\mathbf{u}_k(\mathbf{x}) = \phi(\mathbf{F}_k\mathbf{x})$ with $\phi : \mathbb{R} \to \mathbb{R}$ an activation function acting component-wise and $\mathbf{F}_k \in \mathbb{R}^{p\times d}$ a family of independently sampled random matrices. Besides being an efficient approximation for kernels [238], random features are often studied as a simple model for neural networks in the lazy and neural tangent kernel regimes of deep neural networks [64, 131], in which case the matrices $\mathbf{F}_k$ correspond to different random initialisation of hidden-layer weights. Moreover, the random features model displays some of the exotic behaviours of high-dimensional overparametrised models, such as double-descent [191, 107] and benign overfitting [21], therefore providing an ideal playground to study the interplay between fluctuations and overparametrisation. A broader class of features maps is also discussed in Sec. 11.4.

To provide an exact characterisation of the statistics of the estimators in eq. (11.2), we shall assume data is generated from a target

$$y = f_0\left(\frac{\boldsymbol{\theta}^\top \mathbf{x}}{\sqrt{d}}\right), \qquad \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_d, \rho\mathbf{I}_d), \qquad \rho \in \mathbb{R}_0^+, \tag{11.3}$$

with $f_0 : \mathbb{R} \to \mathcal{Y}$ and $\mathbf{I}_d$ d-dimensional identity matrix. The dataset is then constructed generating *i.i.d.* n vectors $\mathbf{x}^\mu \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, $\mu \in [n]$.

An illustration summary of the setting considered here in given in Figure 11.1. Note that such architecture can be interpreted as a two-layer tree neural network, also known in some contexts as the *tree-committee* or *parity machine* [259].

Main contributions — The results in this manuscript can be listed as follows.

- We provide a sharp asymptotic characterisation of the joint statistics of the ensemble of empirical risk minimisers $\{\hat{\boldsymbol{w}}_k\}_{k\in[K]}$ in the high-dimensional limit where $p, n \to +\infty$ with n/p kept constant, for any convex loss and penalty. In particular, we show that the pre-activations $\{\hat{\boldsymbol{w}}_k^\top \boldsymbol{u}_k\}_{k\in[K]}$ are jointly Gaussian, with sufficient statistics obeying a set of explicit closed-form equations. Note that the analysis of ensembling with non-square losses is out of the grasp of the most commonly adopted theoretical tools (e.g., random matrix theory). Therefore, our proof method based on recent progress on Approximate Message Passing techniques [135, 37, 110] is of independent interest. Different versions of our theorem are discussed throughout the manuscript. First, in Sec. 11.2 for the particular case of independently trained learners on random features (Theorem 20). Later, in Sec. 11.4 for the general case of jointly trained learners on correlated Gaussian covariates (Theorem 21).

- We discuss the role played by fluctuations in the non-monotonic behaviour of the generalisation performance of interpolators (a.k.a. double-descent behaviour). In particular —as discussed in [104, 72] for the ridge case— the interpolation peak arises from the model overfitting the particular realisation of the random weights. We show the test error can be decomposed $\epsilon_g(K=1) = \overline{\epsilon_g} + \delta\epsilon_g$ in terms of a fluctuation-free term $\overline{\epsilon_g}$ and a fluctuation term $\delta\epsilon_g$ responsible for the double-descent behavior, see Fig. 11.2 for the case of max-margin classification.

- In the context of classification, we discuss how *majority vote* and *score averaging*, two popular ensembling procedures, compare in terms of generalisation performance. More specifically, we show

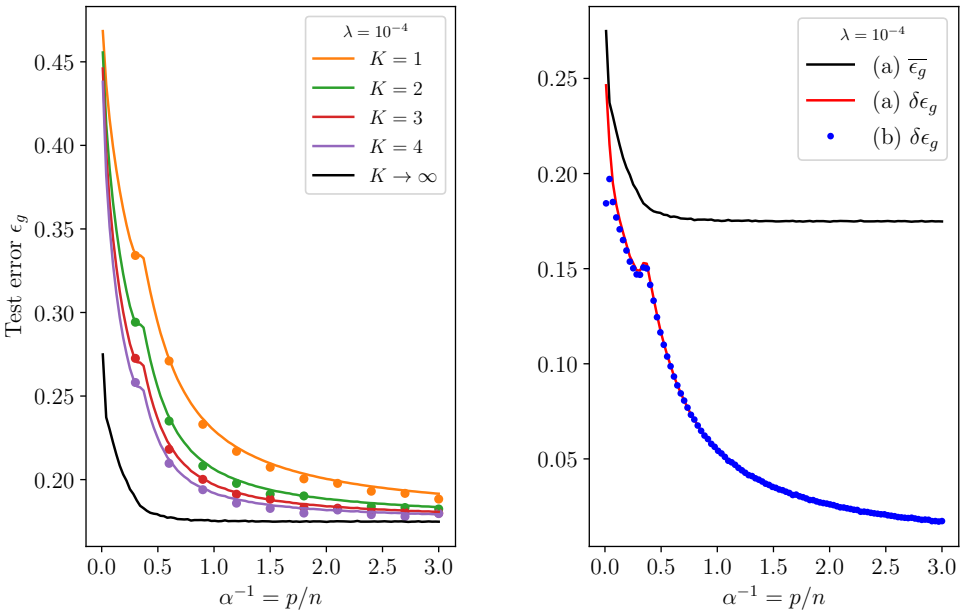


Figure 11.2: *Left.* Test error for logistic regression with $\lambda = 10^{-4}$ and different values of K as function of $p/n = 1/\alpha$ with $n/d = 2$ and $\rho = 1$. Dots represent the average of the outcomes of 10^3 numerical experiments. Here we adopted $\phi(x) = \mathrm{erf}(x)$ and estimator $\hat{f}(\mathbf{v}) = \mathrm{sign}(\sum_k v_k)$. *Right.* Decomposition of the $K = 1$ test error $\epsilon_g = \overline{\epsilon_g} + \delta\epsilon_g$ for the estimator **(a)**, with $n/d = 2$ and $\lambda = 10^{-4}$. We plot also the contribution $\delta\epsilon_g$ corresponding to the estimator **(b)**: we numerically observed that such decomposition coincides in the two cases. Note also the presence of a kink in $\delta\epsilon_g$ at the interpolation transition.

that in the setting we study score averaging consistently outperforms the majority vote predictor. However, for a large number of learners $K \gg 1$ these two predictors agree, see Fig. 11.5 (right).

- Finally, we discuss how ensembling can be used as a tool for uncertainty quantification. In particular, we connect the correlation between two learners to the probability of disagreement, and show that it decreases with overparametrisation, see Fig. 11.5 (center). We provide a full characterisation of the joint probability density of the confidence score between two independent learners, see Fig. 11.5 (left).

Related works — The idea of reducing the variance of a predictor by averaging over independent learners is quite old in Machine Learning [124, 233, 232, 152], and an early asymptotic analysis of the regression case was given in [151]. In particular, a variety of methods to combine an ensemble of learners appeared in the literature [216]. In a very inspiring work, [104] carried out an extensive series of experiments in order to shed light on the generalisation properties of neural networks, and reported many observations and empirical arguments about the role of the variance due to the random initialisation of the weights in the double-descent curve using an ensemble of learners. This was a major motivation for the present work. Closest to our setting is the work of [211, 71, 132] which disentangles the various sources of variance in the process of training deep neural networks. Indeed, here we adopt the model defined by [71], and provide a rigorous justification of their results for the case of ridge regression. A slightly finer decomposition of the variance in terms of the

different sources of randomness in the problem was later proposed by [2]. [171] show that such decomposition is not unique, and can be more generally understood from the point of view of the *analysis of variance* (ANOVA) framework. Interestingly, subsequent papers were able to identity a series of triple (and more) descent, e.g., [72, 1, 60].

The Random Features (RF) model was introduced in the seminal work of [238] as an efficient approximation for kernel methods. Drawing from early ideas of [141], [231] showed that the empirical distribution of the Gram matrix of RF is asymptotically equivalent to a linear model with matched second statistics, and characterised in this way memorisation with RF regression. The learning problem was first analysed by [191], who provided an exact asymptotic characterisation of the training and generalisation errors of RF regression. This analysis was later extended to generic convex losses by [107] using the heuristic replica method, and later proved by [76] using convex Gaussian inequalities.

The aforementioned asymptotic equivalence between the RF model and a Gaussian model with matched moments has been named the *Gaussian Equivalence Principle* (GEP) [118]. Rigorous proofs in the memorisation and learning setting with square loss appeared in [231, 191], and for general convex penalties in [115, 128]. [115] and [178] provided extensive numerical evidence that the GEP holds for more generic feature maps, including features stemming from trained neural networks.

Most of the previously mentioned works deriving exact asymptotics for the RF model in the proportional limit use either Random Matrix Theory techniques or Convex Gaussian inequalities. While these tools have been recently used in many different contexts, they ultimately fall short when considering an ensemble of predictors with generic convex loss and regularisation, along with structured design matrices. Therefore, to prove the results herein we employ an *Approximate Message Passing* (AMP) proof technique [28, 82], leveraging on recently introduced progresses in [178, 110] which enables to capture the full complexity of the problem and obtain the asymptotic joint distribution of the ensemble of predictors. [162] studies ensembles of ordinary least-squares learned from subsamples of a common data matrix, and shows its equivalence to an implicit ridge regularization.

11.2 Learning with an ensemble of random features

In this section give a first formulation of our main result, namely the exact asymptotic characterisation of the statistics of the ensembling estimator introduced in eq. (11.1). We prove that, in the proportional high dimensional limit, the statistics of the arguments of the activation function in eq. (11.1) is simply given by a multivariate Gaussian, whose covariance matrix we can completely specify. This result holds for any convex loss, any convex regularisation, and for all models of generative networks $\mathbf{u}_k \colon \mathbb{R}^d \to \mathbb{R}^p$, as we will show in full generality in Sec. 11.4. However, for simplicity, in this section and in the following we focus on the setting described in Sec. 11.1, in which the statistician averages over an independent ensemble of random features, i.e., $\mathbf{u}_k(\mathbf{x}) = \phi(\mathbf{F}_k\mathbf{x})$. In this case, our result can be formulated as follows:

Theorem 20 (Simplified version)**.** *Assume that in the high-dimensional limit where $d, p, n \to +\infty$ with $\alpha \coloneqq {}^{n}/_{p}$ and $\gamma \coloneqq {}^{d}/_{p}$ kept $\Theta(1)$ constants, the Wishart matrix $\mathbf{F}\mathbf{F}^\intercal$ has a well-defined asymptotic spectral distribution. Then in this limit, for any pseudo-Lispchitz function of order 2 $\varphi \colon \mathbb{R}\times\mathbb{R}^K \to \mathbb{R}$, we have*

$$\mathbb{E}_{(\mathbf{x},y)}\left[\varphi\left(y, \frac{\hat{\mathbf{w}}_1^\intercal \mathbf{u}_1}{\sqrt{p}}, \dots, \frac{\hat{\mathbf{w}}_K^\intercal \mathbf{u}_K}{\sqrt{p}}\right)\right] \xrightarrow{\text{P}} \mathbb{E}_{(\nu,\boldsymbol{\mu})}\left[\varphi\left(f_0(\nu), \boldsymbol{m}\boldsymbol{u}\right)\right], \tag{11.4}$$

where $(\nu, \boldsymbol{mu}) \in \mathbb{R}^{K+1}$ *is a jointly Gaussian vector* $(\nu, \boldsymbol{mu}) \sim \mathcal{N}(\mathbf{0}_{K+1}, \mathbf{\Sigma})$ *with covariance*

$$\mathbf{\Sigma} = \begin{pmatrix} \rho & m\mathbf{1}_K^\intercal \\ m\mathbf{1}_K & \mathbf{Q} \end{pmatrix}, \qquad \mathbf{Q} \coloneqq (q_0 - q_1)\mathbf{I}_K + q_1 \mathbf{1}_{K,K}, \tag{11.5}$$

with $\mathbf{1}_{K,K} \in \mathbb{R}^{K\times K}$ *and* $\mathbf{1}_K \in \mathbb{R}^K$ *are a matrix and a vector of ones respectively. The entries of* $\mathbf{\Sigma}$ *are solutions of a set of self-consistent equations given in Corollary 5.*

As discussed in the introduction, the asymptotic statistics of the *single* learner has been studied in [107, 76, 178]. Their result amounts to the analysis of the estimator solving the empirical risk minimisation problem in eq. (11.2) and it is recovered imposing $K = 1$ in the theorem above. For $K = 1$, $(\nu, \mu) \in \mathbb{R}^2$ is jointly Gaussian with zero mean and covariance $\mathbf{\Sigma} = \left(\begin{smallmatrix} \rho & m \\ m & q_0 \end{smallmatrix}\right)$.

However, such result is not enough to quantify the correlation between different learners, induced by the training on the same dataset, which is required to compute, e.g., the test error associated with an ensembling predictor as in eq. (11.1). For example, in the simple case where $f_0(u) = u$ and $\hat{f}(\mathbf{v}) = \frac{1}{K}\sum_k v_k$, the mean-squared error on the labels is given by $\epsilon_g = \mathbb{E}_{(\mathbf{x},y)}[(y - \hat{y}(\mathbf{x}))^2] = \rho + (q_0 - q_1)K^{-1} + q_1 - 2m$, which crucially depends on the average correlation between two independent learners[1] $q_1 \coloneqq \frac{1}{p}\mathbb{E}[\hat{\mathbf{w}}_1^\intercal \hat{\mathbf{w}}_2]$. Our main result is precisely an exact asymptotic characterisation of this correlation in the proportional limit of the previous theorem. Once m, q_0 and q_1 have been determined, the generalisation error can be computed as

$$\epsilon_g \coloneqq \mathbb{E}_{(\mathbf{x},y)}[\Delta\left(y, \hat{y}(\mathbf{x})\right)] \xrightarrow{n\to+\infty} \mathbb{E}_{(\nu,\boldsymbol{\mu})}\left[\Delta\left(f_0(\nu), \hat{f}(\boldsymbol{\mu})\right)\right] \tag{11.6}$$

for any error measure $\Delta\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.

Suppose now that

$$\hat{f}(\mathbf{v}) \equiv \hat{f}_0\left(\frac{1}{K}\sum_k v_k\right) \tag{11.7}$$

for some $\hat{f}_0\colon \mathbb{R} \to \mathcal{Y}$ activation function of the single learner. In this case we can introduce the random variable $\hat{\mu} \overset{\mathrm{d}}{=} \lim_{K\to+\infty} \frac{1}{K}\sum_k \mu_k$. It is not difficult to see that the joint probability $p(\nu, \hat{\mu}) \sim \mathcal{N}(\mathbf{0}_2, \hat{\mathbf{\Sigma}})$ where $\hat{\mathbf{\Sigma}} = \left(\begin{smallmatrix} \rho & m \\ m & q_1 \end{smallmatrix}\right)$. This formally coincides with the joint distribution for the activation fields for $K = 1$ [107], but with q_0 replaced by $q_1 \leqslant q_0$. The smaller variance is due to the fact that the fluctuations of the activation fields are averaged out by the ensembling process. The test error in the $K \to +\infty$ limit is then

$$\overline{\epsilon_g} \coloneqq \mathbb{E}_{(\nu,\hat{\mu})}[\Delta(f_0(\nu), \hat{f}_0(\hat{\mu}))], \tag{11.8}$$

so that the fluctuation contribution to the test error for $K = 1$ can be defined as

$$\delta\epsilon_g \coloneqq \mathbb{E}_{(\nu,\mu)}[\Delta(f_0(\nu), \hat{f}_0(\mu))] - \overline{\epsilon_g}. \tag{11.9}$$

The term $\delta\epsilon_g$ is by definition the contribution suppressed by ensembling and corresponds to the *ambiguity* introduced by [152] for the square loss. This contribution expresses the variance in the ensemble and it is responsible for the non-monotonic behaviour in the test error of interpolators, also known as the double-descent behavior.

[1] Note that since all learners are here assumed to be statistically equivalent, their pair-wise correlation is the same on average. In the general case, discussed in Sec. 11.4, the correlation matrix $\mathbf{Q} \in \mathbb{R}^{K\times K}$ can have a more complex structure.

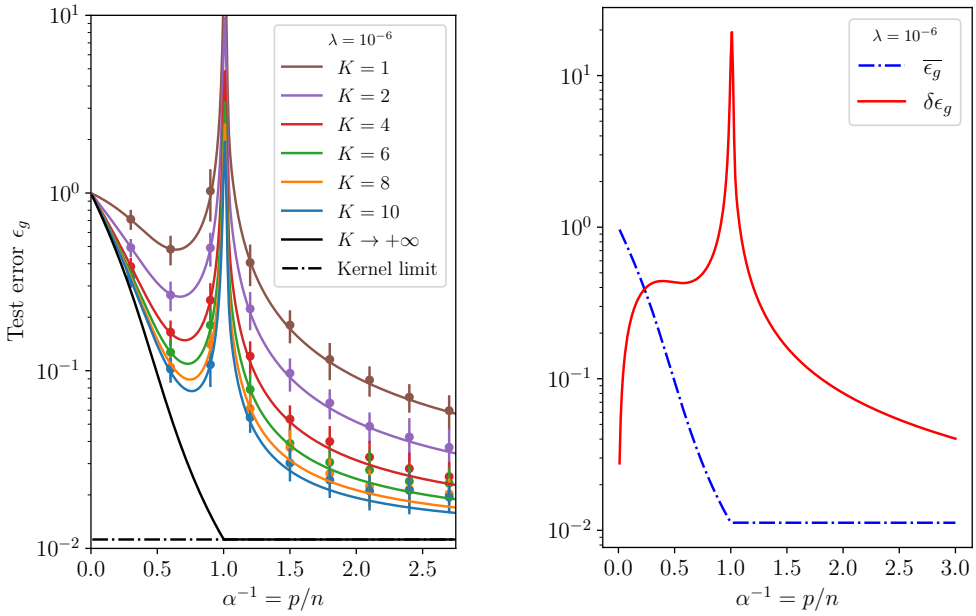


Figure 11.3: *Left.* Test error for ridge regression with $\lambda = 10^{-6}$ and different values of K as function of $p/n = 1/\alpha$ with $n/d = 2$ and $\rho = 1$. Dots represent the average of the outcomes of 50 numerical experiments in which the parameters of the neurons are estimated using $\min(d, p) = 200$. Here we adopted $\phi(x) = \mathrm{erf}(x)$. *Right.* Decomposition of $\epsilon_g = \overline{\epsilon_g} + \delta\epsilon_g$ in the $K = 1$ case

11.3 Applications

We will consider now two relevant examples of separable losses, namely a ridge loss and a logistic loss. In both cases, it is possible to derive the explicit expression of the training loss and generalisation error in terms of the elements of the correlation matrix introduced above.

11.3.1 Ridge regression

If we assume $f_0(x) = x$, $\hat{f}(\mathbf{v}) = \frac{1}{K}\sum_k v_k$, and a quadratic loss of the type $\ell(y, x) = \frac{1}{2}(y - x)^2$, it is possible to write down simple recursive equations for m, q_0 and q_1 (see the appendix of the original paper). Taking $\Delta(y, \hat{y}) = (y - \hat{y})^2$, the generalisation error is easily computed as

$$\epsilon_g = \rho + \frac{q_0 - q_1}{K} + q_1 - 2m \xrightarrow{K\to+\infty} \rho + q_1 - 2m \;\equiv\; \overline{\epsilon_g}. \tag{11.10}$$

Note that in this case the $\lambda \to 0^+$ limit gives the minimum ℓ_2-norm interpolator. In Fig. 11.3 we compare our theoretical prediction with numerical results for $\lambda = 10^{-6}$ and various values of K. It is evident that the divergence of the generalisation error at $\alpha = 1$ is only due to the divergence of q_0, whereas the contribution $\overline{\epsilon_g}$, which is independent on q_0, is smooth everywhere. Alongside with the interpolation divergence, $\delta\epsilon_g = q_0 - q_1$ has an additional bump at $p/n = d/n$, which corresponds to the "linear peak" discussed by [72].

In the plot we present also the so-called kernel limit, corresponding to the limit $n/p = \alpha \to 0$ at fixed n/d. An explicit manipulation (see the appendix of the originaal paper) shows that $q_1 = q_0 \equiv q$ in this limit. This implies that in the kernel limit ϵ_g^{k} does not depend on K, being equal to

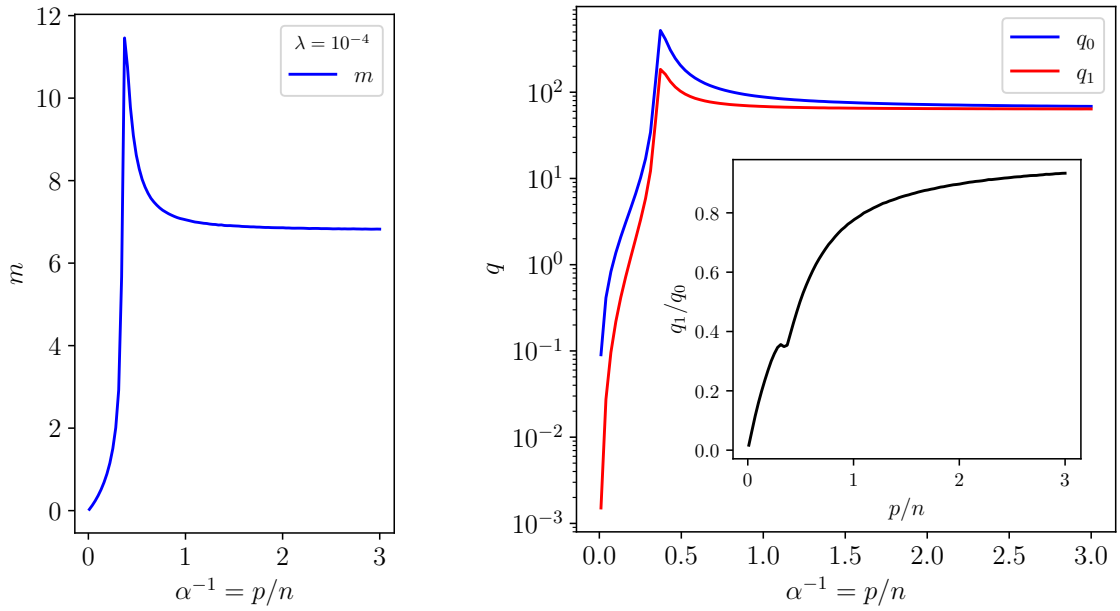


Figure 11.4: Analytical estimation of the covariance parameters characterising the correlation with the oracle m (left), the norm of the predictor in feature space q_0 and the correlation between learners q_1 (right) (see eq. (11.5) for the definition) in a classification task using logistic loss with ridge penalty with $\lambda = 10^{-4}$ at fixed $n/d = 2$ as function of p/n. In the inset, ratio q_1/q_0, quantifying the correlation between two learners. In all parameters the interpolation kink is clearly visible.

$\epsilon_g^{\mathrm{k}} \equiv \rho + q - 2m$. The generalisation error obtained in the kernel limit coincides with $\overline{\epsilon_g}$ for $p > n$: this is expected as in $\overline{\epsilon_g}$ the fluctuations amongst learners are averaged out, effectively recovering the cost obtained in the case of an infinite number of parameters.

11.3.2 Binary classification

Suppose now that we are considering a classification task, such that $\mathcal{Y} = \{-1, 1\}$. For this task we consider $f_0(x) = \mathrm{sign}(x)$. A popular choice of loss in this classification task is the logistic loss,

$$\ell(y, x) = \ln(1 + \mathrm{e}^{-yx}), \tag{11.11}$$

although other choices, e.g, hinge loss, can be considered. Since both the logistic and hinge losses depend only on the *margin* $y\mathbf{w}^\intercal\mathbf{u}$, the empirical risk minimiser for $\lambda \to 0^+$ in both cases give the max-margin interpolator [247]. The explicit saddle-point equations associated to the logistic and hinge loss can be found in the appendix of the original paper, but we will focus our attention on the logistic case for the sake of brevity. For this choice of the loss, we obtained the values of m, q_0 and q_1 showed in Fig. 11.4. Using these values, a number of relevant questions can be addressed.

Alignment of learners

Assuming that the predictor of the learner k is $\hat{y}_k(\mathbf{x}) = \text{sign}(\hat{\mathbf{w}}_k^\intercal \mathbf{u}_k(\mathbf{x}))$, in Fig. 11.5 (center) we estimate the probability that two learners give opposite classification. This is analytically given by

$$\mathbb{P}[\hat{y}_1(\mathbf{x}) \neq \hat{y}_2(\mathbf{x})] = \mathbb{P}[\mu_1\mu_2 < 0] = \frac{1}{\pi}\arccos\left(\frac{q_1}{q_0}\right). \tag{11.12}$$

Note that by definition the ratio q_1/q_0 is a cosine similarity between two learners in the norm induced by the feature space. Therefore, this provides an interesting interpretation of these sufficient statistics in terms of the probability of disagreement. In particular, as illustrated in Fig. 11.5 (center) overparametrisation promotes agreement between the learners, therefore suppressing uncertainty. More generally, ensembling can be used as a technique for uncertainty estimation [156]. In the context of logistic regression, the pre-activation to the sign function is often interpreted as a *confidence score*. Indeed, introducing the logistic function $\varphi_k(\mathbf{x}) = (1+\exp(-p^{-1/2}\hat{\mathbf{w}}_k^\intercal \mathbf{u}_k(\mathbf{x})))^{-1}$, it expresses the confidence of the kth classifier in associating $\hat{y} = 1$ to the input $\mathbf{x}$. Therefore, it is reasonable to ask how reliable is the logistic score as a confidence measure. For instance, what is the variance of the confidence among different learners? This can be quantified by the joint probability density $\rho(\varphi_1, \varphi_2) \coloneqq \mathbb{E}_{\mathbf{x}}[\delta(\varphi_1 - \varphi_1(\mathbf{x}))\delta(\varphi_2 - \varphi_2(\mathbf{x}))]$, which can be readily computed using our Theorem 20. Fig. 11.5 (left) shows one example at fixed p/n and vanishing λ.

Ensemble predictors

In the previous two points, we discussed how ensembling can be used as a tool to quantify fluctuations. However, ensembling methods are also used in practical settings in order to mitigate fluctuations, e.g., [51]. An important question in this context is: given an ensemble of predictors $\{\hat{\mathbf{w}}_k\}_{k\in[K]}$, what is the best way of combining them to produce a point estimate? In our setting, this amounts to choosing the function $\hat{f} : \mathbb{R}^K \to \mathcal{Y}$. Let us consider two popular choices for the estimator $\hat{f}$ in eq. (11.1) used in practice:

$$\textbf{(a)} \qquad \hat{f}(\mathbf{v}) = \text{sign}\left(\sum_k v_k\right), \tag{11.13a}$$

$$\textbf{(b)} \qquad \hat{f}(\mathbf{v}) = \text{sign}\left(\sum_k \text{sign}(v_k)\right). \tag{11.13b}$$

In a sense, **(a)** provides an estimator based on the average of the output fields, whereas **(b)**, which corresponds to a majority rule if K is odd [124], is a function of the average of the estimators of the single learners. For both choices of the estimator we use $\Delta(y, \hat{y}) = \delta_{\hat{y},y}$ to measure the test error. In Fig. 11.5 (right) we compare the test error obtained using **(a)** and **(b)** for $K = 3$ with vanishing regularisation $\lambda = 10^{-4}$. It is observed that the estimator **(a)** has better performances than the estimator **(b)**. As previously discussed, in this case logistic regression is equivalent to max-margin estimation, and in this case the error **(a)** can be intuitively understood in terms of a robust max-margin estimation obtained by averaging the margins associated to different draws of the random features. In the case **(a)** it is easy to show that the generalisation error takes the form

$$\epsilon_g = \frac{1}{\pi}\arccos\left(\frac{\sqrt{K}m}{\sqrt{\rho(q_0 - q_1 + Kq_1)}}\right) \xrightarrow{K\to\infty} \frac{1}{\pi}\arccos\left(\frac{m}{\sqrt{\rho q_1}}\right) \equiv \overline{\epsilon_g}. \tag{11.14}$$

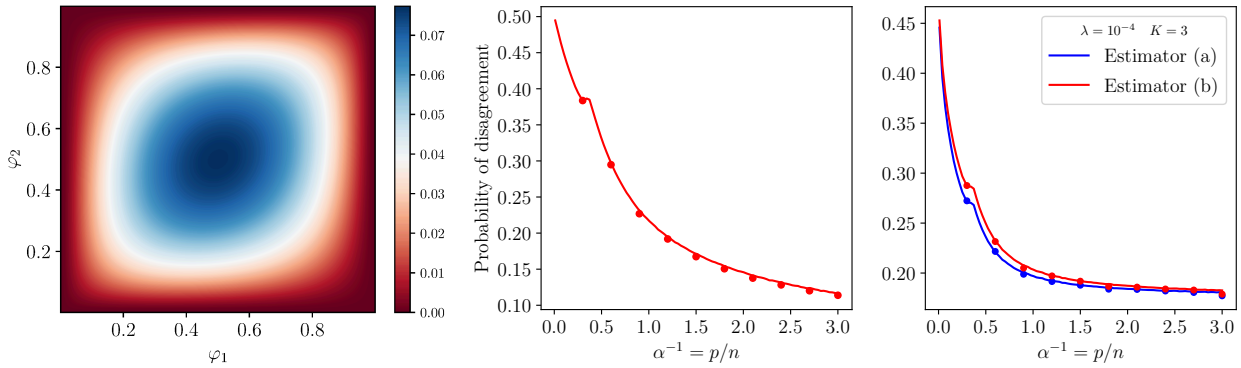


Figure 11.5: *Left.* Joint probability density of the confidence score $\varphi_i(\mathbf{x}) = (1 + \exp(-p^{-1/2}\hat{\mathbf{w}}_i^\intercal \mathbf{u}_i(\mathbf{x})))^{-1}$ of two learners for $p/n \simeq 0.13$. *Center.* Probability that two learners give discordant predictions using logistic regression as function of $p/n = 1/\alpha$ with $n/d = 2$, $\rho = 1$, and $\lambda = 10^{-4}$. *Right.* Test error for logistic regression using the estimators in eq. (11.13) and $K = 3$, with the same parameters. We adopted $\phi(x) = \mathrm{erf}(x)$. We observe that the test error obtained using **(a)** is always smaller than the one obtained using **(b)**. (*Center and right*) Dots represent the average of the outcomes of 10^3 numerical experiments.

This formula is in agreement with numerical experiments, see Fig. 11.2 (left). Unfortunately, we did not find a similar closed-form expression in case **(b)**. However, we can observe that in the $K \to +\infty$ limit the generalisation error in case **(a)** coincides with the generalisation error in case **(b)**, see Fig. 11.2 (right). By comparing with the results in Fig. 11.5 (center), it is evident that the benefit of ensembling in reducing the test error correlates with the tendency of learners to disagree, i.e., for small values of p/n, as stressed by [152]. Finally, we observe a constant value of $\overline{\epsilon_g}$ beyond the interpolation threshold, compatibly with the numerical results of [104].

11.4 The case of general loss and regularisation

In this Section we generalise our results in Sec. 11.2 relaxing the hypothesis on the loss, on the regularisation and on the properties of the feature maps. In the general setting we are going to consider, we denote $P_y^0(y|x)$ the probabilistic law by which y is generated. For example, in Sec. 11.2, $P_y^0(y|x) = \delta(y - f_0(x))$. In the treatment given here, we allow for more general cases (e.g., the presence of noise in the label generation). We make no assumptions on the generative networks $\mathbf{u}_k$, so that the information about the first layer is contained in the following tensors,

$$\mathbf{\Omega} := \mathbb{E}_{\mathbf{x}}[\mathbf{U}(\mathbf{x}) \otimes \mathbf{U}(\mathbf{x})] \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}, \tag{11.15}$$

$$\hat{\mathbf{\Phi}} := \mathbb{E}_{\mathbf{x}}[\mathbf{U}(\mathbf{x})\mathbf{x}^\intercal \boldsymbol{\theta}] \in \mathbb{R}^{p\times K}, \tag{11.16}$$

$$\mathbf{\Theta} = \hat{\mathbf{\Phi}} \otimes \hat{\mathbf{\Phi}} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}. \tag{11.17}$$

In the equations above, $\mathbf{U}(\mathbf{x}) \in \mathbb{R}^{p\times K}$ is the matrix having as concatenated columns $\mathbf{u}_k(\mathbf{x})$. We aim at learning a rule as in eq. (11.1), adopting a general convex loss $\hat{\ell}\colon \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}$, so that the weights are estimated as

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in\mathbb{R}^{p\times K}}{\arg\min} \left[\frac{1}{n} \sum_{\mu=1}^{n} \hat{\ell}\left(y^\mu, \frac{\mathrm{diag}(\mathbf{W}^\intercal \mathbf{U}^\mu)}{\sqrt{p}} \right) + \lambda r(\mathbf{W}) \right] \tag{11.18}$$

where $r\colon \mathbb{R}^{p\times K} \to \mathbb{R}$ is a convex regularisation, $\mathbf{U}^\mu \equiv \mathbf{U}(\mathbf{x}^\mu)$ and $\hat{\mathbf{W}} \in \mathbb{R}^{p\times K}$ matrix of the concatenated columns $\{\hat{\mathbf{w}}_k\}$. Here, since the optimization problem defining the estimator may be non strictly convex, the solution may not be unique. We then denote with $\hat{\mathbf{W}}$ the unique least ℓ_2 norm solution of Eq.(11.18).

In the most general case, the statistical properties of $\hat{\mathbf{W}}$ are captured by a finite set of finite-dimensional order parameters, namely $\mathbf{V}, \hat{\mathbf{V}}, \mathbf{Q}, \hat{\mathbf{Q}} \in \mathbb{R}^{K\times K}$ and $\mathbf{m}, \hat{\mathbf{m}} \in \mathbb{R}^K$. These order parameters satisfy a set of fixed-point equations. To avoid a proliferation of indices in our formulas, let us introduce some notation. Let $\mathbf{A} = (A^{ij}_{kk'})^{i,j\in[p]}_{k,k'\in[K]} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}$ be a tensor, and $\mathbf{X} = (X^i_k)^{i\in[p]}_{k\in[K]}$, $\mathbf{Y} = (Y^i_k)^{i\in[p]}_{k\in[K]}$, $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p\times K}$ two matrices. We will denote

$$\langle\!\langle \mathbf{A} \rangle\!\rangle \coloneqq (\sum_i A^{ii}_{kk'})_{kk'} \in \mathbb{R}^{K\times K}, \tag{11.19a}$$

$$\langle\!\langle \mathbf{X}|\mathbf{A}|\mathbf{Y} \rangle\!\rangle \coloneqq (\sum_{ij} X^i_k A^{ij}_{kk'} Y^j_{k'})_{kk'} \in \mathbb{R}^{K\times K}, \tag{11.19b}$$

$$\langle\!\langle \mathbf{X}|\mathbf{Y} \rangle\!\rangle \coloneqq (\sum_{ij} X^i_k Y^i_k)_k \in \mathbb{R}^K, \tag{11.19c}$$

$$\langle \mathbf{X}|\mathbf{A}|\mathbf{Y} \rangle \coloneqq \sum_{ijk} X^i_k A^{ij}_{kk} Y^j_k \in \mathbb{R} \tag{11.19d}$$

$$\langle \mathbf{X}|\mathbf{Y} \rangle \coloneqq \sum_{ik} X^i_k Y^i_k \in \mathbb{R}. \tag{11.19e}$$

Given a second tensor $\mathbf{B} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}$, we write

$$\mathbf{AB} \coloneqq (\sum_{i'\kappa} A^{ii'}_{k\kappa} B^{i'j}_{\kappa k'})^{ij}_{kk'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}, \tag{11.19f}$$

$$\mathbf{A} \circ \mathbf{B} \coloneqq (\sum_{i'} A^{ii'}_{kk'} B^{i'j}_{k'k})^{ij}_{kk'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}, \tag{11.19g}$$

$$\mathbf{A} \odot \mathbf{B} \coloneqq (A^{ij}_{kk'} B^{ij}_{kk'})^{ij}_{kk'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}. \tag{11.19h}$$

We can now state our general result.

Theorem 21. *Let us consider the random quantities $\boldsymbol{\xi} \in \mathbb{R}^K$ and $\boldsymbol{\Xi} \in \mathbb{R}^{K\times K}$ with entries distributed as $\mathcal{N}(0,1)$. Assume that in the high-dimensional limit where $d, p, n \to +\infty$ with $\alpha \coloneqq {}^n\!/_p$ and $\gamma \coloneqq {}^d\!/_p$ kept $\Theta(1)$ constants. Then in this limit, for any pseudo-Lispchitz functions of order 2 $\varphi\colon \mathbb{R}\times\mathbb{R}^K \to \mathbb{R}$ and $\tilde{\varphi}\colon \mathbb{R}^{K\times p} \to \mathbb{R}$, the estimator $\hat{\mathbf{W}}$ verifies*

$$\begin{aligned}
&\mathbb{E}_{(y,\mathbf{x})}\left[\varphi\left(y, \frac{\langle\!\langle \hat{\mathbf{W}}|\mathbf{U} \rangle\!\rangle}{\sqrt{p}}\right)\right] \xrightarrow{\mathrm{P}} \int_{\mathcal{Y}} dy\, \mathbb{E}_{(\nu,\boldsymbol{\mu})}\left[P^0_y(y|\nu)\varphi\left(y, \boldsymbol{\mu}\right)\right], \\
&\frac{1}{n}\sum_{\mu=1}^n \varphi\left(y^\mu, \frac{\langle\!\langle \hat{\mathbf{W}}|\mathbf{U}^\mu \rangle\!\rangle}{\sqrt{p}}\right) \xrightarrow{\mathrm{P}} \int_{\mathcal{Y}} dy\, \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0\left(y, \omega_0, \sigma_0\right)\varphi(y, \mathbf{h})\right], \\
&\tilde{\varphi}(\hat{\mathbf{W}}) \xrightarrow{\mathrm{P}} \mathbb{E}_{\boldsymbol{\Xi}}\left[\tilde{\varphi}(\mathbf{G})\right],
\end{aligned} \tag{11.20}$$

where $\mathbf{U} \equiv \mathbf{U}(\mathbf{x})$, $(\nu, \boldsymbol{\mu}) \in \mathbb{R}^{1+K}$ are jointly Gaussian random variables with zero mean and covariance matrix

$$(\nu, \boldsymbol{\mu}) \sim \mathcal{N}\left(\mathbf{0}_{1+K}, \begin{pmatrix} \rho & \mathbf{M}^\intercal \\ \mathbf{M} & \mathbf{Q} \end{pmatrix}\right), \tag{11.21}$$

and we have introduced the proximals for the loss and the regularisation:

$$\begin{aligned}\mathbf{h} &\coloneqq \arg\min_{\mathbf{u}} \left[\frac{(\mathbf{u}-\boldsymbol{\omega})\mathbf{V}^{-1}(\mathbf{u}-\boldsymbol{\omega})}{2} + \hat{\ell}(y,\mathbf{u})\right], \\ \mathbf{G} &\coloneqq \arg\min_{\mathbf{U}} \left[\frac{\langle \mathbf{U}|(\mathbf{1}_{p,p}\otimes\hat{\mathbf{V}})\odot\boldsymbol{\Omega}|\mathbf{U}\rangle}{2} - \langle \mathbf{B}|\mathbf{U}\rangle + \lambda r(\mathbf{U})\right],\end{aligned} \tag{11.22}$$

with $\boldsymbol{\omega} \coloneqq \mathbf{Q}^{1/2}\boldsymbol{\xi}$ *and* $\mathbf{B} \coloneqq (\mathbf{1}_p \otimes \hat{\mathbf{m}}^\intercal)\odot\hat{\boldsymbol{\Phi}} + ((\mathbf{1}_{p,p}\otimes\hat{\mathbf{Q}})\odot\boldsymbol{\Omega})^{\frac{1}{2}}\boldsymbol{\Xi}$. *We have also introduced the auxiliary function*

$$\mathcal{Z}^0(y,\mu,\sigma) \coloneqq \int \frac{P_y^0(y|x)\,dx}{\sqrt{2\pi\sigma}}\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma}}. \tag{11.23}$$

and the scalar quantities $\omega_0 \coloneqq \mathbf{M}^\intercal\mathbf{Q}^{-1/2}\boldsymbol{\xi}$ *and* $\sigma_0 \coloneqq \rho - \mathbf{M}^\intercal\mathbf{Q}^{-1}\mathbf{M}$. *The order parameters satisfy the saddle-point equations*

$$\begin{aligned}\hat{\mathbf{V}} &= -\alpha\int_{\mathcal{Y}} dy\, \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\,\partial_{\boldsymbol{\omega}}\mathbf{f}\right], \\ \hat{\mathbf{Q}} &= \alpha\int_{\mathcal{Y}} dy\, \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\,\mathbf{f}\mathbf{f}^\intercal\right], \\ \hat{\mathbf{m}} &= \frac{\alpha}{\sqrt{\gamma}}\int_{\mathcal{Y}} dy\, \mathbb{E}_{\boldsymbol{\xi}}\left[\partial_\mu\mathcal{Z}^0(y,\omega_0,\sigma_0)\mathbf{f}\right],\end{aligned} \tag{11.24}$$

and

$$\begin{aligned}\mathbf{V} &= \frac{2}{p}\mathbb{E}_{\boldsymbol{\Xi}}\Big\langle \mathbf{G}\Big|\frac{D\left((\mathbf{1}_{p,p}\otimes\hat{\mathbf{Q}})\odot\boldsymbol{\Omega}\right)^{1/2}}{D\hat{\mathbf{Q}}}\Big|\boldsymbol{\Xi}\Big\rangle \\ \mathbf{Q} &= \frac{1}{p}\mathbb{E}_{\boldsymbol{\Xi}}\langle\!\langle \mathbf{G}|\boldsymbol{\Omega}|\mathbf{G}\rangle\!\rangle, \\ \mathbf{M} &= \frac{1}{\sqrt{\gamma}p}\mathbb{E}_{\boldsymbol{\Xi}}\langle\!\langle \hat{\boldsymbol{\Phi}}|\mathbf{G}\rangle\!\rangle.\end{aligned} \tag{11.25}$$

In the equation above we have introduced the short-hand notation $\mathbf{f} \coloneqq \mathbf{V}^{-1}(\mathbf{h}-\boldsymbol{\omega})$.

In the theorem above, for a tensor $\hat{\mathbf{A}} \in \mathbb{R}^{p\times p}\otimes\mathbb{R}^{K\times K}$, then $[\frac{\mathrm{D}\hat{\mathbf{A}}}{\mathrm{D}\hat{\mathbf{Q}}}]_{ij}^{kk',\kappa\kappa'} \equiv \frac{\partial\hat{A}_{ij}^{kk'}}{\partial\hat{Q}_{\kappa\kappa'}}$: in the formula, the contractions involve latin indices only. Eqs. (11.24) are typically called *channel equations*, because depend on the form of the loss $\hat{\ell}$. Eqs. (11.25), instead, are usually called *prior equations*, because of their dependence on the prior, i.e., r. In the following Corollary, we specify their expression for a ridge regularisation, $r(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_{\mathrm{F}}^2$.

Corollary 4 (Ridge regularisation)**.** *In the hypotheses of Theorem 21, if* $r(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_{\mathrm{F}}^2$, *then the prior equations are*

$$\begin{aligned}\mathbf{V} &= \frac{1}{p}\langle\!\langle \boldsymbol{\Omega}\circ\mathbf{A}\rangle\!\rangle, \\ \mathbf{Q} &= \frac{1}{p}\langle\!\langle \boldsymbol{\Omega}\circ\Big(\mathbf{A}\left((\mathbf{1}_{p,p}\otimes\hat{\mathbf{m}}\otimes\hat{\mathbf{m}}^\intercal)\odot\boldsymbol{\Theta} + (\mathbf{1}_{p,p}\otimes\hat{\mathbf{Q}})\odot\boldsymbol{\Omega}\right)\mathbf{A}\Big)\rangle\!\rangle, \\ \mathbf{M} &= \frac{1}{\sqrt{\gamma}p}\langle\!\langle \mathbf{A}\left((\mathbf{1}_{p,p}\otimes\hat{\mathbf{m}}\otimes\mathbf{1}_K^\intercal)\odot\boldsymbol{\Theta}\right)\rangle\!\rangle.\end{aligned} \tag{11.26}$$

In the equation above, we have used the auxiliary tensor $\mathbf{A} \equiv \mathbf{A}(\hat{\mathbf{V}};\lambda,\boldsymbol{\Omega}) \coloneqq (\lambda\mathbf{I}_p\otimes\mathbf{I}_K + (\mathbf{1}_{p,p}\otimes\hat{\mathbf{V}})\odot\boldsymbol{\Omega})^{-1} \in \mathbb{R}^{p\times p}\otimes\mathbb{R}^{K\times K}$.

11.4.1 The random feature case

Theorem 21 is given in a very general setting, and, in particular, no assumptions are made on the features $\mathbf{u}_k$. We have anticipated in Sec. 11.2 that, in the case of random features, the structure of the order parameters highly simplifies and the covariance matrix $\mathbf{\Sigma}$ is fully specified by only three scalar order parameters for any $K > 1$. Here will adapt therefore Theorem 21 to the random feature setting in Sec. 11.2, using the notation therein. The motivation of this section is to explicitly present the self-consistent equations that are required to produce the results given in the paper.

Corollary 5. *Assume that in the high-dimensional limit where $d, p, n \to +\infty$ with $\alpha \coloneqq {}^{n}/_{p}$ and $\gamma \coloneqq {}^{d}/_{p}$ kept $\Theta(1)$ constants, the Wishart matrix $\mathbf{F}\mathbf{F}^\intercal$ has a well-defined asymptotic spectral distribution. Then in this limit, for any pseudo-Lispchitz function of finite order $\varphi\colon \mathbb{R} \times \mathbb{R}^K \to \mathbb{R}$, the estimator $\hat{\mathbf{W}}$ verifies*

$$\mathbb{E}_{(\mathbf{x},y)} \left[\varphi \left(y, \frac{\langle\!\langle \hat{\mathbf{W}} | \mathbf{U} \rangle\!\rangle}{\sqrt{p}} \right) \right] \xrightarrow{\text{P}} \mathbb{E}_{(\nu,\boldsymbol{\mu})} \left[\varphi \left(f_0(\nu), \boldsymbol{\mu} \right) \right], \tag{11.27}$$

where $(\nu, \boldsymbol{\mu}) \in \mathbb{R}^{K+1}$ is a jointly Gaussian vector with covariance

$$(\nu, \boldsymbol{\mu}) \sim \mathcal{N} \left(\mathbf{0}_{K+1}, \begin{pmatrix} \rho & m\mathbf{1}_K^\intercal \\ m\mathbf{1}_K & \mathbf{Q} \end{pmatrix} \right), \tag{11.28}$$

and $\mathbf{Q} \coloneqq (q_0 - q_1)\mathbf{I}_K + q_1 \mathbf{1}_{K,K}$. The collection of parameters (q_0, q_1, m) is obtained solving a set of fixed point equations involving the auxiliary variables $(\hat{q}_0, \hat{q}_1, \hat{m}, v, \hat{v})$, namely:

$$\hat{v} = -\alpha \int_{\mathcal{Y}} dy\, \mathbb{E}_\omega \left[\mathcal{Z}^0 \left(y, \frac{m\omega}{q_0}, \rho - \frac{m^2}{q_0} \right) \partial_\omega f \right], \tag{11.29a}$$

$$\hat{m} = \frac{\alpha}{\sqrt{\gamma}} \int_{\mathcal{Y}} dy\, \mathbb{E}_\omega \left[\partial_\mu \mathcal{Z}^0 \left(y, \frac{m\omega}{q_0}, \rho - \frac{m^2}{q_0} \right) f \right], \tag{11.29b}$$

$$\hat{q}_0 = \alpha \int_{\mathcal{Y}} dy\, \mathbb{E}_\omega \left[\mathcal{Z}^0 \left(y, \frac{m\omega}{q_0}, \rho - \frac{m^2}{q_0} \right) f^2 \right], \tag{11.29c}$$

$$\hat{q}_1 = \alpha \int_{\mathcal{Y}} dy \mathbb{E}_{\omega,\omega'} \left[\mathcal{Z}^0 \left(y, m\frac{\omega + \omega'}{q_0 + q_1}, \rho - \frac{2m^2}{q_0 + q_1} \right) f f' \right], \tag{11.29d}$$

$$v = \int \frac{s\varrho(s)\,ds}{\lambda + s\hat{v}}, \tag{11.29e}$$

$$m = \frac{\hat{m}}{\sqrt{\gamma}} \int \frac{s - \kappa_*^2}{\lambda + \hat{v}s} \varrho(s)\,ds, \tag{11.29f}$$

$$q_0 = \int \frac{(\hat{q}_0 + \hat{m}^2)s^2 - \hat{m}^2\kappa_*^2 s}{(\lambda + \hat{v}s)^2} \varrho(s)\,ds, \tag{11.29g}$$

$$q_1 = \left(1 + \frac{\hat{q}_1}{\hat{m}^2} \right) m^2. \tag{11.29h}$$

where ω and ω' are two correlated Gaussian random variables of zero mean and $\mathbb{E}[\omega^2] = \mathbb{E}[\omega'^2] = q_0$, $\mathbb{E}[\omega\omega'] = q_1$. Moreover, we have introduced the proximals

$$f = \frac{\text{Prox}_{v\ell(y,\bullet)}(\omega) - \omega}{v}, \quad f' = \frac{\text{Prox}_{v\ell(y,\bullet)}(\omega') - \omega'}{v}, \tag{11.30}$$

with

$$\text{Prox}_{v\ell(y,\bullet)}(\omega) \coloneqq \arg\min_x \left[\frac{(x - \omega)^2}{2v} + \ell(y, x) \right]. \tag{11.31}$$

Finally, $\varrho(s)$ is the asymptotic spectral density of the features covariance matrix $\boldsymbol{\Omega} \equiv \mathrm{Var}(\mathbf{u}) = \kappa_0^2 \mathbf{1}_{p,p} + \frac{\kappa_1^2}{d}\mathbf{F}\mathbf{F}^\intercal + \kappa_*^2 \mathbf{I}_p$ and the coefficients are given by $\kappa_0 \coloneqq \mathbb{E}_\zeta[\phi(\zeta)]$, $\kappa_1 \coloneqq \mathbb{E}_\zeta[\zeta\phi(\zeta)]$, $\kappa_* \coloneqq \mathbb{E}_\zeta[\phi^2(\zeta)] - \kappa_0^2 - \kappa_1^2$ with $\zeta \sim \mathcal{N}(0,1)$.

The previous corollary recovers the results of [107], [76], and [178] when restricted to the $K = 1$ case by marginalisation.

Chapter 12

Proofs for the ensembling

12.1 Proof of the main theorem

In this section we prove Theorem 21, from which all other analytical results in the paper can be deduced. We start by reminding the learning problem defining the ensemble of estimators with a few auxiliary notations, so that this part is self contained. The sketch of proof is one pioneered in [29, 82] and is the following: the estimator $\mathbf{W}^*$ is expressed as the limit of a carefully chosen sequence, an *approximate message-passing iteration* [28, 300], whose iterates can be asymptotically exactly characterized using an auxiliary, closed form iteration, the *state evolution equations*. We then show that converging trajectories of such an AMP iteration can be systematically found.

12.1.1 The learning problem

We start by reminding the definition of the problem. Consider the following generative model

$$\mathbf{y} = f_0(\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0, \boldsymbol{\epsilon}_0) \tag{12.1}$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X}_0 \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{00}) \in \mathbb{R}^{n\times d}, \mathbf{w}_0 \in \mathbb{R}^d$, $\boldsymbol{\epsilon}_0 \in \mathbb{R}^d$ is a noise vector and $\boldsymbol{\Sigma}_{\mathbf{00}} \in \mathbb{R}^{d\times d}$ is a positive definite matrix. The goal is to learn this generative model using an ensemble of predictors $\mathbf{W} = \left[\mathbf{w}_1|\mathbf{w}_2|...|\mathbf{w}_K\right] \in \mathbb{R}^{p\times K}$ where each predictor $\mathbf{w}_k \in \mathbb{R}^p, k \in [1, K]$ is learned using a sample dataset $\mathbf{X}_k \in \mathbb{R}^{n\times p}$, where, for any $i \in [1, n]$ and $k \in [0, K]$, we have:

$$\mathbb{E}\left[\mathbf{x}_i^k(\mathbf{x}_i^{k'})^\top\right] = \boldsymbol{\Sigma}_{kk'} \tag{12.2}$$

where each sample is Gaussian and we denote :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01} & ... & \boldsymbol{\Sigma}_{0K} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} & ... & \boldsymbol{\Sigma}_{1K} \\ ... & & & \\ \boldsymbol{\Sigma}_{K0} & \boldsymbol{\Sigma}_{K1} & ... & \boldsymbol{\Sigma}_{KK} \end{bmatrix} \in \mathbb{R}^{(Kp+d)\times(Kp+d)}. \tag{12.3}$$

The predictors interact with each sample dataset in a linear way, i.e. we will consider a generalized linear model acting on the ensemble of products $\{\mathbf{X}_k\mathbf{w}_k\}_{k=1}^K$:

$$\mathbf{W}^* \in \underset{\mathbf{W}\in\mathbb{R}^{p\times K}}{\arg\min}\, \mathcal{L}\left(\mathbf{y}, \left\{\frac{1}{\sqrt{p}}\mathbf{X}_k\mathbf{w}_k\right\}_{k=1}^K\right) + r_0(\mathbf{W}) \tag{12.4}$$

where $\mathcal{L}, r_0$ are convex functions. We wish to determine the asymptotic properties of the estimator $\mathbf{W}^*$ in the limit where $n, p, d \to \infty$ with fixed ratios $\alpha = n/p, \gamma = d/p$. We now list the necessary assumptions for our main theorem to hold.

Assumptions –

- the functions $\mathcal{L}, r_0$ are proper, closed, lower-semicontinuous, convex functions. The loss function $\mathcal{L}$ is pseudo-lipschitz of order 2 in both its arguments and the regularisation r_0 is pseudo-Lipschitz of order 2. The cost function $\mathcal{L}(\mathbf{X}.) + r(.)$ is coercive.
- for any $1 \leqslant k \leqslant K$, the matrix $\mathbf{\Sigma}_k \in \mathbb{R}^{p\times p}$ is symmetric and there exist strictly positive constants κ_0, κ_1 such that $\kappa_0 \leqslant \lambda_{min}(\mathbf{\Sigma}_k) \leqslant \lambda_{max}(\mathbf{\Sigma}_k) \leqslant \kappa_1$. We also assume that the matrix $\mathbf{\Sigma}$ is positive definite.
- their exists a positive constant C_{f_0} such that $\left\| f_0(\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0, \boldsymbol{\epsilon}_0)\right\|_2 \leqslant C_{f_0}\left(\left\|\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0\right\|_2 + \|\boldsymbol{\epsilon}_0\|_2\right)$
- the dimensions n, p, d grow linearly with finite ratios $\alpha = n/p$ and $\gamma = d/p$.
- the ground truth vector $\mathbf{w}_0 \in \mathbb{R}^d$ and noise vector $\boldsymbol{\epsilon}_0 \in \mathbb{R}^n$ are sampled from subgaussian probability distributions independent from each other and from all other random quantities of the learning problem.

The proof method we will employ involves expressing the estimator $\mathbf{W}^*$ as the limit of a carefully chosen sequence. In the case of non-strictly convex problems, the estimator may not be unique, making it unclear what estimator is reached by the sequence (at best we know it belongs to the set of zeroes of the subgradient of the cost function). We thus start with the following problem

$$\mathbf{W}^* \in \underset{\mathbf{W}\in\mathbb{R}^{p\times K}}{\arg\min}\, \mathcal{L}(\mathbf{y}, \{\mathbf{X}_k\mathbf{w}_k\}_{k=1}^K) + r_{\lambda_2}(\mathbf{W}) \tag{12.5}$$

$$\text{where, for any } \mathbf{W} \in \mathbb{R}^{p\times K},\ r_{\lambda_2}(\mathbf{W}) = r_0(\mathbf{W}) + \frac{\lambda_2}{2}\|\mathbf{W}\|_F^2 \tag{12.6}$$

i.e. we add a ridge regularisation to the initial problem to make it strongly convex. We will relax this additional strong convexity constraint later on.

12.1.2 Asymptotics for the strongly convex problem

We now reformulate the minimization problem Eq.(12.5) to make it amenable to an approximate message-passing iteration (AMP). The key feature of this ensembling problem, outside of the convexity which will be crucial to control the trajectories of the AMP iteration, is the fact that each predictor only interacts linearly with each design sample, along with the correlation structure of the overall dataset. We are effectively sampling n vectors of size $(Kp + d)$ from the Gaussian distribution with covariance $\mathbf{\Sigma}$, i.e. $\left[\mathbf{x}_0|\mathbf{x}_1|...|\mathbf{x}_K\right] \sim \mathcal{N}(0, \mathbf{\Sigma})$. We then write $\{\mathbf{X}_k\mathbf{w}_k\}_{k=0}^K = \left[\mathbf{X}_0\mathbf{w}_0|...|\mathbf{X}_K\mathbf{w}_K\right] \in \mathbb{R}^{n\times(K+1)}$, such that

$$\left[\mathbf{X}_0\mathbf{w}_0|...|\mathbf{X}_K\mathbf{w}_K\right] = \left[\mathbf{X}_0|...|\mathbf{X}_K\right]\mathbf{W} = \mathbf{Z}\mathbf{\Sigma}^{1/2}\begin{bmatrix}\mathbf{w}_0 & 0\\ 0 & \tilde{\mathbf{W}}\end{bmatrix} \tag{12.7}$$

$$\text{where}\quad \tilde{\mathbf{W}} = \begin{bmatrix}\mathbf{w}_1 & 0 & ... & 0\\ 0 & \mathbf{w}_2 & ... & 0\\ & & ... & \\ 0 & 0 & ... & \mathbf{w}_K\end{bmatrix} \in \mathbb{R}^{Kp\times K} \tag{12.8}$$

and $\mathbf{Z} \in \mathbb{R}^{n\times(Kp+d)}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ elements. Then, any sample $\left[\mathbf{x}_0|\mathbf{x}_1|...|\mathbf{x}_K\right]$ may be rewritten as

$$\mathbf{x}_0 = \Psi^{1/2}\mathbf{a} \quad \text{and} \quad \left[\mathbf{x}_1|...|\mathbf{x}_K\right] = \Phi^\top\Psi^{-1/2}\mathbf{a} + \left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\mathbf{b} \tag{12.9}$$

$$\mathbf{X}_0 = \mathbf{A}\Psi^{1/2} \quad \text{and} \quad \left[\mathbf{X}_1|...|\mathbf{X}_K\right] = \mathbf{A}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2} \tag{12.10}$$

where $\mathbf{a} \in \mathbb{R}^d, \mathbf{b} \in \mathbb{R}^{Kp}$ are vectors with i.i.d. standard normal components, $\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{B} \in \mathbb{R}^{n\times Kp}$ are the corresponding design matrices, and the covariance matrices are given by $\mathbf{\Psi} = \mathbf{\Sigma}_{00} \in \mathbb{R}^{d\times d}, \mathbf{\Phi} = \left[\mathbf{\Sigma}_{11}|\mathbf{\Sigma}_{12}|\mathbf{\Sigma}_{13}...|\mathbf{\Sigma}_{1K}\right] \in \mathbb{R}^{d\times Kp}$ and

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & ... & \mathbf{\Sigma}_{1K} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} & ... & \mathbf{\Sigma}_{2K} \\ ... & & & \\ \mathbf{\Sigma}_{K1} & \mathbf{\Sigma}_{K2} & ... & \mathbf{\Sigma}_{KK} \end{bmatrix} \in \mathbb{R}^{Kp\times Kp} \tag{12.11}$$

The optimization problem may then be written, introducing the appropriate scalings

$$\tilde{\mathbf{W}}^* \in \underset{\tilde{\mathbf{W}}\in\mathbb{R}^{Kp\times K}}{\arg\min}\, \mathcal{L}\left(f_0(\frac{1}{\sqrt{d}}\mathbf{A}\tilde{\mathbf{w}}_0), \frac{1}{\sqrt{p}}\left(\mathbf{A}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) \tag{12.12}$$

where we let $\tilde{\mathbf{w}}_0 = \Psi^{1/2}\mathbf{w}_0$, its scaled norm $\rho_{\tilde{\mathbf{w}}_0} = \frac{1}{d}\|\tilde{\mathbf{w}}_0\|_2^2$ and we introduced the function

$$r : \mathbb{R}^{Kp\times K} \to \mathbb{R} \tag{12.13}$$

$$\tilde{\mathbf{W}} \to r_{\lambda_2}(\mathbf{W}) \tag{12.14}$$

. In order to isolate the contribution correlated with the teacher, we condition the design matrix $\mathbf{A}$ on the teacher distribution $\mathbf{y}$, we can write

$$\mathbf{A} = \mathbb{E}\left[\mathbf{A}|\mathbf{y}\right] + \mathbf{A} - \mathbb{E}\left[\mathbf{A}|\mathbf{y}\right] \tag{12.15}$$

$$= \mathbb{E}\left[\mathbf{A}|\mathbf{A}\tilde{\mathbf{w}}_0\right] + \mathbf{A} - \mathbb{E}\left[\mathbf{A}|\mathbf{A}\tilde{\mathbf{w}}_0\right] \tag{12.16}$$

$$= \mathbf{A}\mathbf{P}_{\tilde{\mathbf{w}}_0} + \tilde{\mathbf{A}}\mathbf{P}^\perp_{\tilde{\mathbf{w}}_0} \tag{12.17}$$

where $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$, see [28] Lemma 11. The cost function then becomes

$$\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \frac{1}{\sqrt{p}}\left(\mathbf{s}\frac{(\Phi^\top\mathbf{w}_0)^\top}{\sqrt{d\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{A}}\mathbf{P}^\perp_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) \tag{12.18}$$

where $\mathbf{s} = \mathbf{A}\frac{\tilde{\mathbf{w}}_0}{\|\tilde{\mathbf{w}}_0\|_2} \in \mathbb{R}^n$ is an i.i.d. standard normal vector.

The term $\tilde{\mathbf{A}}\mathbf{P}^\perp_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}$ can then be represented as a $\mathbb{R}^{n\times Kp}$ Gaussian matrix with covariance

$$\Phi^\top\Psi^{-1/2}\mathbf{P}^\perp_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi + \Omega - \Phi^\top\Psi^{-1}\Phi = \Omega - \Phi^\top\Psi^{-1/2}\mathbf{P}_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi \tag{12.19}$$

$$= \Omega - \Phi^\top\Psi^{-1/2}\frac{\tilde{\mathbf{w}}_0\tilde{\mathbf{w}}_0^\top}{\|\tilde{\mathbf{w}}_0\|_2^2}\Psi^{-1/2}\Phi = \Omega - \frac{\mathbf{c}\mathbf{c}^\top}{\|\tilde{\mathbf{w}}_0\|_2^2} \tag{12.20}$$

where we introduced $\mathbf{c} = \Phi^\top \mathbf{w}_0 \in \mathbb{R}^{Kp}$ and $\rho_{\mathbf{c}} = \frac{1}{p}\|\mathbf{c}\|_2^2$, reaching the cost function

$$\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \frac{1}{\sqrt{p}}\left(\mathbf{s}\frac{\mathbf{c}^\top}{\sqrt{d\rho_{\tilde{\mathbf{w}}_0}}} + \mathbf{Z}\left(\Omega - \frac{\mathbf{c}\mathbf{c}^\top}{\|\tilde{\mathbf{w}}_0\|_2^2}\right)^{1/2}\right)\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) \tag{12.21}$$

Introducing $\mathbf{m} = \frac{1}{\sqrt{dp}}\tilde{\mathbf{W}}^\top \mathbf{c} \in \mathbb{R}^K, \mathbf{C} = \Omega - \frac{\mathbf{c}\mathbf{c}^\top}{\|\tilde{\mathbf{w}}_0\|_2^2} \in \mathbb{R}^{Kp\times Kp}$, and the Lagrange multiplier $\boldsymbol{\nu}$ associated to $\mathbf{m}$, the optimization problem can equivalently be written

$$\inf_{\mathbf{m}\in\mathbb{R}^K, \tilde{\mathbf{W}}\in\mathbb{R}^{Kp\times K}} \sup_{\boldsymbol{\nu}\in\mathbb{R}^K} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{1}{\sqrt{p}}\mathbf{Z}\mathbf{C}^{1/2}\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) - \boldsymbol{\nu}^\top\left(\tilde{\mathbf{W}}^\top\mathbf{c} - \sqrt{dp}\mathbf{m}\right) \tag{12.22}$$

We now look for an explicit expression of the matrix square root $\mathbf{C}^{1/2}$

$$\mathbf{C} = \Omega^{1/2}\left(Id - \frac{\Omega^{-1/2}\mathbf{c}(\Omega^{-1/2}\mathbf{c})^\top}{\|\tilde{\mathbf{w}}_0\|_2^2}\right)\Omega^{1/2} \quad \text{let} \quad \tilde{\mathbf{c}} = \Omega^{-1/2}\mathbf{c} \tag{12.23}$$

$$= \Omega^{1/2}\left(\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp} + \kappa\mathbf{P}_{\tilde{\mathbf{c}}}\right)\Omega^{1/2} \quad \text{where} \quad \kappa = 1 - \frac{\|\tilde{\mathbf{c}}\|_2^2}{\|\tilde{\mathbf{w}}_0\|_2^2} \tag{12.24}$$

$$= \Omega^{1/2}\left(\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp} + \sqrt{\kappa}\mathbf{P}_{\tilde{\mathbf{c}}}\right)\left(\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp} + \sqrt{\kappa}\mathbf{P}_{\tilde{\mathbf{c}}}\right)\Omega^{1/2} \tag{12.25}$$

where the positivity of κ is ensured by the positive-definiteness of $\boldsymbol{\Sigma}$. The problem then becomes

$$\inf_{\mathbf{m},\tilde{\mathbf{W}}} \sup_{\boldsymbol{\nu}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\sqrt{\kappa}}{\sqrt{p}}\mathbf{Z}\mathbf{P}_{\tilde{\mathbf{c}}}\Omega^{1/2}\tilde{\mathbf{W}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}\Omega^{1/2}\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) - \boldsymbol{\nu}^\top\left(\tilde{\mathbf{W}}^\top\mathbf{c} - \sqrt{dp}\mathbf{m}\right) \tag{12.26}$$

where $\tilde{\mathbf{Z}}$ is an independent copy of $\mathbf{Z}$, see [28] Lemma 11. Then

$$\frac{\sqrt{\kappa}}{\sqrt{p}}\mathbf{Z}\mathbf{P}_{\tilde{\mathbf{c}}}\Omega^{1/2}\tilde{\mathbf{W}} = \frac{\sqrt{\kappa}}{\sqrt{p}}\tilde{\mathbf{s}}\frac{\mathbf{c}^\top\tilde{\mathbf{W}}}{\|\tilde{\mathbf{c}}\|_2} \tag{12.27}$$

$$= \sqrt{\kappa}\tilde{\mathbf{s}}\frac{\mathbf{c}^\top\tilde{\mathbf{W}}}{p\sqrt{\rho_{\tilde{\mathbf{c}}}}} \tag{12.28}$$

$$= \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} \tag{12.29}$$

where $\tilde{\mathbf{s}} = \mathbf{Z}\frac{\tilde{\mathbf{c}}}{\|\tilde{\mathbf{c}}\|_2}$ is an i.i.d. standard normal vector and $\rho_{\tilde{\mathbf{c}}} = \frac{1}{p}\|\tilde{\mathbf{c}}\|_2^2$ such that the optimization problem becomes

$$\inf_{\mathbf{m},\tilde{\mathbf{W}}} \sup_{\boldsymbol{\nu}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}\Omega^{1/2}\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) - \boldsymbol{\nu}^\top\left(\tilde{\mathbf{W}}^\top\mathbf{c} - \sqrt{dp}\mathbf{m}\right) \tag{12.30}$$

Now let $\mathbf{U} = \mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}\Omega^{1/2}\tilde{\mathbf{W}}$, such that $\tilde{\mathbf{W}} = \Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)$. The equivalent problem in $\mathbf{U}$ reads

$$\inf_{\mathbf{m},\mathbf{U}} \sup_{\boldsymbol{\nu}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right) + r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) - \boldsymbol{\nu}^\top\mathbf{U}^\top\tilde{\mathbf{c}} \tag{12.31}$$

Note that the constraint defining $\mathbf{m}$ automatically enforces the orthogonality constraint on $\mathbf{U}$ w.r.t. $\tilde{\mathbf{c}}$. The following lemma characterizes properties of the feasibility sets of $\mathbf{U}, \mathbf{m}, \boldsymbol{\nu}$.

Lemma 47. *Consider the optimization problem Eq.(12.31). Then there exist constants $C_{\mathbf{U}}, C_{\mathbf{m}}, C_{\boldsymbol{\nu}}$ such that*

$$\frac{1}{\sqrt{p}}\|\mathbf{U}\|_F \leqslant C_{\mathbf{U}}, \quad \|\mathbf{m}\|_2 \leqslant C_{\mathbf{m}}, \quad \|\boldsymbol{\nu}\|_2 \leqslant C_{\boldsymbol{\nu}} \tag{12.32}$$

with high probability as $n, p, d \to \infty$.

Proof. Consider the optimization problem defining $\tilde{\mathbf{W}}^*$

$$\tilde{\mathbf{W}}^* \in \underset{\tilde{\mathbf{W}}\in\mathbb{R}^{Kp\times K}}{\arg\min}\ \mathcal{L}(\mathbf{y}, \mathbf{X}\tilde{\mathbf{W}}) + \tilde{r}_0(\mathbf{W}) + \frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}\right\|_F^2 \tag{12.33}$$

which, owing to the convexity of the cost function, verifies

$$\frac{1}{p}\left(\mathcal{L}(\mathbf{y}, \mathbf{X}\tilde{\mathbf{W}}^*) + \tilde{r}_0(\tilde{\mathbf{W}}^*) + \frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2\right) \leqslant \frac{1}{p}\left(\mathcal{L}(\mathbf{y}, 0) + \tilde{r}_0(0)\right) \tag{12.34}$$

The functions $\mathcal{L}$ and $\tilde{r}_0$ are assumed to be proper, thus their sum is bounded below for any value of their arguments and we may write

$$\frac{1}{p}\frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2 \leqslant \frac{1}{p}\left(\mathcal{L}(\mathbf{y}, 0) + \tilde{r}_0(0)\right) \tag{12.35}$$

The pseudo-Lipschitz assumption on $\mathcal{L}$ and $\tilde{r}_0$ then implies that there exist positive constants $C_{\mathcal{L}}$ and $C_{\tilde{r}_0}$ such that

$$\frac{1}{p}\frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2 \leqslant \frac{1}{p}\left(C_{\mathcal{L}}\left(1 + \|\mathbf{y}\|_2^2\right)\right) + C_{\tilde{r}_0} \tag{12.36}$$

$$\leqslant \frac{1}{p}\left(C_{\mathcal{L}}\left(1 + C_{f_0}\left\|\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0\right\|_2^2 + C_{f_0}\left\|\boldsymbol{\epsilon}_0^2\right\|\right)\right) + C_{\tilde{r}_0} \tag{12.37}$$

where the second line follows from the scaling assumption on the teacher function f_0. Hence

$$\frac{1}{p}\frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2 \leqslant C_{\mathcal{L}}\left(1 + C_{f_0}\left\|\frac{1}{\sqrt{d}}\mathbf{A}\right\|_{op}^2\left\|\Psi^{1/2}\right\|_{op}^2\frac{\gamma}{d}\|\mathbf{w}\|_0^2 + C_{f_0}\frac{\alpha}{n}\|\boldsymbol{\epsilon}_0\|_2^2\right) + C_{\tilde{r}_0} \tag{12.38}$$

where $\|\bullet\|_{op}$ denotes the operator norm of a given matrix, and we remind that $\mathbf{A}$ has i.i.d. $\mathcal{N}(0,1)$ elements. By assumption the maximum singular value of $\Psi^{1/2}$ is bounded. The maximum singular value of a random matrix with i.i.d. $\mathcal{N}(0, \frac{1}{d})$ elements is bounded with high probability as $n, p, d \to \infty$, see e.g., [287]. Finally, $\mathbf{w}_0$ and $\boldsymbol{\epsilon}_0$ are sampled from subgaussian probability distributions, thus their scaled norms are bounded with high probability as $n, p, d \to \infty$ according to Bernstein's inequality, see e.g., [288]. An application of the union bound then leads to the following statement:

there exists a constant $C_{\tilde{\mathbf{W}}}$ such that $\frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_2^2 \leqslant C_{\tilde{\mathbf{W}}}$, with high probability as $n, p, d \to \infty$. Now using the definition of $\mathbf{U}$

$$\frac{1}{p}\|\mathbf{U}\|_F^2 = \frac{1}{p}\left\|\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}\Omega^{1/2}\tilde{\mathbf{W}}\right\|_F^2 \tag{12.39}$$

$$\leqslant \left\|\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}\right\|_{op}^2 \left\|\Omega^{1/2}\right\|_{op}^2 \frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_F \tag{12.40}$$

where the singular values of $\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}$ and $\Omega^{1/2}$ are bounded with probability one. Therefore there exists a constant $C_{\mathbf{U}}$ such that $\frac{1}{\sqrt{p}}\|\mathbf{U}\| \leqslant C_{\mathbf{U}}$ with high probability as $n, p, d \to \infty$. Then, by definition of $\mathbf{m}$ and the Cauchy-Schwarz inequality

$$\|\mathbf{m}\|_2^2 \leqslant \frac{1}{d}\|\mathbf{c}\|_2^2 \frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_F^2 \tag{12.41}$$

$$\leqslant \|\Phi\|_{op}^2 \frac{1}{d}\|\mathbf{w}_0\|_2^2 \frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_F^2 \tag{12.42}$$

combining the results previously established on $\tilde{\mathbf{W}}$ and $\mathbf{w}_0$ by the fact that the maximum singular value of Φ is bounded, there exists a positive constant $C_{\mathbf{m}}$ such that $\|\mathbf{m}\|_2 \leqslant C_{\mathbf{m}}$ with high probability as $n, p, d \to \infty$. We finally turn to $\boldsymbol{\nu}$. The optimality condition for $\mathbf{m}$ in problem Eq.(12.22) gives

$$\boldsymbol{\nu} = -\frac{1}{\sqrt{dp}}\frac{\mathbf{s}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}\partial\mathcal{L}\left(\mathbf{y}, \frac{\mathbf{s}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{1}{\sqrt{p}}\mathbf{Z}\mathbf{C}^{1/2}\tilde{\mathbf{W}}^*\right) \tag{12.43}$$

The pseudo-Lipschtiz assumption on $\mathcal{L}$ implies that we can find a constant $\mathcal{C}_{\partial\mathcal{L}}$ such that

$$\|\boldsymbol{\nu}\|_2^2 = \frac{1}{dp}\frac{\|\mathbf{s}\|_2^2}{\rho_{\tilde{\mathbf{w}}_0}} C_{\mathcal{L}}\left(1 + \|\mathbf{y}\|_2^2 + \left\|\frac{\mathbf{s}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{1}{\sqrt{p}}\mathbf{Z}\mathbf{C}^{1/2}\tilde{\mathbf{W}}^*\right\|_2^2\right) \tag{12.44}$$

the last bound then follows from similar arguments as those employed above. □

The optimization problem Eq.(12.31) is convex and feasible. Furthermore, we may reduce the feasibility sets of $\mathbf{m}, \boldsymbol{\nu}$ to compact spaces, and the function of $\mathbf{U}$ is coercive and thus has bounded lower level sets. Strong duality then implies we can invert the order of minimization to obtain the equivalent problem

$$\inf_{\mathbf{m}}\sup_{\boldsymbol{\nu}}\inf_{\mathbf{U}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right) + r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) - \boldsymbol{\nu}^\top\mathbf{U}^\top\tilde{\mathbf{c}} \tag{12.45}$$

and study the optimization problem in $\mathbf{U}$ at fixed $\mathbf{m}, \boldsymbol{\nu}$:

$$\inf_{\mathbf{U}\in\mathbb{R}^{Kp\times K}} \tilde{\mathcal{L}}(\frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}) + \tilde{r}(\mathbf{U}) \tag{12.46}$$

where we defined the functions

$$\tilde{\mathcal{L}} : \mathbb{R}^{n\times K} \to \mathbb{R} \tag{12.47}$$

$$\frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U} \to \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right) \tag{12.48}$$

$$\tilde{r} : \mathbb{R}^{Kp\times K} \to \mathbb{R} \tag{12.49}$$

$$\mathbf{U} \to r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) - \boldsymbol{\nu}^\top\mathbf{U}^\top\tilde{\mathbf{c}} \tag{12.50}$$

and the random matrix $\tilde{\mathbf{Z}}$ with i.i.d. $\mathcal{N}(0,1)$ elements is independent from all other random quantities in the problem. The asymptotic properties of the unique solution to this optimization problem can now be studied with a non-separable, matrix-valued approximate message passing iteration. The AMP iteration solving problem Eq.(12.46) is given in the following lemma

Lemma 48. *Consider the following AMP iteration*

$$\mathbf{u}^{t+1} = \tilde{\mathbf{Z}}^\top\mathbf{h}_t(\mathbf{v}^t) - \mathbf{e}_t(\mathbf{u}^t)\langle\mathbf{h}_t'\rangle^\top \tag{12.51}$$

$$\mathbf{v}^t = \tilde{\mathbf{Z}}\mathbf{e}_t(\mathbf{u}^t) - \mathbf{h}_{t-1}(\mathbf{v}^{t-1})\langle\mathbf{e}_t'\rangle^\top \tag{12.52}$$

where for any $t \in \mathbb{N}$

$$\mathbf{h}_t(\mathbf{v}^t) = \left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{S}^t}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}^t) - \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}^t\right)\right)(\mathbf{S}^t)^{-1} \tag{12.53}$$

$$\mathbf{e}_t(\mathbf{u}^t) = \mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}^t}\left(\mathbf{u}^t\hat{\mathbf{S}}^t + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}}^t + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \tag{12.54}$$

$$\textit{and} \quad \mathbf{S}^t = \langle(\mathbf{e}^t)'\rangle^\top, \quad \hat{\mathbf{S}}^t = -\left(\langle(\mathbf{h}^t)'\rangle^\top\right)^{-1} \tag{12.55}$$

Then the fixed point $(\mathbf{u}^\infty, \mathbf{v}^\infty)$ *of this iteration verifies*

$$\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}^\infty}\left(\mathbf{u}^\infty\hat{\mathbf{S}}^\infty + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}}^\infty + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} = \mathbf{U}^* \tag{12.56}$$

$$\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{S}^\infty}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}^\infty) - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} = \tilde{\mathbf{Z}}\mathbf{U}^* \tag{12.57}$$

where $\mathbf{U}^*$ *is the unique solution to the optimization problem Eq.*(12.46)*.*

Proof. To find the correct form of the non-linearities in the AMP iteration, we match the optimality condition of problem Eq.(12.46) with the generic form of the fixed point of the AMP iteration Eq.(10.13). In the subsequent derivation, we absorb the scaling $\frac{1}{\sqrt{d}}$ in the matrix $\tilde{\mathbf{Z}}$, such that its elements are i.i.d. $\mathcal{N}(0, 1/d)$, and omit time indices for simplicity. Going back to problem Eq. (12.46), its optimality condition reads :

$$\tilde{\mathbf{Z}}^\top\partial\tilde{\mathcal{L}}(\tilde{\mathbf{Z}}\mathbf{U}) + \partial\tilde{r}(\mathbf{U}) = 0 \tag{12.58}$$

For any pair of $K \times K$ symmetric positive definite matrices $\mathbf{S}, \hat{\mathbf{S}}$, this optimality condition is equivalent to

$$\tilde{\mathbf{Z}}^\top\left(\partial\tilde{\mathcal{L}}(\tilde{\mathbf{Z}}\mathbf{U})\mathbf{S} + \tilde{\mathbf{Z}}\mathbf{U}\right)\mathbf{S}^{-1} + \left(\partial\tilde{r}(\mathbf{U})\hat{\mathbf{S}} + \mathbf{U}\right)\hat{\mathbf{S}}^{-1} = \tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}\mathbf{U}\mathbf{S}^{-1} + \mathbf{U}\hat{\mathbf{S}}^{-1} \tag{12.59}$$

where we added the same quantity on both sides of the equality. For the loss function, we can then introduce the resolvent, formally D-resolvent:

$$\hat{\mathbf{v}} = \partial\tilde{\mathcal{L}}(\tilde{\mathbf{Z}}\mathbf{U})\mathbf{S} + \mathbf{Z}\mathbf{U} \iff \tilde{\mathbf{Z}}\mathbf{U} = \mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\hat{\mathbf{v}}) \tag{12.60}$$

such that

$$\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\hat{\mathbf{v}}) = (\mathrm{Id} + \partial\tilde{\mathcal{L}}(\bullet)\mathbf{S})^{-1}(\hat{\mathbf{v}}) = \underset{\mathbf{T}\in\mathbb{R}^{n\times K}}{\arg\min} \left\{ \tilde{\mathcal{L}}(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\hat{\mathbf{v}})\mathbf{S}^{-1}(\mathbf{T}-\hat{\mathbf{v}})^\top\right)\right\} \tag{12.61}$$

Similarly for the regularisation, introduce

$$\hat{\mathbf{u}} \equiv \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}}\right)(\mathbf{U}) \qquad \mathbf{U} = \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}) \tag{12.62}$$

where $\mathbf{S} \in \mathbb{R}^{K\times K}$ is a positive definite matrix, and

$$\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{v}}) = \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}}\right)^{-1}(\hat{\mathbf{v}}) = \underset{\mathbf{T}\in\mathbb{R}^{Kp\times K}}{\arg\min} \left\{ \tilde{r}(\mathbf{T}) + \frac{1}{2}\mathrm{tr}\left((\mathbf{T}-\hat{\mathbf{v}})\hat{\mathbf{S}}^{-1}(\mathbf{T}-\hat{\mathbf{v}})^\top\right)\right\} \tag{12.63}$$

where $\hat{\mathbf{S}} \in \mathbb{R}^{K\times K}$ is a positive definite matrix, and $\hat{\mathbf{v}} \in \mathbb{R}^{d\times K}$. The optimality condition Eq.(12.59) may then be rewritten as:

$$\tilde{\mathbf{Z}}^\top\left(\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\hat{\mathbf{v}}) - \hat{\mathbf{v}}\right)\mathbf{S}^{-1} = (\hat{\mathbf{u}} - \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}))\hat{\mathbf{S}}^{-1} \tag{12.64}$$

$$\tilde{\mathbf{Z}}\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}) = \mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\hat{\mathbf{v}}) \tag{12.65}$$

where both equations should be satisfied. We can now define update functions based on the previously obtained block decomposition. The fixed point of the matrix-valued AMP Eq.(10.13), omitting the time indices for simplicity, reads:

$$\mathbf{u} + \mathbf{e}(\mathbf{u})\langle\mathbf{h}'\rangle^\top = \tilde{\mathbf{Z}}^\top\mathbf{h}(\mathbf{v}) \tag{12.66}$$

$$\mathbf{v} + \mathbf{h}(\mathbf{v})\langle\mathbf{e}'\rangle^\top = \tilde{\mathbf{Z}}\mathbf{e}(\mathbf{u}) \tag{12.67}$$

Matching this fixed point with the optimality condition Eq.(12.64) suggests the following mapping:

$$\begin{aligned} \mathbf{h}(\mathbf{v}) &= \left(\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\mathbf{v}) - \mathbf{v}\right)\mathbf{S}^{-1}, & \mathbf{S} &= \langle\mathbf{e}'\rangle^\top, \\ \mathbf{e}(\mathbf{u}) &= \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}\hat{\mathbf{S}}), & \hat{\mathbf{S}} &= -(\langle\mathbf{h}'\rangle^\top)^{-1}, \end{aligned} \tag{12.68}$$

where we redefined $\hat{\mathbf{u}} \equiv \hat{\mathbf{u}}\hat{\mathbf{S}}$ in (12.62). We are now left with the task of evaluating the resolvents of $\tilde{\mathcal{L}}, \tilde{r}$ as expressions of the original functions $\mathcal{L}, r$. Starting with the loss function, we get

$$\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\mathbf{v}) = \underset{\mathbf{x}\in\mathbb{R}^{n\times K}}{\arg\min} \left\{ \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{x}\right) + \frac{1}{2}\mathrm{tr}\left((\mathbf{x}-\mathbf{v})\mathbf{S}^{-1}(\mathbf{x}-\mathbf{v})\right)^\top \right\} \tag{12.69}$$

letting $\tilde{\mathbf{x}} = \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{x}$, the problem is equivalent to

$$\begin{aligned}
\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\mathbf{v}) = \underset{\tilde{\mathbf{x}}\in\mathbb{R}^{n\times K}}{\arg\min}\Bigg\{ & \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \tilde{\mathbf{x}}\right) \\
& + \frac{1}{2}\mathrm{tr}\left((\tilde{\mathbf{x}} - (\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}))\mathbf{S}^{-1}(\tilde{\mathbf{x}} - (\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}))^\top\right)\Bigg\} \\
& - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} - \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} \qquad (12.70) \\
& = \mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{S}}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}) - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} - \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} \qquad (12.71)
\end{aligned}$$

and the corresponding non-linearity will then be

$$\mathbf{h}(\mathbf{v}) = \left(R_{\mathcal{L}(\mathbf{y},.),\mathbf{S}}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}) - \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}\right)\right)\mathbf{S}^{-1} \qquad (12.72)$$

Moving to the regularization, the resolvent reads

$$\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}) = \underset{\mathbf{x}\in\mathbb{R}^{Kp\times K}}{\arg\min}\left\{r\left(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{x}\right)\right) - \boldsymbol{\nu}^\top\mathbf{x}^\top\Omega^{-1/2}\mathbf{c} + \frac{1}{2}\mathrm{tr}\left((\mathbf{x}-\mathbf{u})\hat{\mathbf{S}}^{-1}(\mathbf{x}-\mathbf{u})^\top\right)\right\} \qquad (12.73)$$

letting $\tilde{\mathbf{x}} = \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{x}$, we obtain

$$\begin{aligned}
\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}) = \underset{\tilde{\mathbf{x}}\in\mathbb{R}^{Kp\times K}}{\arg\min}\Bigg\{ & r\left(\Omega^{-1/2}\tilde{\mathbf{x}}\right) - \boldsymbol{\nu}^\top\tilde{\mathbf{x}}^\top\Omega^{-1/2}\mathbf{c} \qquad (12.74) \\
& + \frac{1}{2}\mathrm{tr}\left((\tilde{\mathbf{x}} - \left(\mathbf{u} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\hat{\mathbf{S}}^{-1}(\tilde{\mathbf{x}} - \left(\mathbf{u} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top\right)\Bigg\} - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \qquad (12.75) \\
= \underset{\tilde{\mathbf{x}}\in\mathbb{R}^{Kp\times K}}{\arg\min}\Bigg\{ & r\left(\Omega^{-1/2}\tilde{\mathbf{x}}\right) \qquad (12.76) \\
& + \frac{1}{2}\mathrm{tr}\left((\tilde{\mathbf{x}} - \left(\mathbf{u} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\hat{\mathbf{S}}^{-1}(\tilde{\mathbf{x}} - \left(\mathbf{u} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top\right)\Bigg\} \qquad (12.77) \\
& - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \qquad (12.78) \\
\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}}\left(\mathbf{u} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) & - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \qquad (12.79)
\end{aligned}$$

Which gives the following non-linearity for the AMP iteration

$$e(\mathbf{u}) = \mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}}\left(\mathbf{u}\hat{\mathbf{S}} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\mathbf{V} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \qquad (12.80)$$

□

The following lemma then gives the exact asymptotics at each time step of the AMP iteration solving problem Eq.(12.46) : its *state evolution equations*.

Lemma 49. *Consider the AMP iteration Eq.(12.51-12.55). Assume it is initialized with* $\mathbf{u}^0$ *such that* $\lim_{d\to\infty} \frac{1}{d}\left\|\mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0)\right\|_{\mathrm{F}}$ *exists, a positive definite matrix* $\hat{\mathbf{S}}_0$*, and* $\mathbf{h}_{-1} \equiv 0$*. Then for any* $t \in \mathbb{N}$*, and any pair of seqeunces of uniformly pseudo-Lipschitz functions* $\phi_{1,n} : \mathbb{R}^{Kp\times K}$ *and* $\phi_{2,n} : \mathbb{R}^{n\times K}$*, the following holds*

$$\phi_{1,n}\left(\mathbf{u}^t\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{1,n}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}\right)\right] \tag{12.81}$$

$$\phi_{2,n}\left(\mathbf{v}^t\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{2,n}\left(\mathbf{H}(\mathbf{Q}^t)^{1/2}\right)\right] \tag{12.82}$$

where $\mathbf{G} \in \mathbb{R}^{Kp\times K}$ *and* $\mathbf{H} \in \mathbb{R}^{n\times K}$ *are independent random matrices with i.i.d. standard normal elements, and* $\mathbf{Q}^t, \hat{\mathbf{Q}}^t, \mathbf{V}^t, \hat{\mathbf{V}}^t$ *are given by the equations*

$$\mathbf{Q}^t = \frac{1}{p}\mathbb{E}\Bigg[\left(\mathbf{R}_{r(\Omega^{-1/2}.),(\hat{\mathbf{V}}^t)^{-1}}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top(\hat{\mathbf{V}}^t)^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top$$
$$\left(\mathbf{R}_{r(\Omega^{-1/2}.),(\hat{\mathbf{V}}^t)^{-1}}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top(\hat{\mathbf{V}}^t)^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\Bigg] \tag{12.83}$$

$$\hat{\mathbf{Q}}^t = \frac{1}{p}\mathbb{E}\Bigg[\left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}^{t-1}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2}\right)(\mathbf{V}^{t-1})^{-1}\right)^\top \tag{12.84}$$

$$\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}^{t-1}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2}\right)(\mathbf{V}^{t-1})^{-1}\Bigg] \tag{12.85}$$

$$\mathbf{V}^t = \frac{1}{p}\mathbb{E}\left[(\hat{\mathbf{Q}}^t)^{-1/2}\mathbf{G}^\top R_{r(\Omega^{-1/2}.),(\hat{\mathbf{V}}^t)^{-1}}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top(\hat{\mathbf{V}}^t)^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{12.86}$$

$$\hat{\mathbf{V}}^t = -\frac{1}{p}\mathbb{E}\Bigg[(\mathbf{Q}^{t-1})^{-1/2}\mathbf{H}^\top\left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}^{t-1}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2}\right)\right)$$
$$(\mathbf{V}^{t-1})^{-1}\Bigg] \tag{12.87}$$

Proof. Owing to the properties of Bregman proximity operators [24, 23], the update functions in the AMP iteration Eq.(12.51-12.55) are Lipschitz continuous. Thus under the assumptions made on the initialization, the assumptions of Theorem 19 are verified, which gives the desired result. □

Lemma 50. *Consider iteration Eq.(12.51-12.55), where the parameters* $\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{V}, \hat{\mathbf{V}}$ *are initialized at any fixed point of the state evolution equations of Lemma 49. For any sequence initialized with* $\hat{\mathbf{V}}_0 = \hat{\mathbf{V}}$ *and* $\mathbf{u}_0$ *such that*

$$\lim_{d\to\infty} \frac{1}{d}\mathbf{e}_0(\mathbf{u}_0)^\top \mathbf{e}_0(\mathbf{u}_0) = \mathbf{Q} \tag{12.88}$$

the following holds

$$\lim_{t\to\infty}\lim_{p\to\infty} \frac{1}{\sqrt{p}}\left\|\mathbf{u}^t - \mathbf{u}^\star\right\|_{\mathrm{F}} = 0 \quad \lim_{t\to\infty}\lim_{d\to\infty} \frac{1}{\sqrt{p}}\left\|\mathbf{v}^t - \mathbf{v}^\star\right\|_{\mathrm{F}} = 0 \tag{12.89}$$

Proof. The proof of this lemma is identical to that of Lemma 7 from [178]. □

Combining these results, we obtain the following asymptotic characterization of $\mathbf{U}^*$.

Lemma 51. *For any fixed* $\mathbf{m}$ *and* $\boldsymbol{\nu}$ *in their feasibility sets, let* $\mathbf{U}^*$ *be the unique solution to the optimization problem Eq.(12.46). Then, for any sequences (in the problem dimension) of pseudo-Lipschitz functions of order 2* $\phi_{1,n} : \mathbb{R}^{n\times K} \to \mathbb{R}$ *and* $\phi_{2,n} : \mathbb{R}^{Kp\times K} \to \mathbb{R}$*, the following holds*

$$\phi_{1,n}\left(\mathbf{U}^*\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{1,n}\left(\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{12.90}$$

$$\phi_{2,n}\left(\frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}^*\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{2,n}\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\hat{\mathbf{Q}}^{1/2}) - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} - \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}\right)\right] \tag{12.91}$$

where $\mathbf{G} \in \mathbb{R}^{Kp\times K}$ *and* $\mathbf{H} \in \mathbb{R}^{n\times K}$ *are independent random matrices with i.i.d. standard normal elements, and* $\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{V}, \hat{\mathbf{V}}$ *are given by the fixed point (assumed to be unique) of the following set of self consistent equations*

$$\mathbf{Q} = \frac{1}{p}\mathbb{E}\Big[\left(\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top \tag{12.92}$$

$$\left(\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\Big] \tag{12.93}$$

$$\hat{\mathbf{Q}} = \frac{1}{p}\mathbb{E}\Big[\left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)^\top$$

$$\left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)\Big] \tag{12.94}$$

$$\mathbf{V} = \frac{1}{p}\mathbb{E}\left[\hat{\mathbf{Q}}^{-1/2}\mathbf{G}^\top\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{12.95}$$

$$\hat{\mathbf{V}} = -\frac{1}{p}\mathbb{E}\left[\mathbf{Q}^{-1/2}\mathbf{H}^\top\left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)\right] \tag{12.96}$$

Proof. Combining the results of the previous lemmas, this proof is close to that of Theorem 1.5 in [29]. □

Returning to the optimization problem on $\mathbf{m}, \boldsymbol{\nu}$ in Eq.(12.45), the solution $\mathbf{U}^*$, at any dimension, verifies the zero gradient conditions on $\mathbf{m}, \boldsymbol{\nu}$:

$$\partial\boldsymbol{\nu} = 0 \iff \left(\mathbf{U}^*\right)^\top\tilde{\mathbf{c}} = 0 \tag{12.97}$$

$$\partial\mathbf{m} = 0 \iff \left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^\top\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right)$$

$$+ \frac{\sqrt{\gamma}\tilde{\mathbf{v}}^\top}{\rho_{\tilde{\mathbf{c}}}}\Omega^{-1/2}\partial r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) = 0 \tag{12.98}$$

Using Lemma 51 while assuming the subgradients of $\mathcal{L}, r$ are pseudo-Lipschitz (we discuss this assumption in subsection 12.1.4), we obtain for $\mathbf{m}$

$$\frac{1}{p}\mathbb{E}\left[\left(\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}+\Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^{\top}\hat{\mathbf{V}}^{-1}+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^{\top}}{\rho_{\tilde{\mathbf{c}}}}\right)-\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^{\top}}{\rho_{\tilde{\mathbf{c}}}}\right)^{\top}\tilde{\mathbf{c}}\right]=0 \tag{12.99}$$

$$\iff \mathbf{m}=\frac{1}{\sqrt{dp}}\mathbb{E}\left[\tilde{\mathbf{c}}^{\top}\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}+\Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^{\top}\hat{\mathbf{V}}^{-1}+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^{\top}}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{12.100}$$

and for $\boldsymbol{\nu}$

$$\frac{1}{p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^{\top}\partial\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right),\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(\mathbf{s}\frac{\mathbf{m}^{\top}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^{\top}}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}+\mathbf{H}\hat{\mathbf{Q}}^{1/2})\right) \tag{12.101}$$

$$+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}^{\top}}{\rho_{\tilde{\mathbf{c}}}}\Omega^{-1/2}\partial r\left(\Omega^{-1/2}\left(\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}+\Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^{\top}\hat{\mathbf{V}}^{-1}+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^{\top}}{\rho_{\tilde{\mathbf{c}}}}\right)\right)\right)\Bigg]=0 \tag{12.102}$$

Using the definition of D-resolvents, this is equivalent to

$$\frac{1}{p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^{\top}\left(Id-\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)\right)\left(\mathbf{s}\frac{\mathbf{m}^{\top}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^{\top}}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}+\mathbf{H}\hat{\mathbf{Q}}^{1/2}\right)\mathbf{V}^{-1} \tag{12.103}$$

$$+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}^{\top}}{\rho_{\tilde{\mathbf{c}}}}\left(Id-\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}(.)\right)\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}+\Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^{\top}\hat{\mathbf{V}}^{-1}+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^{\top}}{\rho_{\tilde{\mathbf{c}}}}\right)\hat{\mathbf{V}}\Bigg]=0 \tag{12.104}$$

which simplifies to

$$\boldsymbol{\nu}^{\top}=-\frac{1}{\sqrt{\gamma}p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^{\top}\left(Id-\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)\right)\left(\mathbf{s}\frac{\mathbf{m}^{\top}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^{\top}}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}+\mathbf{H}\hat{\mathbf{Q}}^{1/2}\right)\mathbf{V}^{-1}\Bigg] \tag{12.105}$$

which brings us to the following set of six self consistent equations

$$\mathbf{Q} = \frac{1}{p}\mathbb{E}\Bigg[\left(\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}} \left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \right)^\top \tag{12.106}$$

$$\left(\mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}} \left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \right) \Bigg] \tag{12.107}$$

$$\hat{\mathbf{Q}} = \frac{1}{p}\mathbb{E}\Bigg[\left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id \right) \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2} \right) \mathbf{V}^{-1} \right)^\top \tag{12.108}$$

$$\left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id \right) \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2} \right) \mathbf{V}^{-1} \right) \Bigg]$$

$$\mathbf{V} = \frac{1}{p}\mathbb{E}\left[\hat{\mathbf{Q}}^{-1/2}\mathbf{G}^\top \mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}} \left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \right) \right] \tag{12.109}$$

$$\hat{\mathbf{V}} = -\frac{1}{p}\mathbb{E}\left[\mathbf{Q}^{-1/2}\mathbf{H}^\top \left(\left(\mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id \right) \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2} \right) \mathbf{V}^{-1} \right) \right] \tag{12.110}$$

$$\mathbf{m} = \frac{1}{\sqrt{dp}}\mathbb{E}\left[\tilde{\mathbf{c}}^\top \mathbf{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}} \left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \right) \right] \tag{12.111}$$

$$\boldsymbol{\nu}^\top = -\frac{1}{\sqrt{\gamma}p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}} \right)^\top \left(Id - \mathbf{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) \right) \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\hat{\mathbf{Q}}^{1/2} \right) \mathbf{V}^{-1} \Bigg] \tag{12.112}$$

This set of equations then characterizes the asymptotic distribution of the estimator $\mathbf{U}^*$ in the sense of Lemma 51, with the optimal values of $\mathbf{m}$ and $\boldsymbol{\nu}$. Using the definition of $\mathbf{U}^*$ and $\tilde{\mathbf{Z}}\mathbf{U}^*$, along with the definition of the function r w.r.t. the original regularization function, a tedious but straightforward calculation allows reconstruct the asymptotic properties of $\mathbf{W}^*$ and of the set $\{\mathbf{X}_k\mathbf{w}_k^*\}_{k=1}^K$ given in the main text.

12.1.3 Relaxing the strong convexity constraint

Assuming the set of self consistent equations (12.106) have a unique fixed point regardless of the strong convexity assumption, this solution defines a unique set of six order parameters for the $\lambda_2 = 0$ case. Furthermore, using Proposition 12, the unique estimator $\mathbf{W}^*(\lambda_2)$ solving problem Eq.(12.5) for strictly positive λ_2 converges to the least-norm solution to the convex (but not strongly) Eq.(12.4). Thus, for any pseudo-Lipschitz observable of $\mathbf{U}^*(\lambda_2)$, we have, one the one side a continuous function of λ_2 with a unique continuous extension at $\lambda_2 = 0$, and on the other side a function of λ_2 prescribed by the expectation taken w.r.t. the asymptotic Gaussian model parametrised by the state evolution parameters which is defined for all positive values of λ_2. Since both functions match for any strictly positive λ_2, continuity implies they also match for $\lambda_2 = 0$ and we obtain the exact asymptotics of the least ℓ_2 norm solution of problem Eq.(12.4). Regarding the uniqueness of the solution to the fixed point equations (12.106), it is shown in [176] that a similar set of equations, although for a vector valued variable, i.e. no ensembling, the solution is unique even if the original problem is not strictly convex. This is proven by showing that the fixed point equations are the solution of a strictly convex problem. We expect this to be true here as well, and leave this part for a longer version of this paper.

12.1.4 A comment on non-pseudo-Lipschitz subgradients

Provided the subgradients in Eq.(12.97) are pseudo-Lipschitz continuous, the proof goes through. However some convex functions commonly used in machine learning, such as the hinge loss or the ℓ_1 norm for the penalty, have non-pseudo-Lipschitz gradient. To circumvent this issue, one can consider the optimization problem where both loss and regularization are replaced by their Moreau envelopes with strictly positive parameters τ_1, τ_2, as is done in [57] for the LASSO. Moreau envelopes are everywhere differentiable and have Lipschitz gradient for strictly positive values of their parameter [25], thus the asymptotic characterization holds. One can then take the parameters to zero, using the fact that the limit at zero in the parameters of Moreau envelopes is well defined [25], recovering the original function. Since proximity operators are defined as strongly convex problems, the sequence of problems defined by the proximal operator of a Moreau envelope with decreasing parameter converges to the proximal operator of the original function when the parameter is taken to zero. Finally, inverting the expectations on random quantities with the limit taking the parameters of the Moreau envelopes to zero can be done by verifying the dominated convergence theorem using the firm-nonexpansiveness of proximity operators and the corresponding bounds on their norms, see [25] Chapter 4, Section 1. We leave the details of this part to a longer version of this paper.

12.1.5 Toolbox

The required tools for this proof are the same as those given in section 10.1 of Chapter 10, with the added following lemma :

A useful result from convex analysis Here we remind a result from [25] describing the limiting behavior of regularized estimators for vanishing regularization.

Proposition 8. *(Theorem 26.20 from [25]) Let f and h be proper, lower semi-continuous, convex functions. Suppose that $\arg\min f \cap dom(g) \neq \emptyset$ and that h is coercive and strictly convex. Then g admits a unique minimizer $\mathbf{x}_0$ over $\arg\min f$ and , for every $\epsilon \in]0,1[$, the regularized problem*

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) + \epsilon h(\mathbf{x}) \tag{12.113}$$

admits a unique solution $\mathbf{x}_\epsilon$. If we assume further that h is uniformly convex on any closed ball of the input space, then $\lim_{\epsilon \to 0} \mathbf{x}_\epsilon = \mathbf{x}_0$.

Part III

Convex GLMs with left and right orthogonally invariant matrices

Chapter 13

How to prove Kabashima's replica formula

The results in this chapter are based on the paper [109]. Preliminary results on a simpler models were published in [108] and are not reproduced since they are included in the more general statement presented here. The reader curious to see how the formulas reduce on the simpler case of convex penalized least-squares regression may consult the main theorem and discussion in [108].

There has been a recent surge of interest in the study of asymptotic reconstruction performance in various cases of generalized linear estimation problems in the teacher-student setting, especially for the case of i.i.d standard normal matrices. Here, we go beyond these matrices, and prove an analytical formula for the reconstruction performance of convex generalized linear models with rotationally-invariant data matrices with arbitrary bounded spectrum, rigorously confirming, under suitable assumptions, a conjecture originally derived using the replica method from statistical physics. The proof is achieved by leveraging on message passing algorithms and the statistical properties of their iterates, allowing to characterize the asymptotic empirical distribution of the estimator. For sufficiently strongly convex problems, we show that the two-layer vector approximate message passing algorithm (2-MLVAMP) converges, where the convergence analysis is done by checking the stability of an equivalent dynamical system, which gives the result for such problems. We then show that, under a concentration assumption, an analytical continuation may be carried out to extend the result to convex (non-strongly) problems. We illustrate our claim with numerical examples on mainstream learning methods such as sparse logistic regression and linear support vector classifiers, showing excellent agreement between moderate size simulation and the asymptotic prediction.

13.1 Introduction

13.1.1 Background and motivation

In the modern era of statistics and machine learning, data analysis often requires solving high-dimensional estimation problems with a very large number of parameters. Developing algorithms for this task and understanding their limitations has become a major challenge. In this paper, we consider this question in the framework of supervised learning under the teacher-student scenario: (i) the data is synthetic and labels are generated by a "teacher"rule and (ii) training is done with

a convex Generalized Linear Model (GLM) . Such problems are ubiquitous in machine learning, statistics, communications, and signal processing.

The study of asymptotic (i.e. large-dimensional) reconstruction performance of generalized linear estimation in the teacher-student setting has been the subject of a significant body of work over the past few decades [263, 294, 94, 29, 90, 82, 300], and is currently witnessing a renewal of interest, especially for the case of identically and independently distributed (i.i.d.) standard normal data matrices, see e.g. [274, 125, 190]. The aim of this paper is to provide a general analytical formula describing the reconstruction performance of such convex generalized linear models, but for a broader class of more adaptable matrices.

The problem is defined as follows: we aim at reconstructing a given i.i.d. weight vector $\mathbf{x}_0 \in \mathbb{R}^N$ from outputs $\mathbf{y} \in \mathbb{R}^M$ generated using a training set $(\mathbf{f}_\mu)_{\mu=1,\dots,M}$ and the "teacher" rule:

$$\mathbf{y} = \varphi(\mathbf{F}\mathbf{x}_0, \boldsymbol{\omega_0}) \tag{13.1}$$

where φ is a proper, closed, continuous function and $\boldsymbol{\omega_0} \sim \mathcal{N}(0, \Delta_0 \mathrm{Id})$ is an i.i.d. noise vector. To go beyond the Gaussian i.i.d. case tackled in a majority of theoretical works, we shall allow matrices of arbitrary spectrum. We consider the data matrix $\mathbf{F} \in \mathbb{R}^{M\times N}$, obtained by concatenating the vectors of the training set, to be *rotationally invariant*: its singular value decomposition reads $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{M\times M}, \mathbf{V} \in \mathbb{R}^{N\times N}$ are uniformly sampled from the orthogonal groups $O(M)$ and $O(N)$ respectively. $\mathbf{D} \in \mathbb{R}^{M\times N}$ contains the singular values of $\mathbf{F}$ on its diagonal. Our analysis encompasses any singular value distribution with compact support. We place ourselves in the so-called high-dimensional regime, so that $M, N \to \infty$ while the ratio $\alpha \equiv M/N$ is kept finite. Our goal is to study the reconstruction performance of the generalized linear estimation method:

$$\hat{\mathbf{x}} \in \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \{g(\mathbf{F}\mathbf{x}, \mathbf{y}) + f(\mathbf{x})\} \tag{13.2}$$

where g and f are proper, closed, convex and separable functions. This type of procedure is an instance of empirical risk minimizationa and is one of the building blocks of modern machine learning. It encompasses several mainstream methods such as logistic regression, the LASSO or linear support vector machines. More precisely, the quantities of interest representing the reconstruction performance are the mean squared error $E = \mathbb{E}\left[\frac{1}{N}\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2\right]$ for regression problems, and the reconstruction angle $\theta_x = \arccos \frac{\mathbf{x}_0^T\hat{\mathbf{x}}}{\|\mathbf{x}_0\|_2\|\hat{\mathbf{x}}\|_2}$ for classification problems.

13.1.2 Main contributions

- We provide a set of equations characterizing the asymptotic statistical properties of the estimator defined by problem (13.2) with data generated by (13.1) in the asymptotic setup, for separable, convex losses and penalties (including for instance Logistic, Hinge, LASSO and Elastic net), for rotationally invariant sequences of matrices $\mathbf{F}$. For sufficiently strongly convex problems (in the sense of Lemma 54), our assumptions are classical with respect to earlier work. To extend the result to convex problems however, we require a concentration assumption that we discuss further in section 13.3.
- By doing so, we give, under the aforementioned set of assumptions, a mathematically rigorous proof, of a replica formula obtained heuristically through statistical physics for this problem, notably by Y. Kabashima[138]. This is a significant step beyond the setting of most rigorous work on replica results, which assume matrices to be i.i.d. random Gaussian ones.

- Our proof method builds on a detailed mapping between alternating directions descent methods [50] from convex optimization and a set of algorithms called multi-layer vector approximate message-passing algorithms [188, 256]. This enables us to use convergence results from convex analysis and dynamical systems to study the trajectories of vector approximate message-passing algorithms.
- Beyond the high-dimensional result on the estimator defined by the GLM, our convergence analysis provides a generic condition for the convergence of 2-layer MLVAMP, regardless of the randomness of the design matrix and of the dimensions of the problem, for sufficiently strongly convex problems.

13.1.3 Related work

The simplest case of the present question, when both f and g are quadratic functions, can be mapped to a random matrix theory problem and solved rigorously, as in e.g. [125]. Handling non-linearity is, however, more challenging. A long history of research tackles this difficulty in the high-dimensional limit, especially in the statistical physics literature where this setup is common. The usual analytical approach in statistical physics of learning [263, 294, 94] is a heuristic, non-rigorous but very adaptable technique called the replica method [196, 195]. In particular, it has been applied on many variations of the present problem, and laid the foundation of a large number of deep, non-trivial results in machine learning, signal processing and statistics, e.g. [103, 217, 39, 140, 101, 3, 205, 93]. Among them, a generic formula for the present problem has been conjectured by Y. Kabashima, providing sharp asymptotics for the reconstruction performance of the signal $\mathbf{x_0}$ [138].

Proving the validity of a replica prediction is a difficult task altogether. There has been recent progress in the particular case of Gaussian data, where the matrix $\mathbf{F}$ is made of i.i.d. standard Gaussian coefficients. In this case, the asymptotic performance of the LASSO was rigorously derived in [28], and the existence of the logistic estimator discussed in [274]. A set of papers managed to extend this study to a large set of convex losses g, using the so-called Gordon comparison theorem [281]. We broaden those results here by proving the Kabashima formula, valid for the set of rotationally invariant matrices introduced above and any convex, separable loss g and sufficiently strongly convex regularizer f under classical conditions. We extend this result to any convex, separable g and f under stronger assumptions.

Our proof strategy is based on the use of approximate-message-passing [83, 240], as pioneered in [29], and is similar to a recent work [109] on a simpler setting. This family of algorithms is a statistical physics-inspired variant of belief propagation [193, 137, 139] where local beliefs are approximated by Gaussian distributions. A key feature of these algorithms is the existence of the state evolution equations, a scalar equivalent model which allows to track the asymptotic statistical properties of the iterates at every time step. A series of groundbreaking papers initiated with [28] proved that these equations are exact in the large system limit, and extended the method to treat nonlinear problems [240] and handle rotationally invariant matrices [242, 277]. We shall use a variant of these algorithms called multi-layer vector approximate message-passing (MLVAMP) [256, 97]. The key technical point in our approach is an analysis of the convergence of MLVAMP. This is achieved by phrasing the algorithm as a dynamical system, and then determining sufficient conditions for convergence with linear rate. Our analysis guarantees converging trajectories above a threshold value of the strong convexity parameter of the problem, which is sufficient to complete the proof in that region. We use an analytic continuation to extend the result to convex problems, at the cost of an additional condition discussed after stating our main set of assumption.

13.2 Background on MLVAMP

In this section, we present background on the multilayer vector approximate message-passing algorithm developed in [97]. In doing so, we will introduce the key quantities involved in our main theorem. MLVAMP was initially designed as a probabilistic inference algorithm in multilayer architectures. Here, we only focus on the 2-layer version for inference in GLMs, and use the notations of [277]. The algorithm can be derived in several ways, notably from expectation-consistent variational inference frameworks such as expectation propagation [203], where the target posterior distribution is approximated by a simpler one with moment matching constraints. In the maximum a posteriori setting (MAP), the frequentist optimization framework is recovered, with additional parameter prescriptions due to the probabilistic models, as we will see below. The derivation of the algorithm is, however, not our point of interest. We focus on providing a self-contained interpretation from the convex optimization point of view, in particular in terms of variable splitting.

13.2.1 Link with variable splitting and proximal descent

A common procedure to tackle nonlinear optimization problems involving several functions is variable splitting, so that each non-linearity may be treated independently. Augmenting the Lagrangian with a square penalty on the slack variable equality constraint leads to the family of alternating direction methods of multipliers (ADMM) [50], where the objective is iteratively minimized in the direction of each initial variable and slack variable. The descent steps then take the form of proximal operators of the non-linearities. For example, on problem (13.2), adding a slack variable $\mathbf{z} = \mathbf{F}\mathbf{x}$ would lead to the augmented Lagrangian:

$$g(\mathbf{z}, \mathbf{y}) + f(\mathbf{x}) + \theta^T(\mathbf{z} - \mathbf{F}\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{z} - \mathbf{F}\mathbf{x}\|_2^2 \tag{13.3}$$

where $\alpha > 0$ is a free parameter that can enforce strong convexity of the objective if large enough and θ is a Lagrange multiplier. Updating $\mathbf{x}$ from an update on $\mathbf{z}$ amounts to a linear estimation problem, which can be solved by least squares. This is implemented, for example, in linearized ADMM [50], where the proximal descent steps are coupled to least-square ones.
MLVAMP solves problem (13.2) by introducing the same splitting as in (13.3) with an additional trivial splitting for each variable: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2$ such that $\mathbf{x}_1 = \mathbf{x}_2$, $\mathbf{z}_1 = \mathbf{F}\mathbf{x}_1$, $\mathbf{z}_2 = \mathbf{F}\mathbf{x}_2$. In the convex optimization framework, parameters like gradient step sizes, or proximal parameters need to be chosen. In the expectation propagation framework, they are prescribed by expectation-consistency constraints, which leads to additional steps in the algorithm. MLVAMP thus consists in four descent steps on $\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2$, and the updates on the parameters of the functions corresponding to those descent steps. This is shown in the MLVAMP iterations (see (1) further), where $\mathbf{x}_1, \mathbf{z}_1$ are updated using the proximal operators of the loss and regularizer, while $\mathbf{z}_2$ and $\mathbf{x}_2$ are obtained through least-squares. As mentioned above, the parameters of proximal operators (or denoisers in the signal processing literature) and least-squares are set by probabilistic inference rules (here moment-matching of marginal distributions). It is shown in [98] that, in the MAP setting, these updates amount to adapting the parameters to the local curvature of the cost function.

13.2.2 2-layer MLVAMP and its state evolution

We lay out the full iterations of the MLVAMP algorithm from [97] applied to a 2-layer network in Algorithm 1. For a given operator $T : \mathcal{X} \to \mathbb{R}^d$ where d is M or N in our setting, the brackets $\langle T(\mathbf{x}) \rangle = \frac{1}{d}\sum_{i=1}^{d} T(\mathbf{x})_i$ denote element-wise averaging operations. For a given matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, the

brackets amount to $\langle \mathbf{M} \rangle = \frac{1}{d}\mathbf{Tr}(\mathbf{M})$. For a given function, for example g_{1x}, we use the shorthand $g_{1x}(...)$ when the arguments have been made clear in a line above and are left unchanged. The

Algorithm 1 2-layer MLVAMP

Require: Initialize $\mathbf{h}_{1x}^{(0)}, \mathbf{h}_{2z}^{(0)}, \hat{Q}_{1x}^{(0)}, \hat{Q}_{2z}^{(0)}$, number of iterations T.

for t=0,1...,T **do**

// Denoising $\mathbf{x}$

$\hat{\mathbf{x}}_1^{(t)} = g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$

$\chi_{1x}^{(t)} = \left\langle \partial_{\mathbf{h}_{1x}^{(t)}} g_{1x}(...) \right\rangle / \hat{Q}_{1x}^{(t)}$

$\hat{Q}_{2x}^{(t)} = 1/\chi_{1x}^{(t)} - \hat{Q}_{1x}^{(t)}$

$\mathbf{h}_{2x}^{(t)} = (\hat{\mathbf{x}}_1^{(t)}/\chi_{1x}^{(t)} - \hat{Q}_{1x}^{(t)}\mathbf{h}_{1x}^{(t)})/\hat{Q}_{2x}^{(t)}$

// LMMSE estimation of $\mathbf{z}$

$\hat{\mathbf{z}}_2^{(t)} = g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})$

$\chi_{2z}^{(t)} = \left\langle \partial_{\mathbf{h}_{2z}^{(t)}} g_{2z}(...) \right\rangle / \hat{Q}_{2z}^{(t)}$

$\hat{Q}_{1z}^{(t)} = 1/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)}$

$\mathbf{h}_{1z}^{(t)} = (\hat{\mathbf{z}}_{\mathbf{2}}^{(\mathbf{t})}/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)}\mathbf{h}_{2z}^{(t)})/\hat{Q}_{1z}^{(t)}$

// Denoising $\mathbf{z}$

$\hat{\mathbf{z}}_1^{(t)} = g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)})$,

$\chi_{1z}^{(t)} = \left\langle \partial_{\mathbf{h}_{1z}^{(t)}} g_{1z}(...) \right\rangle / \hat{Q}_{1z}^{(t)}$

$\hat{Q}_{2z}^{(t+1)} = 1/\chi_{1z}^{(t)} - \hat{Q}_{1z}^{(t)}$

$\mathbf{h}_{2z}^{(t+1)} = (\hat{\mathbf{z}}_1^{(t)}/\chi_{1z}^{(t)} - \hat{Q}_{1z}^{(t)}\mathbf{h}_{1z}^{(t)})/\hat{Q}_{2z}^{(t+1)}$

// LMMSE estimation of $\mathbf{x}$

$\hat{\mathbf{x}}_2^{(t+1)} = g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)})$

$\chi_{2x}^{(t+1)} = \left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2x}(...) \right\rangle / \hat{Q}_{2x}^{(t)}$

$\hat{Q}_{1x}^{(t+1)} = 1/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)}$

$\mathbf{h}_{1x}^{(t+1)} = (\hat{\mathbf{x}}_2^{(t+1)}/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)})/\hat{Q}_{1x}^{(t+1)}$

end for

return $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2$

denoising functions g_{1x} and g_{1z} can be written as proximal operators in the MAP setting:

$$g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ f(\mathbf{x}) + \frac{\hat{Q}_{1x}^{(t)}}{2} \left\| \mathbf{x} - \mathbf{h}_{1x}^{(t)} \right\|_2^2 \right\} \tag{13.4}$$

$$= \mathrm{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)}) \tag{13.5}$$

and

$$g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}) = \underset{\mathbf{z}\in\mathbb{R}^M}{\arg\min} \left\{ g(\mathbf{y}, \mathbf{z}) + \frac{\hat{Q}_{1z}^{(t)}}{2} \left\| \mathbf{z} - \mathbf{h}_{1z}^{(t)} \right\|_2^2 \right\} \tag{13.6}$$

$$= \text{Prox}_{g(.,\mathbf{y})/\hat{Q}_{1z}^{(t)}}(\mathbf{h}_{1z}^{(t)}). \tag{13.7}$$

The LMMSE denoisers g_{2z} and g_{2x} in the MAP setting read (see [256]):

$$g_{2z}(...) = \mathbf{F}\mathbf{M}_1^{(t)}(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{h}_{2z}^{(t)}) \tag{13.8}$$

$$g_{2x}(...) = \mathbf{M}_2^{(t)}(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{h}_{2z}^{(t+1)}). \tag{13.9}$$

where we defined the matrices $\mathbf{M}_1^{(t)} = (\hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\text{Id})^{-1}$, and $\mathbf{M}_2^{(t)} = (\hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\text{Id})^{-1}$. As mentioned in the previous section, MLVAMP returns at each iteration two sets of estimators $(\hat{\mathbf{x}}_1^{(t)}, \hat{\mathbf{x}}_2^{(t)})$ and $(\hat{\mathbf{z}}_1^{(t)}, \hat{\mathbf{z}}_2^{(t)})$ which respectively aim at reconstructing the minimizer $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. At the fixed point, we have $\hat{\mathbf{x}}_1^{(t)} = \hat{\mathbf{x}}_2^{(t)}$ and $\hat{\mathbf{z}}_1^{(t)} = \hat{\mathbf{z}}_2^{(t)}$, as proven in [222]. The intermediate vectors $\mathbf{h}_{1x}^{(t)}$, $\mathbf{h}_{2x}^{(t)}$, $\mathbf{h}_{1z}^{(t)}$ and $\mathbf{h}_{2z}^{(t)}$ have the key feature that they behave asymptotically as Gaussian centered around $\mathbf{x}_0$ and $\mathbf{z}_0 = \mathbf{F}\mathbf{x}_0$, under the set of assumptions given in appendix 14.5.2. More precisely, at each iteration, they converge empirically with second order moment (PL2) towards Gaussian variables:

$$\lim_{M,N\to\infty} \hat{Q}_{1x}^{(t)}\mathbf{h}_{1x}^{(t)} - \hat{m}_{1x}^{(t)}\mathbf{x_0} \overset{PL(2)}{=} \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)} \tag{13.10a}$$

$$\lim_{M,N\to\infty} \mathbf{V}^T(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} - \hat{m}_{2x}^{(t)}\mathbf{x_0}) \overset{PL(2)}{=} \sqrt{\hat{\chi}_{2x}^{(t)}}\xi_{2x}^{(t)} \tag{13.10b}$$

$$\lim_{M,N\to\infty} \mathbf{U}^T(\hat{Q}_{1z}^{(t)}\mathbf{h}_{1z}^{(t)} - \hat{m}_{1z}^{(t)}\mathbf{z_0}) \overset{PL(2)}{=} \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)} \tag{13.10c}$$

$$\lim_{M,N\to\infty} \hat{Q}_{2z}^{(t)}\mathbf{h}_{2z}^{(t)} - \hat{m}_{2z}^{(t)}\mathbf{z_0} \overset{PL(2)}{=} \sqrt{\hat{\chi}_{2z}^{(t)}}\xi_{2z}^{(t)} \tag{13.10d}$$

where $\xi_{1x}^{(t)}, \xi_{2x}^{(t)}, \xi_{1z}^{(t)}, \xi_{2z}^{(t)}$ are i.i.d standard normal random variables independent of all other quantities. The definition of PL(2) convergence is reminded in Appendix 14.1, and we use the notation $\overset{PL(2)}{=}$ following [242, 97]. We can roughly say that the $\hat{Q}, \hat{m}, \hat{\chi}$'s parameters characterize the distributions of the $\mathbf{h}$'s. Using the representation (13.10) in the iterations of MLVAMP results in a scalar recursion that tracks the evolution of the parameters of the aforementioned Gaussian distributions. This recursion provides the so-called state evolution equations. The existence of state evolution equations is the reason why we use 2-layer MLVAMP in our proof. Indeed, they allow the construction of iterate paths that lead to the solution of problem (13.2), while knowing their statistical properties.

13.3 Main result

Our main result characterizes the asymptotic empirical distribution of the estimator $\hat{\mathbf{x}}$ defined in (13.2) with data generated by (13.1), and of $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. We start by stating the necessary assumptions.

Assumption 2.

(a) the functions f and g are proper, closed, convex and separable functions.

(b) the cost function $g(\mathbf{F}.,\mathbf{y}) + f(.)$ is coercive, i.e. $\lim_{\|\mathbf{x}\|\to\infty} g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) = +\infty$.

(c) there exists a finite constant B_1 such that $\frac{1}{N}\|\hat{\mathbf{x}}\|_2^2 \leqslant B_1$ almost surely as $N \to \infty$. We also assume that, for any pseudo-Lipschitz function of order 2, if there exists a finite constant B_2 such that $\forall N \in \mathbb{N}$, $\frac{1}{N}\sum_{i=1}^N \phi(\hat{x}_i) \leqslant B_2$, then the limit $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \phi(\hat{x}_i)$ exists.

(d) for any $\mathbf{x} \in dom(f)$ and any $\mathbf{x}' \in \partial f(\mathbf{x})$, there exists a constant C such that $\|\mathbf{x}'\|_2 \leqslant C(1 + \|\mathbf{x}\|_2)$. The same holds for g on its domain.

(e) there exist sequences of real analytic functions g_ϵ, f_ϵ such that for any x, $\lim_{\epsilon\to 0} g_\epsilon(x) = g(x)$, $\lim_{\epsilon\to 0} f_\epsilon(x) = f(x)$, and for all $\epsilon > 0$, g_ϵ'' and f_ϵ'' belong to the Schwartz space.

(f) the empirical distributions of the underlying truth $\mathbf{x}_0$, eigenvalues of $\mathbf{F}^T\mathbf{F}$, and noise vector w_0, respectively converge empirically with second order moments, as defined in appendix 14.1, to independent scalar random variables x_0, w_0, λ with distributions p_{x_0}, p_λ, p_{w_0}. We assume that the distribution p_λ is not all-zero and has compact support.

(g) the design matrix $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{M\times N}$ is rotationally invariant, as defined in the introduction, where the elements of the Haar distributed matrices $\mathbf{U}, \mathbf{V}$ are independent of the elements of the ground truth vector $\mathbf{x}_0$, noise $\boldsymbol{\omega}_0$ and elements of $\mathbf{D}$.

(h) the solution to the set of fixed point equations (13.13) exists and is unique, for any convex g and f verifying the assumptions above

(i) finally assume that $M, N \to \infty$ with fixed ratio $\alpha = M/N$.

The coercivity assumption (b) ensures that the minimization problem Eq.(13.2) is feasible and that the estimator exists. Most machine learning cost functions verify this assumption, including any convex loss which is bounded below and regularized with a coercive term such as the ℓ_1 or ℓ_2 norm, see [25] Corollary 11.15. Non-coercive problems include unregularized logistic regression and unregularized, underspecified least-squares for example. The scaling assumptions (d) are required for the state evolution equations of the MLVAMP iteration corresponding to the optimization problem Eq.(13.2) to hold, as discussed in appendix 14.5.2. Such conditions are often encountered in high dimensional analysis of M-estimators, see, e.g. [281], and are verified by the setups proposed in the experiments section. The convergence of averaged sumes of PL2 observables in assumption (c) and the analytic approximation in assumption (e) are required for our analytic continuation to hold, and we show that any combination of hinge, logistic and square loss with ℓ_1 or ℓ_2 regularization verifies the latter in Appendix 14.8, subsection 14.8.6. We show in Lemma 55 that, for sufficiently strongly convex problems, these two assumptions are not required. The concentration assumption we require has been proven to hold for a number of convex problems with Gaussian random design regardless of the strong convexity of the problem (see the related work section), and we believe rotationally invariant matrices do not change this behaviour. However, since we are unable to prove it below the threshold value of the strong convexity parameter, it remains an assumption. Additional detail on the notion of empirical convergence is given in appendix 14.1. This analysis framework is mainly due to [28] and is related to convergence in Wasserstein metric as pointed out in [93]. We are now ready to state our main theorem.

Theorem 22 (Fixed point equations)**.** *Under assumption 2, consider the ground-truth* $\mathbf{x_0}$ *and let* $\mathbf{z_0} = \mathbf{F}\mathbf{x_0}$*,* $\rho_x \equiv \|\mathbf{x_0}\|_2^2/N$ *and* $\rho_z \equiv \|\mathbf{z_0}\|_2^2/M$*. For a strictly convex instance of problem (13.2), let* $\hat{\mathbf{x}}$ *be its unique solution. For a convex (non-strictly) instance of problem (13.2), let* $\hat{\mathbf{x}}$ *be its unique least* ℓ_2 *norm solution. Then let* $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$*. Then, for any real analytic, pseudo-Lipschitz function of order 2* ϕ *whose second derivative belongs to the Schwartz space, the following holds :*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \phi(x_{0,i}, \hat{x}_i) \overset{a.s.}{=} \mathbb{E}[\phi(x_0, Prox_{f/\hat{Q}_{1x}^{(*)}}(H_x))] \tag{13.11}$$

$$\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \phi(z_{0,i}, \hat{z}_i) \overset{a.s.}{=} \mathbb{E}[\phi(z_0, Prox_{f/\hat{Q}_{1z}^{(*)}}(H_z))] \tag{13.12}$$

where $H_x = \frac{\hat{m}_{1x}^* x_0 + \sqrt{\hat{\chi}_{1x}^*}\xi_{1x}}{\hat{Q}_{1x}^*}$*,* $H_z = \frac{\hat{m}_{1z}^* z_0 + \sqrt{\hat{\chi}_{1z}^*}\xi_{1z}}{\hat{Q}_{1z}^*}$ *and expectations are taken with respect to the random variables* $x_0 \sim p_{x_0}$*,* $z_0 \sim \mathcal{N}(0, \sqrt{\rho_z})$*,* $\xi_{1x}, \xi_{1z} \sim \mathcal{N}(0,1)$*.*
The parameters $\hat{Q}_{1x}^*, \hat{Q}_{1z}^*, \hat{m}_{1x}^*, \hat{m}_{1z}^*, \hat{\chi}_{1x}^*, \hat{\chi}_{1z}^*$ *are determined by the fixed point of the system:*

$$\hat{Q}_{2x} = \hat{Q}_{1x}(\mathbb{E}\left[\eta'_{f/\hat{Q}_{1x}}(H_x)\right]^{-1} - 1) \tag{13.13a}$$

$$\hat{Q}_{2z} = \hat{Q}_{1z}(\mathbb{E}\left[\eta'_{g(.,y)/\hat{Q}_{1z}}(H_z)\right]^{-1} - 1) \tag{13.13b}$$

$$\hat{m}_{2x} = \frac{\mathbb{E}\left[x_0\eta_{f/\hat{Q}_{1x}}(H_x)\right]}{\rho_x\chi_x} - \hat{m}_{1x} \tag{13.13c}$$

$$\hat{m}_{2z} = \frac{\mathbb{E}\left[z_0\eta_{g(.,y)/\hat{Q}_{1z}}(H_z)\right]}{\rho_z\chi_z} - \hat{m}_{1z} \tag{13.13d}$$

$$\hat{\chi}_{2x} = \frac{\mathbb{E}\left[\eta^2_{f/\hat{Q}_{1x}}(H_x)\right]}{\chi_x^2} \tag{13.13e}$$

$$- \rho_x(\hat{m}_{1x} + \hat{m}_{2x})^2 - \hat{\chi}_{1x} \tag{13.13f}$$

$$\hat{\chi}_{2z} = \frac{\mathbb{E}\left[\eta^2_{g(.,y)/\hat{Q}_{1z}}(H_z)\right]}{\chi_z^2} \tag{13.13g}$$

$$- \rho_z(\hat{m}_{1z} + \hat{m}_{2z})^2 - \hat{\chi}_{1z} \tag{13.13h}$$

$$\hat{Q}_{1x} = \mathbb{E}\left[\frac{1}{\hat{Q}_{2x} + \lambda \hat{Q}_{2z}}\right]^{-1} - \hat{Q}_{2x} \tag{13.13i}$$

$$\hat{Q}_{1z} = \alpha \mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x} + \lambda \hat{Q}_{2z}}\right]^{-1} - \hat{Q}_{2z} \tag{13.13j}$$

$$\hat{m}_{1x} = \frac{1}{\chi_x}\mathbb{E}\left[\frac{\hat{m}_{2x} + \lambda \hat{m}_{2z}}{\hat{Q}_{2x} + \lambda \hat{Q}_{2z}}\right] - \hat{m}_{2x} \tag{13.13k}$$

$$\hat{m}_{1z} = \frac{\rho_x}{\alpha \chi_z \rho_z}\mathbb{E}\left[\frac{\lambda(\hat{m}_{2x} + \lambda \hat{m}_{2z})}{\hat{Q}_{2x} + \lambda \hat{Q}_{2z}}\right] - \hat{m}_{2z} \tag{13.13l}$$

$$\hat{\chi}_{1x} = \frac{1}{\chi_x^2}\mathbb{E}\left[\frac{\hat{\chi}_{2x} + \lambda \hat{\chi}_{2z} + \rho_x(\hat{m}_{2x} + \lambda \hat{m}_{2z})^2}{(\hat{Q}_{2x} + \lambda \hat{Q}_{2z})^2}\right] - \rho_x(\hat{m}_{1x} + \hat{m}_{2x})^2 - \hat{\chi}_{2x} \tag{13.13m}$$

$$\hat{\chi}_{1z} = \frac{1}{\alpha \chi_z^2}\mathbb{E}\left[\frac{\lambda(\hat{\chi}_{2x} + \lambda \hat{\chi}_{2z} + \rho_x(\hat{m}_{2x} + \lambda \hat{m}_{2z})^2)}{(\hat{Q}_{2x} + \lambda \hat{Q}_{2z})^2}\right] - \rho_z(\hat{m}_{1z} + \hat{m}_{2z})^2 - \hat{\chi}_{2z}, \tag{13.13n}$$

where $\chi_x = (\hat{Q}_{1x} + \hat{Q}_{2x})^{-1}$, $\chi_z = (\hat{Q}_{1z} + \hat{Q}_{2z})^{-1}$, *and expectations are taken with respect to the random variables* $x_0 \sim p_{x_0}$, $z_0 \sim \mathcal{N}(0, \sqrt{\rho_z})$, $y \sim \varphi(z_0, \omega_0)$, $\xi_{1x}, \xi_{1z} \sim \mathcal{N}(0,1)$, *and eigenvalues* $\lambda \sim p_\lambda$. η *is a shorthand for the scalar proximal operator:*

$$\eta_{\gamma f}(z) = \arg\min_{x \in \mathcal{X}} \left\{ \gamma f(x) + \frac{1}{2}(x - z)^2 \right\}. \tag{13.14}$$

The set of fixed point equations from Theorem 22 naturally stems from the "replica-symmetric" free energy commonly used in the statistical physics community [196, 195]. The free energy depends on a set of parameters, and extremizing it with respect to all parameters, i.e. writing the zero gradient condition for each parameter, provides the set of equations (13.13). We state this correspondence in the following corollary to Theorem 22 :

Corollary 6 (The Kabashima formula)**.**
The fixed point equations from theorem 22 can equivalently be rewritten as the solution of the extreme value problem (13.15) defined by the replica free energy from [277].

β is a parameter that corresponds in the physics approach to an inverse temperature. In the $\beta \to \infty$ limit (the so-called zero temperature limit), the integrals defining ϕ_x and ϕ_z concentrate on their extremal value. Note that they are closely related to the Moreau envelopes $\mathcal{M}$ [224, 25] of f and g, which represent a smoothed form of the objective function with the same minimizers:

$$\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) = \frac{\hat{Q}_{1x}}{2} H_x^2 - \mathcal{M}_{\frac{f}{\hat{Q}_{1x}}}(H_x) \tag{13.18}$$

$$\text{where } \forall\, \gamma \geqslant 0,\ \mathcal{M}_{\gamma f}(z) = \inf_x \left\{ f(x) + \frac{1}{2\gamma}\|x - z\|_2^2 \right\}, \tag{13.19}$$

We provide details on this correspondence in appendix 14.3. In the zero-temperature limit we consider, it is possible to have more precise information on the geometry of the cost function defining

$$
\begin{aligned}
f &= -\operatorname*{extr}_{m_x,\chi_x,q_x,m_z,\chi_z,q_z} \{g_{\rm F} + g_{\rm G} - g_{\rm S}\}, \qquad (13.15)\\
g_{\rm F} &= \operatorname*{extr}_{\hat{m}_{1x},\hat{\chi}_{1x},\hat{Q}_{1x},\hat{m}_{1z},\hat{\chi}_{1z},\hat{Q}_{1z}} \left\{ \frac{1}{2} q_x \hat{Q}_{1x} - \frac{1}{2}\chi_x\hat{\chi}_{1x} - \hat{m}_{1x} m_x - \alpha \hat{m}_{1z} m_z + \frac{\alpha}{2}\left(q_z\hat{Q}_{1z} - \chi_z\hat{\chi}_{1z}\right) \right.\\
&\quad \left. + \mathbb{E}\left[\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x})\right] + \alpha\mathbb{E}\left[\phi_z(\hat{m}_{1z},\hat{Q}_{1z},\hat{\chi}_{1z};z_0,\xi_{1z})\right]\right\},\\
g_{\rm G} &= \operatorname*{extr}_{\hat{m}_{2x},\hat{\chi}_{2x},\hat{Q}_{2x},\hat{m}_{2z},\hat{\chi}_{2z},\hat{Q}_{2z}} \left\{ \frac{1}{2} q_x \hat{Q}_{2x} - \frac{1}{2}\chi_x\hat{\chi}_{2x} - m_x \hat{m}_{2x} - \alpha m_z \hat{m}_{2z} + \frac{\alpha}{2}\left(q_z\hat{Q}_{2z} - \chi_z\hat{\chi}_{2z}\right) \right.\\
&\quad \left. -\frac{1}{2}\left(\mathbb{E}\left[\log\left(\hat{Q}_{2x}+\lambda\hat{Q}_{2z}\right)\right] - \mathbb{E}\left[\frac{\hat{\chi}_{2x}+\lambda\hat{\chi}_{2z}}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right] - \mathbb{E}\left[\frac{\rho_x(\hat{m}_{2x}+\lambda\hat{m}_{2z})^2}{(\hat{Q}_{2x}+\lambda\hat{Q}_{2z})}\right]\right)\right\},\\
g_{\rm S} &= \frac{1}{2}\left(\frac{q_x}{\chi_x} - \frac{m_x^2}{\rho_x\chi_x}\right) + \frac{\alpha}{2}\left(\frac{q_z}{\chi_z} - \frac{m_z^2}{\rho_z\chi_z}\right),
\end{aligned}
$$

where ϕ_x and ϕ_z are the potential functions

$$\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x}) = \lim_{\beta\to\infty}\frac{1}{\beta}\log\int e^{-\frac{\beta\hat{Q}_{1x}}{2}x^2+\beta(\hat{m}_{1x}x_0+\sqrt{\hat{\chi}_{1x}}\xi_{1x})x-\beta f(x)}dx, \qquad (13.16)$$

$$\phi_z(\hat{m}_{1z},\hat{Q}_{1z},\hat{\chi}_{1z};z_0,\chi_{1z}) = \lim_{\beta\to\infty}\frac{1}{\beta}\log\int e^{-\frac{\beta\hat{Q}_{1z}}{2}z^2+\beta(\hat{m}_{1z}z_0+\sqrt{\hat{\chi}_{1z}}\xi_{1z})z-\beta g(y,z)}dz. \qquad (13.17)$$

---

the optimization problem in Corollary 6. Indeed, it is composed of functions whose convexity or concavity are staightforward to establish : linear terms, inverses, logarithms, squares and expectation of Moreau envelopes. The convexity of the latter is well documented in [281]. First, note that the parameters $\chi_x,\chi_z,\hat{\chi}_{1x},\hat{\chi}_{2x},\hat{\chi}_{1z},\hat{\chi}_{2z},q_x,q_z,\hat{Q}_{1x},\hat{Q}_{2x},\hat{Q}_{1z},\hat{Q}_{2z}$ are positive so we may restrict their feasibility set to $\mathbb{R}^+$, while $m_x,m_z,\hat{m}_{1x},\hat{m}_{1z},\hat{m}_{2x},\hat{m}_{2z}$ can take any value in $\mathbb{R}$. Then, $q_x^* = \frac{1}{N}\|\hat{\mathbf{x}}\|^2$ and $m_x^* = \frac{1}{N}\mathbf{x}_0^\top\hat{\mathbf{x}}$. The Cauchy-Schwarz inequality thus gives $q_x^* \geqslant \frac{(m_x^*)^2}{\rho_x}$. Similarly with $\hat{\mathbf{z}}$, $q_z^* \geqslant \frac{(m_z^*)^2}{\rho_z}$. We may thus restrict the feasibility sets of q_x,q_z,m_x,m_z such that they verify these inequalities. In these regions, the function g_s is convex in χ_x,χ_z, linear in q_x,q_z and concave in m_x,m_z. The terms involving $q_x,q_z,m_x,m_z,\chi_x,\chi_z$ in g_G and g_F are all linear. Moving to g_g, the cost function defining it is convex in $\hat{Q}_{2x},\hat{Q}_{2z}$ (negative logarithm and inverse function on $\mathbb{R}^+$), linear in $\hat{\chi}_{2x},\hat{\chi}_{2z}$ and convex in $\hat{m}_{2x},\hat{m}_{2z}$. Regarding g_F, all terms are linear except for the replica potentials. Using Moreau's identity, we may write $\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x}) = \mathcal{M}_{\hat{Q}_{1x}f^*}\left(\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}\right)$ where f^* is the conjugate of f. Using the properties summarized in [281], the cost function defining g_F is convex in $\hat{m}_{1x},\hat{m}_{1z},\hat{Q}_{1x},\hat{Q}_{1z}$. The convexity with respect to χ_{1x},χ_{1z} is harder to characterize due to the composition of the Moreau envelope with the square root, and should be studied locally for more information. The extremization may then be rewritten as a maximization over the variables in which the cost function is concave and minimization over the variables in which the cost function is convex. Note that this does not give information on the uniqueness of the solution, which would require joint strict convexity and strict concavity.

As immediate corollaries to Theorem 22, we can determine the asymptotic errors of the GLM and the optimal value of the loss function. To characterize the asymptotic reconstruction errors and angles, we can define the norms of the estimators and their overlaps with the ground-truth vectors

as the limits

$$m_x^* \equiv \lim_{N\to\infty} \frac{\hat{\mathbf{x}}^T \mathbf{x}_0}{N} \qquad m_z^* \equiv \lim_{M\to\infty} \frac{\hat{\mathbf{z}}^T \mathbf{z}_0}{M} \tag{13.20}$$

$$q_x^* \equiv \lim_{N\to\infty} \frac{\|\hat{\mathbf{x}}\|_2^2}{N} \qquad q_z^* \equiv \lim_{N\to\infty} \frac{\|\hat{\mathbf{z}}\|_2^2}{M}. \tag{13.21}$$

We then have :

Corollary 7.
Under the set of Assumptions 2, the squared norms m_x^*, m_z^* of estimator $\hat{\mathbf{x}}$ defined by (13.2) and $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$, and their overlaps q_x^*, q_z^* with ground-truth vectors are almost surely given by:

$$m_x^* = \mathbb{E}\left[x_0 \eta_{\frac{f}{\hat{Q}_{1x}^*}}(H_x)\right], \; q_x^* = \mathbb{E}\left[\eta^2_{\frac{f}{\hat{Q}_{1x}^*}}(H_x)\right] \tag{13.22}$$

$$m_z^* = \mathbb{E}\left[z_0 \eta_{\frac{g(.,y)}{\hat{Q}_{1z}^*}}(H_z)\right], \; q_z^* = \mathbb{E}\left[\eta^2_{\frac{g(.,y)}{\hat{Q}_{1z}^*}}(H_z)\right] \tag{13.23}$$

with H_x and H_z defined as in Theorem 22.

With the knowledge of the asymptotic overlap m_x^*, and squared norms q_x^*, ρ_x, most quantities of interest can be determined. For instance, the quadratic reconstruction error is obtained from its definition as $\mathrm{E} = \rho_x + q_x^* - 2m_x^*$, while the angle between the ground-truth vector and the estimator is $\theta = \arccos(m_x^*/(\sqrt{\rho_x q_x^*}))$. One can also evaluate the generalization error for new random Gaussian samples, as advocated in [94], or compute similar errors for the denoising of $\mathbf{z_0}$.

13.4 Numerical results

Obtaining a stable implementation of the fixed point equations can be challenging. We provide simulation details in appendix 14.6 along with a link to the script we used to produce the figures. Theoretical predictions (full lines) are compared with numerical experiments (points) conducted using standard convex optimization solvers from [229]. The comparison with finite size ($N \equiv$ a few hundreds) numerical experiments shows that, despite being asymptotic in nature, the predictions are accurate even at moderate system sizes. All experimental points were done with $N = 200$ and averaged one hundred times.

13.4.1 Validity of the replica prediction

We start with a simple verification of the replica prediction in Figure13.1, on a classification problem where data is generated as $\mathbf{y} = \mathrm{sign}(\mathbf{F}\mathbf{x_0})$. We consider two types of singular value distributions for $\mathbf{F}$ and three types of losses: a square loss, a linear support vector classification (SVC) loss and a logistic loss. Technical details and expressions are given in appendix 14.6. We use ridge regularization with penalty $f = \frac{\lambda_2}{2}\|\cdot\|_2^2$. We plot the reconstruction angle θ as a function of the aspect ratio of the problem α in Figure 13.1. A first plot is done with a Marchenko-Pastur eigenvalue distribution for $\mathbf{F}^T\mathbf{F}$ corresponding to $\mathbf{F}$ being i.i.d Gaussian. We then move out of the Gaussian setting and change the eigenvalue distribution for (14.76), which has a qualitatively similar behaviour: it has bounded support, and includes vanishing singular values at a given value $\alpha = 1$ of the aspect ratio. We recover a result close to the i.i.d. Gaussian one, including the error peak for the square loss

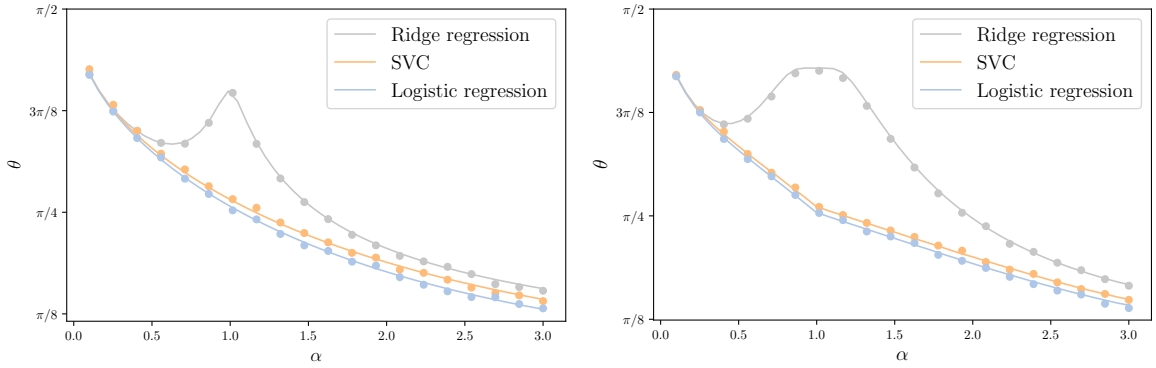


Figure 13.1: Illustration of Theorem 22 in a binary classification problem with data generated as $\mathbf{y} = \phi(\mathbf{F}\mathbf{x}_0)$ with the data matrix $\mathbf{F}$ being **Left :** a Gaussian i.i.d. matrix and **Right :** a random orthogonal invariant matrix with a squared uniform density of singular values. We plot the angle between the estimator and the ground-truth vector $\theta = \arccos(m_x^*/(\sqrt{\rho_x q_x^*}))$ as a function of the aspect ratio $\alpha = M/N$ with three different losses: ridge regression, a Support Vector Machine with linear kernel and a logistic regression. f is a ℓ_2 penalty with parameter $\lambda_2 = 10^{-3}$. The theoretical prediction (full line) is compared with numerical experiments (points) conducted using standard convex optimization solvers from [229].

when $\alpha = 1$. In both cases, the SVC and the logistic regression perform similarly. Note that error peaks can also be obtained for the max-margin solution as shown in [106], using a more elaborate teacher.

13.4.2 Sparse logistic regression

We now use the replica prediction to study sparse logistic regression with i.i.d Gaussian and row-orthogonal data, the latter being ubiquitous in signal processing. Row-orthogonal data gives rise to a discrete eigenvalue distribution for $\mathbf{F}^T\mathbf{F}$ of zeroes and ones:

$$\lambda_{\mathbf{F}^T\mathbf{F}} \sim \max(0, 1-\alpha)\delta(0) + \min(1, \alpha)\delta(1) \tag{13.24}$$

and is often found to outperform Gaussian sensing matrices for recovery tasks, see e.g. [140] or [109]. In what follows, we define the sparsity ρ of the ground truth vector as the fraction of non-zero components which are sampled from a standard normal distribution. Labels are generated with $\mathbf{y} = \text{sign}(\mathbf{F}\mathbf{x_0})$ as for Figure 13.1.

Effect of sparsity

In Figure 13.2, we start by plotting the reconstruction angle against the aspect ratio of the measurement matrix for different values of the sparsity of the teacher vector, for ℓ_2 regularization $f = \frac{\lambda_2}{2}\|\cdot\|_2^2$ and ℓ_1 regularization $f = \lambda_1\|\cdot\|_1$, and a fixed value of regularization parameters λ_1, λ_2. In the case of ℓ_2-regularization, we observe that the reconstruction performance remains the same whatever the sparsity of the original teacher vector as all curves collapse together (top and bottom left). The ridge regularization is thus unable to differentiate sparse and non-sparse problems. For ℓ_1, better performance is observed when the sparsity increases. Comparing the values for ℓ_2 and ℓ_1 also shows

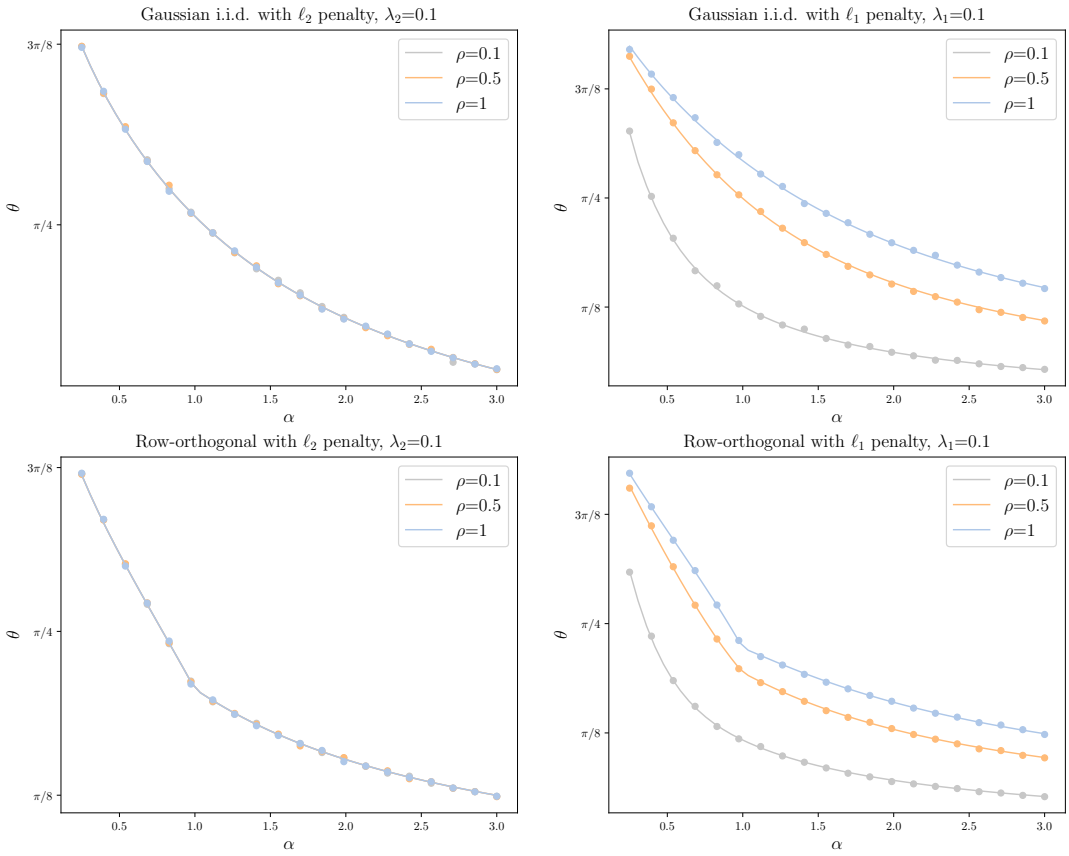


Figure 13.2: Effect of the sparsity of the planted vector. We plot the angle between the estimator and the ground truth in a binary classification problem with $\mathbf{y} = \text{sign}(\mathbf{F}\mathbf{x}_0)$ as a function of $\alpha = M/N$, for different values of sparsity ρ. We use logistic regression. Figures in the top are for $\mathbf{F}$ Gaussian i.i.d., while figures in the bottom are for $\mathbf{F}$ row-orthogonal. **Left :** we use a ℓ_2 penalty with parameter $\lambda_2 = 0.1$, and notice that the angle is the same for any sparsity. **Right :** we use a ℓ_1 penalty with parameter $\lambda_1 = 0.1$. The theoretical prediction (full line) is compared with numerical experiments (points) conducted using standard convex optimization solvers from [229].

that, for a non-sparse signal, ℓ_2 and ℓ_1 reconstruction perform similarly. The largest difference is observed at $\rho = 0.1$, where the ℓ_1 penalized logistic regression significantly outperforms the ridge one. We thus keep this value of the sparsity parameter for the next figures.

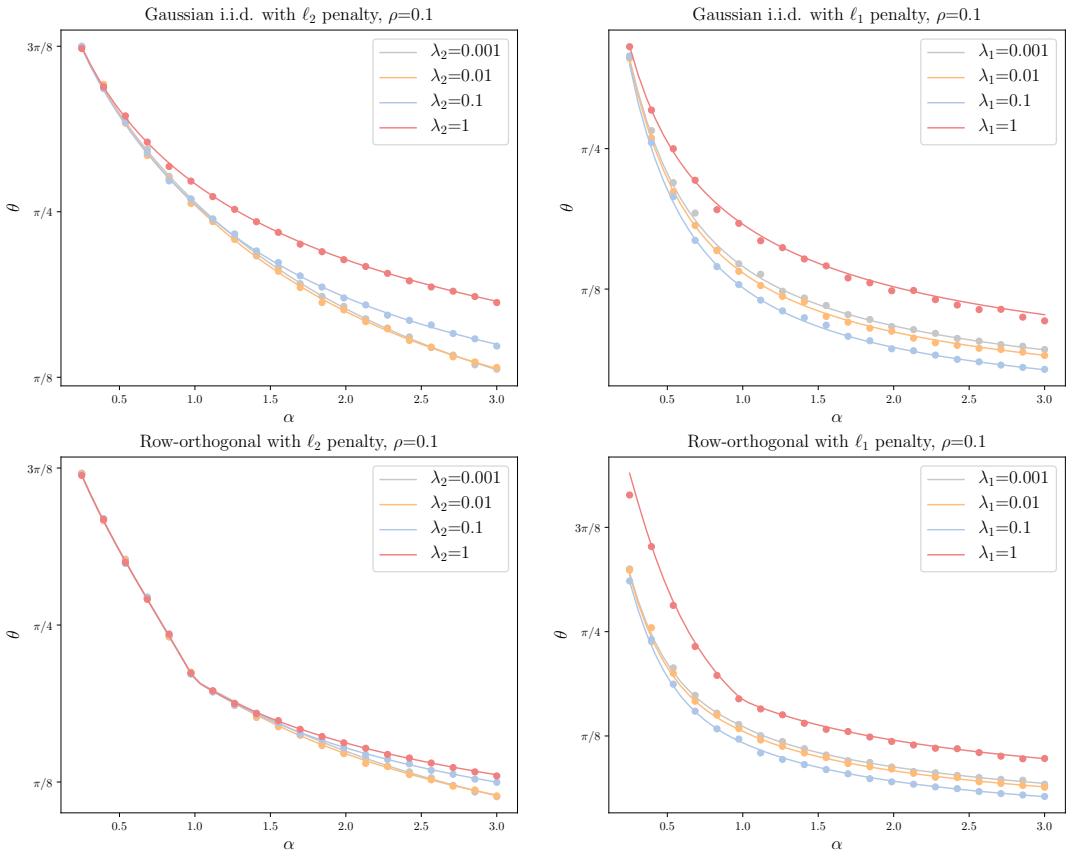


Figure 13.3: Tuning the regularization parameter. We still plot the angle between the estimator and the ground truth in a binary classification problem with $\mathbf{y} = \text{sign}(\mathbf{F}\mathbf{x}_0)$ as a function of $\alpha = M/N$, for a fixed sparsity of planted vector $\rho = 0.1$, for different values of regularization parameters. Figures in the top are for $\mathbf{F}$ Gaussian i.i.d., while figures in the bottom are for $\mathbf{F}$ row-orthogonal. **Left :** ℓ_2 penalty with different values of regularization parameter λ_2. **Right :** ℓ_1 penalty with different values of regularization parameter λ_1.

Varying the regularization parameter at constant sparsity

In Figure 13.3, keeping the sparsity of the teacher constant at $\rho = 0.1$, we look to tune the regularization strength. An interesting effect appears in the ridge-regularized case with row-orthogonal measurements : the curves collapse to a single one when the aspect ratio goes below $\alpha = 1$. We find that the optimal regularization strength for the ℓ_2 penalty lies around $\lambda_2 = 0.01$, and for the ℓ_1-penalty around $\lambda_1 = 0.1$, for both types of matrices.

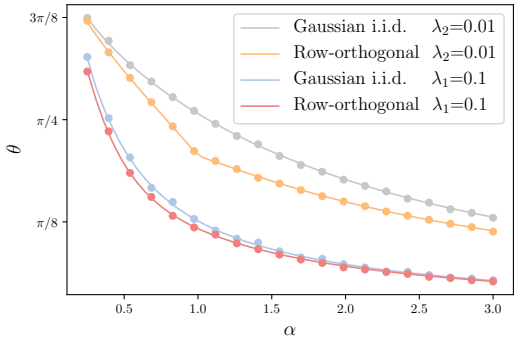


Figure 13.4: Comparing reconstruction performance for Gaussian i.i.d. and row-orthogonal matrices. In this figure, we compare the reconstruction angles between the estimator and the ground-truth for binary classification obtained with ℓ_1 and ℓ_2 penalties. We use logistic regression. The sparsity of the sparse vector is fixed to $\rho = 0.1$. For both Gaussian i.i.d. and row-orthogonal data matrices, we see that ℓ_1 penalty with $\lambda_1 = 0.1$ performs better than the ℓ_2 penalty with $\lambda_2 = 0.01$. For those two penalties, row-orthogonal matrices allow to obtain smaller reconstruction angles than Gaussian i.i.d. matrices.

Comparing case

In Figure 13.4, we directly compare the reconstruction performance of logistic regression on a sparse problem with previously tuned regularization parameter of ℓ_2 and ℓ_1 penalties, with the two types of measurement matrices. We naturally observe that the ℓ_1 penalty leads to better reconstruction of the sparse vector. Row-orthogonal matrices outperform the i.i.d. Gaussian ones with both regularization, although the gap is less significant with the ℓ_1 penalty.

Discussion

Several non-trivial effects are observed when studying the interplay between eigenvalue distribution of the design matrix, loss function, regularization and structure of the underlying teacher vector. Looking for analytical simplifications of the fixed point equations from Theorem 22 in specific cases would be interesting to understand how the key quantities interact and lead, for example, to the collapsing observed in ℓ_2-penalized problems. This further motivates the use of these equations to determine reconstruction limits of generalized-linear modeling. Some examples include limits of sparse recovery for different types of measurement matrices, or finding if optimal losses can be designed to achieve performances close to Bayes optimal errors.

13.5 Sketch of proof of Theorem 22

Our proof follows an approach pioneered in [29] where the LASSO risk for i.i.d. Gaussian matrices is determined. The idea is to build a sequence of iterates that provably converges towards the estimator $\hat{\mathbf{x}}$, while also knowing the statistical properties of those iterates through a set of equations. We must therefore concern ourselves with three fundamental aspects:

(i) construct a sequence of iterates with a rigorous statistical characterization that matches their equations of Theorem 22 at the fixed point,
(ii) verify that the sequence's fixed point corresponds to the estimator $\hat{\mathbf{x}}$,
(iii) check that this sequence is provably convergent, otherwise the iterates might drift off on a diverging trajectory, and the fixed point would never be reached. We thus make sure the statistical characterization indeed applies to the point of interest $\hat{\mathbf{x}}$.

In short, we have a sequence of estimates $(\mathbf{x}_k)_{k\in\mathbb{N}}$ taking values in $\mathbb{R}^N$, and their exact asymptotic (in N) distribution for any $k > 0$. To show that these statistics extend to $\hat{\mathbf{x}}$, we need to show that $\lim_{k\to\infty} \mathbf{x}_k = \hat{\mathbf{x}}$. To do so, we need the sequence to converge (i.e. point iii), and its fixed point to be $\hat{\mathbf{x}}$ (point ii). As indicated in the introduction, we will use an instance of the 2-layer MLVAMP algorithm to construct this sequence. Note that, for the sake of brevity, we do not verify that limiting points of 2-layer MLVAMP trajectories $\lim_{k\to\infty} \mathbf{x}_k$ converge empirically to the Gaussian distribution prescribed by the state evolution equations. This point is treated explicitly in [93].

The following lemma establishes the link between the state evolution equations and our main theorem.

Lemma 52. *(Fixed point of 2-layer MLVAMP state evolution equations) The state evolution equations of 2-layer MLVAMP from [97], reminded in appendix 14.5, match the equations of Theorem 22 at their fixed point.*

Proof. See appendix 14.5. □

This confirms that 2-layer MLVAMP is a good choice to design the sequences that we seek. We know that the iterates of 2-layer MLVAMP can be characterized by state evolution equations which correspond, at their fixed point, to the equations of Theorem 22 by virtue of Lemma 52. The necessary assumptions for the state evolution equations to hold are verified in appendix 14.5.2. We must now show that the estimator of interest defined by (13.1) and (13.2) can be reached using 2-layer MLVAMP. We thus continue with point (ii).

Lemma 53. *(Fixed point of 2-layer MLVAMP) The fixed point of algorithm (1) matches the optimality condition of the unconstrained convex problem Eq.(13.2)*

Proof. See appendix 14.4. □

This part is a consequence of the structure of the algorithm and properties of proximal operators. We now move to point (iii) and seek to characterize the convergence properties of 2-layer MLVAMP. Instead of directly tackling the convergence of 2-layer MLVAMP on any convex GLM, we take a detour and focus on a constrained problem, where functions f and g are augmented by a ℓ_2 norm with ridge parameters λ_2, $\tilde{\lambda}_2$. The called on intuition is that the algorithm will be more likely to converge in a strongly convex problem. We start by showing the convergence of MLVAMP in the constrained strongly convex setting, for values of λ_2 larger than a certain threshold, and any strictly positive $\tilde{\lambda}_2$.

Lemma 54. *(Linear convergence of 2-layer MLVAMP for strongly convex problems) Assume f and g are twice differentiable. Define the constrained problem*

$$\hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \tilde{g}(\mathbf{F}\mathbf{x}, \mathbf{y}) + \tilde{f}(\mathbf{x}) \right\} \tag{13.25}$$

where $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$ *and* $\tilde{g}(\mathbf{x},\mathbf{y}) = g(\mathbf{x},\mathbf{y}) + \frac{\tilde{\lambda}_2}{2}\|\mathbf{x}\|_2^2$. *Consider 2-layer MLVAMP applied to find (13.25), from which we extract at each iteration the vector* $\mathbf{h}^{(t)} = \left[\mathbf{h}_{2z}^{(t)}, \mathbf{h}_{1x}^{(t)}\right]^T$. *Let* $\mathbf{h}^*$ *be its value at the fixed point of algorithm (1). We then have that, for any* $\tilde{\lambda}_2 > 0$, *there exists a value* λ_2^* *such that, for any* $\lambda_2 > \lambda_2^*$, *there exists a strictly positive constant* c *verifying* $0 < c < \lambda_2$, *such that for any* $t \in \mathbb{N}$:

$$\left\|\mathbf{h}^{(t)} - \mathbf{h}^*\right\|_2^2 \leqslant \left(\frac{c}{\lambda_2}\right)^t \left\|\mathbf{h}^{(0)} - \mathbf{h}^*\right\|_2^2, \tag{13.26}$$

The convergence of $\mathbf{h}^{(t)}$ *implies that estimators* $\hat{\mathbf{x}}_1^{(t)}$ *and* $\hat{\mathbf{x}}_2^{(t)}$ *returned by 2-layer MLVAMP also converge to the desired* $\hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2)$, *i.e., under the conditions listed above*

$$\lim_{t\to\infty} \left\|\hat{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2)\right\|_2^2 = 0. \tag{13.27}$$

Proof. See appendix 14.7. □

For a loss function $\tilde{g}$ with any non-zero strong convexity constant, and a regularization $\tilde{f}$ with a sufficiently strong convexity, 2-layer MLVAMP converges linearly towards its unique fixed point. Note that this convergence result is independent from the dimension. We elaborate on this lemma in the next section. An immediate consequence is the following lemma, which claims that Theorem 22 holds when 2-layer MLVAMP converges. Since this result does not rely on an analytic continuation, the assumptions on the concentration of PL2 observables of $\hat{\mathbf{x}}$, given by the state evolution property, and approximation of the cost function by analytic functions with fast decaying higher order derivatives are not required. The result can also be stated for any PL2 observable, with no restriction on its derivability and decay of higher order derivatives. We summarize the necessary assumptions in the following list:

Assumption 3.

(a) the functions f *and* g *are proper, closed, convex and separable functions.*

(b) the cost function $g(\mathbf{F}.,\mathbf{y}) + f(.)$ *is coercive, i.e.* $\lim_{\|\mathbf{x}\|\to\infty} g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) = +\infty$.

(c) there exists a constant B_1 *such that* $\frac{1}{N}\|\hat{\mathbf{x}}\|_2^2 \leqslant B_1$ *almost surely as* $N \to \infty$.

(d) for any $\mathbf{x} \in dom(f)$ *and any* $\mathbf{x}' \in \partial f(\mathbf{x})$, *there exists a constant* C *such that* $\|\mathbf{x}'\|_2 \leqslant C(1 + \|\mathbf{x}\|_2)$. *The same holds for* g *on its domain.*

(e) the empirical distributions of the underlying truth $\mathbf{x}_0$, *eigenvalues of* $\mathbf{F}^T\mathbf{F}$, *and noise vector* w_0, *respectively converge empirically with second order moments, as defined in appendix 14.1, to independent scalar random variables* x_0, w_0, λ *with distributions* p_{x_0}, p_λ, p_{w_0}. *We assume that the distribution* p_λ *is not all-zero and has compact support.*

(f) the design matrix $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{M\times N}$ *is rotationally invariant, as defined in the introduction, where the elements of the Haar distributed matrices* $\mathbf{U}, \mathbf{V}$ *are independent of the elements of the ground truth vector* $\mathbf{x}_0$, *noise* $\boldsymbol{\omega}_0$ *and elements of* $\mathbf{D}$.

(g) the solution to the set of fixed point equations (13.13) exists and is unique for any convex functions f, g *verifying the*

(h) finally assume that $M, N \to \infty$ with fixed ratio $\alpha = M/N$.

Lemma 55. *(Asymptotic error for the twice differentiable, sufficiently strongly convex problem) Consider the strongly convex minimization problem with twice differentiable f and g (13.25). Under the set of assumptions 3, for any $\tilde{\lambda}_2 > 0$, there exists a λ_2^* such that, for any $\lambda_2 > \lambda_2^*$, Then, for any pseudo-Lipschitz function of order 2 ϕ, the following holds :*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \phi(x_{0,i}, \hat{x}_i) \overset{a.s.}{=} \mathbb{E}[\phi(x_0, Prox_{f/\hat{Q}_{1x}^{(t)}}(H_x))] \tag{13.28}$$

$$\lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \phi(z_{0,i}, \hat{z}_i) \overset{a.s.}{=} \mathbb{E}[\phi(z_0, Prox_{f/\hat{Q}_{1z}^{(t)}}(H_z))] \tag{13.29}$$

where the scalars $\hat{Q}_{1x}, \hat{Q}_{1z}$ and the random variables H_x, H_z are defined as in Theorem 22.

Proof. Using the result from Lemma 54, we have $\lim_{t\to\infty} \lim_{N\to\infty} \frac{1}{N} \left\| \mathbf{x}^{(t)} - \hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2) \right\|_2^2 = 0$. As proven in [93], the state evolution parameters will converge to those of the fixed point of the state evolution equations along a converging trajectory of 2-layer MLVAMP. Using the assumption on the bounded averaged norm of $\hat{\mathbf{x}}$, the state evolution equations to show that the averaged norm of the iterates are bounded along a converging trajectory, and the state evolution equations to obtain the exact asymptotics of each iterate along the converging trajectory, an identical argument to that of the proof of Theorem 1.5 from [28] gives Lemma 55. □

We are now left to prove Theorem 22, for any range of parameters $(\lambda_2, \tilde{\lambda}_2)$. $\tilde{\lambda}_2$ can already be chosen arbitrarily small. This means we need to relax the threshold value on λ_2 for the validity of the scalar quantities involved in Theorem 1. To do so, we start by introducing another modification of the original problem, where the objective functions are assumed to be real analytic. Lemma 55 naturally holds for real analytic convex functions. Proving Theorem 22 on the real analytic problem then boils down to performing an analytic continuation on the λ_2 parameter, and is detailed in Appendix 14.8. We thus have the following intermediate result :

Lemma 56. *(Asymptotics of the real analytic problem) Consider assumption 2 is verified. Suppose additionally that f and g are real analytic. Then Theorem 1 holds for any $\tilde{\lambda}_2 > 0$ and any $\lambda_2 > 0$.*

Theorem 22 can then be proven from Lemma 56 by showing that the solutions of the original problem and of its real analytic approximation are arbitrarily close, and by carefully studying the limits $\tilde{\lambda}_2 \to 0$ and $\lambda_2 \to 0$. This is deferred to Appendix 14.8. Note that the proof of the analytic continuation presented here makes the one from [109], which was incomplete, rigorous.
The remaining technical part is the proof of the convergence Lemma 54. For this purpose, we use a dynamical system reformulation of 2-layer MLVAMP and a result from control theory, adapted to machine learning in [166] and more specifically to ADMM in [214].

13.6 Convergence analysis of 2-layer MLVAMP

The key idea of the approach pioneered in [166] is to recast any non-linear dynamical system as a linear one, where convergence will be naturally characterized by a matrix norm. For a given non-linearity $\tilde{\mathcal{O}}$ and iterate $\mathbf{v}$, we define the variable $\mathbf{u} = \tilde{\mathcal{O}}(\mathbf{v})$ and rewrite the initial algorithm

in terms of this trivial transform. Any property of $\tilde{\mathcal{O}}$ is then summarized in a constraint matrix linking $\mathbf{v}$ and $\mathbf{u}$. For example, if $\tilde{\mathcal{O}}$ has Lipschitz constant ω, then for all t:

$$\left\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\right\|_2^2 \leqslant \omega^2\left\|\mathbf{v}^{(t+1)}-\mathbf{v}^{(t)}\right\|_2^2, \tag{13.30}$$

which can be rewritten in matrix form:

$$\mathbf{U}^T\begin{bmatrix}\omega^2\mathbf{I}_{d_v} & 0\\ 0 & -\mathbf{I}_{d_u}\end{bmatrix}\mathbf{U}\succcurlyeq 0 \tag{13.31}$$

$$\text{where } \mathbf{U}=\begin{bmatrix}\mathbf{v}^{(t+1)}-\mathbf{v}^{(t)}\\ \mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\end{bmatrix} \tag{13.32}$$

where $\mathbf{I}_{d_v}, \mathbf{I}_{d_u}$ are the identity matrices with dimensions of $\mathbf{v}, \mathbf{u}$, i.e. M or N in our case. Any co-coercivity property (verified by proximal operators) can be rewritten in matrix form but yields non block diagonal constraint matrices. We will thus directly use the Lipschitz constants for our proof, as they lead to simpler derivations and suffice to prove the required result. The main theorem from [166], adapted to ADMM in [214], then establishes a sufficient condition for convergence with a linear matrix inequality, involving the matrices defining the linear recast of the algorithm and the constraints. Let us now detail how this approach can be used on 2-layer MLVAMP.

13.6.1 2-layer MLVAMP as a dynamical system : sketch of proof of Lemma 3

We start by rewriting 2-layer MLVAMP in a more compact form:

$$\begin{aligned}
&\text{Initialize } \mathbf{h}_{1x}^{(0)}, \mathbf{h}_{2z}^{(0)}\\
&\mathbf{h}_{1x}^{(t+1)} = \mathbf{W}_1^{(t)}\tilde{\mathcal{O}}_1^{(t)}\mathbf{h}_{1x}^{(t)} + \mathbf{W}_2^{(t)}\tilde{\mathcal{O}}_2^{(t)}(\mathbf{W}_3^{(t)}\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4^{(t)}\tilde{\mathcal{O}}_1^{(t)}(\mathbf{h}_{1x}^{(t)})) && (13.33)\\
&\mathbf{h}_{2z}^{(t+1)} = \tilde{\mathcal{O}}_2^{(t)}(\mathbf{W}_3^{(t)}\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4^{(t)}\tilde{\mathcal{O}}_1^{(t)}(\mathbf{h}_{1x}^{(t)})) && (13.34)
\end{aligned}$$

where

$$\mathbf{W_1}^{(t)} = \frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t+1)}}\left(\frac{1}{\chi_{2x}^{(t+1)}}(\hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1} - \mathrm{Id}\right) \tag{13.35}$$

$$\mathbf{W_2}^{(t)} = \frac{\hat{Q}_{2z}^{(t+1)}}{\chi_{2x}^{(t+1)}\hat{Q}_{1x}^{(t+1)}}(\hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}\mathbf{F}^T \tag{13.36}$$

$$\mathbf{W_3}^{(t)} = \frac{\hat{Q}_{2z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\left(\frac{1}{\chi_{2z}^{(t)}}\mathbf{F}(\hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}\mathbf{F}^T - \mathrm{Id}\right) \tag{13.37}$$

$$\mathbf{W_4}^{(t)} = \frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1z}^{(t)}\chi_{2z}^{(t)}}\mathbf{F}(\hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1} \tag{13.38}$$

$$\tilde{\mathbf{O}}_\mathbf{1}^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\left(\frac{1}{\chi_{1x}^{(t)}\hat{Q}_{1x}^{(t)}}\mathrm{Prox}_{\mathbf{f}/\hat{Q}_{1x}^{(t)}}(\cdot) - \mathrm{Id}\right) \tag{13.39}$$

$$\tilde{\mathbf{O}}_\mathbf{2}^{(t)} = \frac{\hat{Q}_{1z}^{(t)}}{\hat{Q}_{2z}^{(t+1)}}\left(\frac{1}{\chi_{1z}^{(t)}\hat{Q}_{1z}^{(t)}}\mathrm{Prox}_{\mathbf{g}(.,y)/\hat{Q}_{1z}^{(t)}}(\cdot) - \mathrm{Id}\right). \tag{13.40}$$

For the linear recast, we then define the variables:

$$\mathbf{u}_1^{(t)} = \tilde{\mathcal{O}}_1^{(t)}(\mathbf{h}_{1x}^{(t)}),\ \mathbf{v}^{(t)} = \mathbf{W}_3^{(t)}\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4^{(t)}\mathbf{u}_1^{(t)}, \tag{13.41}$$

$$\mathbf{u}_2^{(t)} = \tilde{\mathcal{O}}_2^{(t)}(\mathbf{v}^{(t)}), \tag{13.42}$$

$$\text{s.t. } \mathbf{h}_{2z}^{(t+1)} = \mathbf{u}_2^{(t)}, \mathbf{h}_{1x}^{(t+1)} = \mathbf{W}_1^{(t)}\mathbf{u}_1^{(t)} + \mathbf{W}_2^{(t)}\mathbf{u}_2^{(t)}. \tag{13.43}$$

where $\mathbf{u}_1, \mathbf{h}_{1x} \in \mathbb{R}^N$; and $\mathbf{v}, \mathbf{u}_2, \mathbf{h}_{2z} \in \mathbb{R}^M$. We then define as new variables the vectors

$$\mathbf{h}^{(t)} = \begin{bmatrix} \mathbf{h}_{2z}^{(t)} \\ \mathbf{h}_{1x}^{(t)} \end{bmatrix}, \quad \mathbf{u}^{(t)} = \begin{bmatrix} \mathbf{u}_2^{(t)} \\ \mathbf{u}_1^{(t)} \end{bmatrix}, \tag{13.44}$$

$$\mathbf{w}_1^{(t)} = \begin{bmatrix} \mathbf{h}_{1x}^{(t)} \\ \mathbf{u}_1^{(t)} \end{bmatrix}, \quad \mathbf{w}_2^{(t)} = \begin{bmatrix} \mathbf{v}^{(t)} \\ \mathbf{u}_2^{(t)} \end{bmatrix}. \tag{13.45}$$

This leads to the following linear dynamical system recast of (13.33)-(13.34):

$$\mathbf{h}^{(t+1)} = \mathbf{A}^{(t)}\mathbf{h}^{(t)} + \mathbf{B}^{(t)}\mathbf{u}^{(t)} \tag{13.46}$$

$$\mathbf{w}_1^{(t)} = \mathbf{C}_1^{(t)}\mathbf{h}^{(t)} + \mathbf{D}_1^{(t)}\mathbf{u}^{(t)} \tag{13.47}$$

$$\mathbf{w}_2^{(t)} = \mathbf{C}_2^{(t)}\mathbf{h}^{(t)} + \mathbf{D}_2^{(t)}\mathbf{u}^{(t)} \tag{13.48}$$

where

$$\mathbf{A}^{(t)} = \mathbf{0}_{(M+N)\times(M+N)}\ \mathbf{B}^{(t)} = \begin{bmatrix} \mathbf{I}_M & \mathbf{0}_{M\times N} \\ \mathbf{W}_2^{(t)} & \mathbf{W}_1^{(t)} \end{bmatrix} \tag{13.49}$$

$$\mathbf{C}_1^{(t)} = \begin{bmatrix} \mathbf{0}_{N\times M} & \mathbf{I}_N \\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N} \end{bmatrix} \mathbf{D}_1^{(t)} = \begin{bmatrix} \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times M} & \mathbf{I}_N \end{bmatrix} \tag{13.50}$$

$$\mathbf{C}_2^{(t)} = \begin{bmatrix} \mathbf{W}_3^{(t)} & \mathbf{0}_{M\times N} \\ \mathbf{0}_{M\times M} & \mathbf{0}_{M\times N} \end{bmatrix} \mathbf{D}_2^{(t)} = \begin{bmatrix} \mathbf{0}_{M\times M} & \mathbf{W}_4^{(t)} \\ \mathbf{I}_M & \mathbf{0}_{M\times N} \end{bmatrix}. \tag{13.51}$$

$\mathbf{O}$ denotes a matrix with only zeros. The next step is to impose the properties of the non-linearities $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$ through constraint matrices. The Lipschitz constants $\omega_1^{(t)}, \omega_2^{(t)}$ of $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$ can be determined using properties of proximal operators [114] and are directly linked to the strong convexity and smoothness of the cost function and regularization. The relevant properties of proximal operators are reminded in appendix 14.2, while the subsequent derivation of the Lipschitz constants is detailed in appendix 14.7 and yields:

$$\omega_1^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}} \sqrt{1 + \frac{(\hat{Q}_{2x}^{(t)})^2 - (\hat{Q}_{1x}^{(t)})^2}{(\hat{Q}_{1x}^{(t)} + \lambda_2)^2}} \tag{13.52}$$

$$\omega_2^{(t)} = \frac{\hat{Q}_{1z}^{(t)}}{\hat{Q}_{2z}^{(t)}} \sqrt{1 + \frac{(\hat{Q}_{2z}^{(t)})^2 - (\hat{Q}_{1z}^{(t)})^2}{(\hat{Q}_{1z}^{(t)} + \tilde{\lambda}_2)^2}}. \tag{13.53}$$

We thus define the constraints matrices

$$\mathbf{M}_1^{(t)} = \begin{bmatrix} (\omega_1^{(t)})^2 & 0 \\ 0 & -1 \end{bmatrix} \otimes \mathbf{I}_N \quad \mathbf{M}_2^{(t)} = \begin{bmatrix} (\omega_2^{(t)})^2 & 0 \\ 0 & -1 \end{bmatrix} \otimes \mathbf{I}_M \tag{13.54}$$

where $\otimes$ denotes the Kronecker product. We then use a time dependent form of Theorem 4 from [166] in the appropriate form for 2-layer MLVAMP, as was done in [214] for ADMM.

Proposition 9. *(Time dependent version of Theorem 4 from [166]) Consider, at each time step $t \in \mathbb{N}$, the following linear matrix inequality with $\tau_{(t)} \in [0,1]$:*

$$0 \succeq \begin{bmatrix} (\mathbf{A}^{(t)})^T \mathbf{P} \mathbf{A}^{(t)} - (\tau_{(t)})^2 \mathbf{P} & (\mathbf{A}^{(t)})^T \mathbf{P} \mathbf{B}^{(t)} \\ (\mathbf{B}^{(t)})^T \mathbf{P} \mathbf{A}^{(t)} & (\mathbf{B}^{(t)})^T \mathbf{P} \mathbf{B}^{(t)} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1^{(t)} & \mathbf{D}_1^{(t)} \\ \mathbf{C}_2^{(t)} & \mathbf{D}_2^{(t)} \end{bmatrix}^T \begin{bmatrix} \beta_1^{(t)} \mathbf{M}_1^{(t)} & \mathbf{0}_{2N \times 2M} \\ \mathbf{0}_{2M \times 2N} & \beta_2^{(t)} \mathbf{M}_2^{(t)} \end{bmatrix} \begin{bmatrix} \mathbf{C}_1^{(t)} & \mathbf{D}_1^{(t)} \\ \mathbf{C}_2^{(t)} & \mathbf{D}_2^{(t)} \end{bmatrix} \tag{13.55}$$

If, at each time step, (13.55) is feasible for some $\mathbf{P} \succ 0$ and $\beta_1^{(t)}, \beta_2^{(t)} \geqslant 0$, then for any initialization $\mathbf{h}^{(0)}$, $\mathbf{h}^{(t)}$ converges to $\mathbf{h}^*$, the fixed point of (13.46)-(13.48):

$$\forall t, \quad \left\| \mathbf{h}^{(t)} - \mathbf{h}^* \right\| \leqslant \sqrt{\kappa(\mathbf{P})} (\tau^*)^t \left\| \mathbf{h}^{(0)} - \mathbf{h}^* \right\| \tag{13.56}$$

where $\kappa(\mathbf{P})$ is the condition number of $\mathbf{P}$ and we defined $\tau^* = \sup_t \tau_{(t)}$.

Proof. see appendix 14.7.1 □

We show in appendix 14.7 how the additional ridge penalties from the constrained problem (13.25) parametrized by $\lambda_2, \tilde{\lambda}_2$ can be used to make (13.55) feasible and prove Lemma 54. The core idea is to leverage on the Lipschitz constants (13.52), the operator norms of the matrices defined in (13.35) and the following upper and lower bounds on the $\hat{Q}$ parameters defined by the fixed point of state evolution equations:

$$\lambda_{min}(\mathcal{H}_f) \leqslant \hat{Q}_{2x}^{(t)} \leqslant \lambda_{max}(\mathcal{H}_f) \tag{13.57}$$

$$\lambda_{min}(\mathcal{H}_g) \leqslant \hat{Q}_{2z}^{(t+1)} \leqslant \lambda_{max}(\mathcal{H}_g) \tag{13.58}$$

$$\hat{Q}_{2z}^{(t)} \lambda_{min}(\mathbf{F}^T \mathbf{F}) \leqslant \hat{Q}_{1x}^{(t+1)} \leqslant \hat{Q}_{2z}^{(t)} \lambda_{max}(\mathbf{F}^T \mathbf{F}) \tag{13.59}$$

$$\frac{\hat{Q}_{2x}^{(t)}}{\lambda_{max}(\mathbf{F}\mathbf{F}^T)} \leqslant \hat{Q}_{1z}^{(t)} \leqslant \frac{\hat{Q}_{2x}^{(t)}}{\lambda_{min}(\mathbf{F}\mathbf{F}^T)}, \tag{13.60}$$

where $\mathcal{H}_f, \mathcal{H}_g$ are the Hessian of the loss and regularization functions taken at the fixed point. These bounds are obtained from the definitions of χ_x, χ_z in the state evolution equations (or equivalently in Theorem 22), and the fact that the derivative of a proximal operator reads, for a twice differentiable function:

$$\mathcal{D}\eta_{\gamma f}(\mathbf{x}) = (\mathrm{Id} + \gamma \mathcal{H}_f(\eta_{\gamma f}(\mathbf{x})))^{-1}. \tag{13.61}$$

Detail of this derivation can also be found in appendices 14.2 and 14.7. For the constrained problem (13.25), the maximum and minimum eigenvalues of the Hessians are directly augmented by $\tilde{\lambda}_2, \lambda_2$, which allows us to control the scaling of the $\hat{Q}$ parameters. The rest of the convergence proof is then based on successive application of Schur's lemma [127] on the linear matrix inequality (13.55); and translating the resulting conditions on inequalities which can be verified by choosing the appropriate $\tilde{\lambda}_2, \lambda_2, \beta_1^{(t)}, \beta_2^{(t)}$. Convergence of gradient-based descent methods for sufficiently strongly-convex objectives is a coherent result from an optimization point of view. This is corroborated by the symbolic convergence rates derived for ADMM in [214], where a sufficiently strongly convex objective is also considered.

13.6.2 Numerical experiments for Lemma 54

Here we provide numerical evidence for the linear convergence condition proved in Lemma 3. We consider a logistic regression penalized with the ℓ_1 norm ($\lambda_1 = 0.1$) with an ill-conditioned design matrix, with i.i.d. standard normal elements. This corresponds to the setting of Figure 13.3. Since the logistic loss is strongly convex on any compact space, we do not need to add $\tilde{\lambda}_2$. We follow the convergence of 2-layer MLVAMP for this problem for increasing values of an additional ridge penalty $\lambda_2 = 0, 0.01, 0.05, 0.1$ and plot the average distance between successive iterates $\frac{1}{N}\left\|\mathbf{h}_{1x}^{(t+1)} - \mathbf{h}_{1x}^{(t+1)}\right\|_2^2$ and the evolution of the reconstruction angle θ as a function of the number of iterations. We perform two experiments with aspect ratios $\alpha = 1$ and $\alpha = 0.2$. For $\alpha = 1$, 2-layer MLVAMP converges without any additional ridge penalty, and convergence is accelerated by larger values of λ_2. As a sanity check, note that the reconstruction angle of the estimator returned by the algorithm for $\lambda_2 = 0$ (grey line on the lower left plot) converges to the value predicted at Figure 13.3 for $\alpha = 1, \lambda_1 = 0.1$ and a Gaussian matrix. For $\alpha = 0.2$n the design matrix is ill-conditioned and we see that 2-layer MLVAMP diverges. Adding the ridge penalty leads to converging trajectories for a sufficiently large value of λ_2, as shown on the upper right block. Larger values of λ_2 again lead to faster convergence.

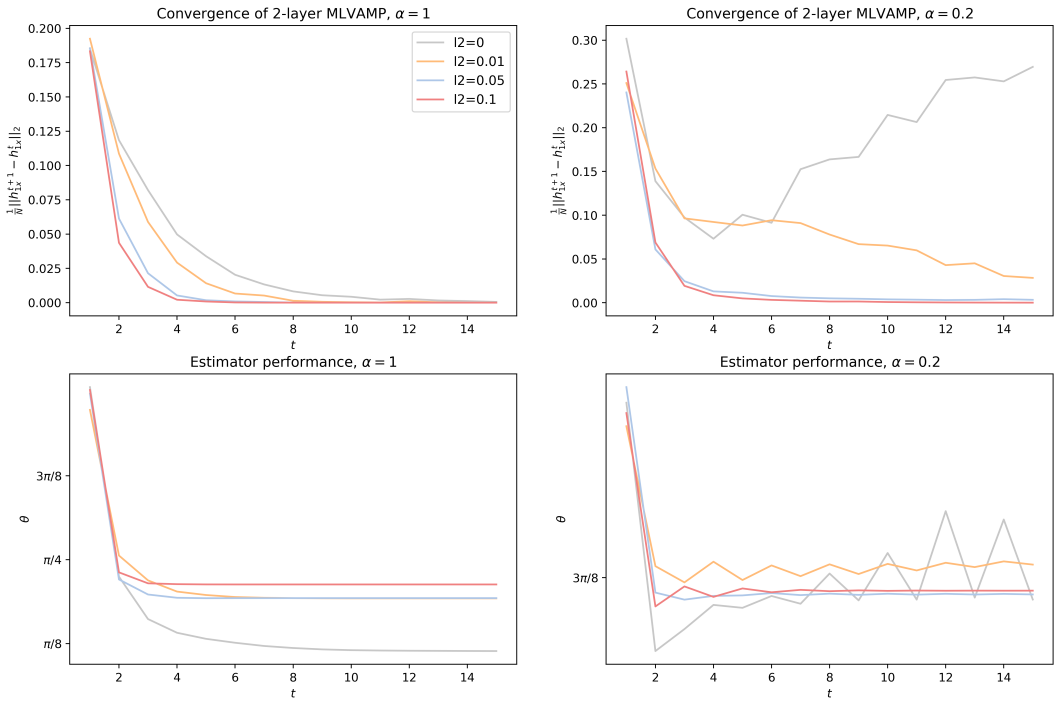


Figure 13.5: Convergence of 2-layer MLVAMP on a logistic regression with ℓ_1 penalty with $\lambda_1 = 0.1$, a Gaussian design matrix and two values of the aspect ratio $\alpha = 1$ (left) and $\alpha = 0.2$ (right). For $\alpha = 1$, the algorithm converges regardless of the additional ridge penalty and we recover the performance predicted by Theorem 22 for the plain ℓ_1 regularization. For $\alpha = 0.2$, the plain ℓ_1 leads to an unstable iteration and a sufficiently large additional ridge indeed leads to convergence. In both cases, the larger the additional ridge, the faster the algorithm converges.

Chapter 14

Proofs for the Kabashima formula

14.1 Convergence of vector sequences

This section is a brief summary of the framework originally introduced in [28] and used in [97, 242]. We review the key definitions and verify that they apply in our setting. We remind the full set of state evolution equations from [97] at (14.57), when applied to learning a GLM, in appendix 14.5, along with the required assumptions for them to hold in appendix 14.5.2.
The main building blocks are the notions of *vector sequence* and *pseudo-Lipschitz function*, which allow to define the *empirical convergence with p-th order moment*. Consider a vector of the form

$$\mathbf{x}(N) = (\mathbf{x}_1(N), ..., \mathbf{x}_N(N)) \tag{14.1}$$

where each sub-vector $\mathbf{x}_n(N) \in \mathbb{R}^r$ for any given $r \in \mathbb{N}^*$. For r=1, which we use in Theorem 22, $\mathbf{x}(N)$ is denoted a *vector sequence*.
Given $p \geqslant 1$, a function $\mathbf{f} : \mathbb{R}^r \to \mathbb{R}^s$ is said to be *pseudo-Lipschitz continuous of order p* if there exists a constant $C > 0$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^s$:

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\| \leqslant C\|\mathbf{x}_1 - \mathbf{x}_2\| \left[1 + \|\mathbf{x}_1\|^{p-1} + \|\mathbf{x}_2\|^{p-1}\right] \tag{14.2}$$

Then, a given vector sequence $\mathbf{x}(N)$ *converges empirically with p-th order moment* if there exists a random variable $X \in \mathbb{R}^r$ such that:

- $\mathbb{E}\|X\|_p^p < \infty$; and
- for any scalar-valued pseudo-Lipschitz continuous $\mathbf{f} : \mathbb{R}^r \to \mathbb{R}$ of order p,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{f}(x_n(N)) = \mathbb{E}[f(X)] \tag{14.3}$$

Note that defining an empirically converging singular value distribution implicitly defines a sequence of matrices $\mathbf{F}(N)$ using the definition of rotational invariance from the introduction. This naturally brings us back to the original definitions from [28]. An important point is that the almost sure convergence of the second condition holds for random vector sequences, such as the ones we consider in the introduction. Note that the noise vector ω_0 must also satisfy these conditions, and naturally does when it is an i.i.d. Gaussian one. We also remind the definition of *uniform Lipschitz continuity*.

For a given mapping $\phi(\mathbf{x}, A)$ defined on $\mathbf{x} \in \mathcal{X}$ and $A \in \mathbb{R}$, we say it is *uniformly Lipschitz continuous* in $\mathbf{x}$ at $A = \bar{A}$ if there exists constants L_1 and $L_2 \geqslant 0$ and an open neighborhood U of $\bar{A}$ such that:

$$\|\phi(\mathbf{x}_1, A) - \phi(\mathbf{x}_2, A)\| \leqslant L_1 \|\mathbf{x}_1 - \mathbf{x}_2\| \tag{14.4}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $A \in U$; and

$$\|\phi(\mathbf{x}, A_1) - \phi(\mathbf{x}, A_2)\| \leqslant L_2(1 + \|\mathbf{x}\|)|A_1 - A_2| \tag{14.5}$$

for all $\mathbf{x} \in \mathcal{X}$ and $A_1, A_2 \in U$.

We discuss the required assumptions for the state evolution equations to hold in detail, and why they are verified in our setting, in appendix 14.5.2.

14.2 Convex analysis and properties of proximal operators

We start this section with a few useful definitions from convex analysis, which can all be found in textbooks such as [25]. We then remind important properties of proximal operators, which we use in appendix 14.7 to derive upper bounds on the Lipschitz constants of the non-linear operators $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$. In what follows, we denote $\mathcal{X}$ the Hilbert space with scalar inner product serving as input and output space, here $\mathbb{R}^N$ or $\mathbb{R}^M$. For simplicity, we will write all operators as going from $\mathcal{X}$ to $\mathcal{X}$.

Definition 16. *(Strong convexity) A proper closed function is σ-strongly convex with $\sigma > 0$ if $f - \frac{\sigma}{2}\|.\|^2$ is convex. If f is differentiable, the definition is equivalent to*

$$f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\sigma}{2}\|x - y\|^2 \tag{14.6}$$

for all $x, y \in \mathcal{X}$.

Definition 17. *(Smoothness for convex functions) A proper closed function f is β-smooth with $\beta > 0$ if $\frac{\beta}{2}\|.\|^2 - f$ is convex. If f is differentiable, the definition is equivalent to*

$$f(x) \leqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\beta}{2}\|x - y\|^2 \tag{14.7}$$

for all $x, y \in \mathcal{X}$.

An immediate consequence of those definitions is the following second order condition: for twice differentiable functions, f is σ-strongly convex and β-smooth if and only if:

$$\sigma \mathrm{Id} \preceq \mathcal{H}_{\mathrm{f}} \preceq \beta \mathrm{Id}. \tag{14.8}$$

Definition 18. *(Co-coercivity) Let $T : \mathcal{X} \to \mathcal{X}$ and $\beta \in \mathbb{R}_+^*$. Then T is β co-coercive if βT is firmly-nonexpansive, i.e.*

$$\langle \mathbf{x} - \mathbf{y}, T(\mathbf{x}) - T(\mathbf{y}) \rangle \geqslant \beta \|T(\mathbf{x}) - T(\mathbf{y})\|_2^2 \tag{14.9}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

Proximal operators are 1 co-coercive or equivalently firmly-nonexpansive.

Corollary 8. *(Remark 4.24 [25]) A mapping $T : \mathcal{X} \to \mathcal{X}$ is β-cocoercive if and only if βT is half-averaged. This means that T can be expressed as:*

$$T = \frac{1}{2\beta}(\mathrm{Id} + \mathrm{S}) \tag{14.10}$$

where S is a nonexpansive operator.

Proposition 10. *(Resolvent of the sub-differential [25]) The proximal mapping of a convex function f is the resolvent of the sub-differential ∂f of f:*

$$Prox_{\gamma f} = (\mathrm{Id} + \gamma \partial \mathrm{f})^{-1}. \tag{14.11}$$

The following proposition is due to [114], and is useful to determine upper bounds on the Lipschitz constant of update functions involving proximal operators.

Proposition 11. *(Proposition 2 from [114]) Assume that f is σ-strongly convex and β-smooth and that $\gamma \in]0, \infty[$. Then $Prox_{\gamma f} - \frac{1}{1+\gamma\beta}\mathrm{Id}$ is $\frac{1}{\frac{1}{1+\gamma\beta} - \frac{1}{1+\gamma\sigma}}$-cocoercive if $\beta > \sigma$ and 0-Lipschitz if $\beta = \sigma$. If f has no smoothness constant, the same holds by taking $\beta = +\infty$.*

We will use these definitions and properties to derive the Lipschitz constants of $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$ in appendix 14.7.

Lemma 57. *Jacobian of the proximal*
Using proposition 10, the proximal operator can be written, for any parameter $\gamma \in \mathbb{R}^+$ and $\mathbf{x}$ in the input space $\mathcal{X}$:

$$Prox_{\gamma f}(\mathbf{x}) = (\mathrm{Id} + \gamma \partial \mathrm{f})^{-1}(\mathbf{x}). \tag{14.12}$$

For any convex and differentiable function f, we have:

$$Prox_{\gamma f}(\mathbf{x}) + \gamma \nabla f(Prox_{\gamma f}(\mathbf{x})) = \mathbf{x} \tag{14.13}$$

For a twice differentiable f, applying the chain rule then yields:

$$\mathcal{D}_{Prox_{\gamma f}}(\mathbf{x}) + \gamma \mathcal{H}_f(Prox_{\gamma f}(\mathbf{x}))\mathcal{D}_{Prox_{\gamma f}}(\mathbf{x}) = \mathrm{Id} \tag{14.14}$$

where $\mathcal{D}$ is the Jacobian matrix and $\mathcal{H}$ the Hessian. Since f is a convex function, its Hessian is positive semi-definite, and, knowing that γ is strictly positive, the matrix $(\mathrm{Id} + \gamma \mathcal{H}_{\mathrm{f}}(Prox_{\gamma f}))$ is invertible. We thus have:

$$\mathcal{D}_{Prox_{\gamma f}}(\mathbf{x}) = (\mathrm{Id} + \gamma \mathcal{H}_{\mathrm{f}}(Prox_{\gamma f}(\mathbf{x})))^{-1} \tag{14.15}$$

Lemma 58. *Proximal of ridge regularized functions*
Since we consider only separable functions, we can work with scalar version of the proximal operators. The scalar proximal of a given function with an added ridge regularization can be written:

$$Prox_{\gamma(f + \frac{\lambda_2}{2}\|.\|_2^2)}(x) = (\mathrm{Id} + \gamma(\partial \mathrm{f} + \lambda_2))^{-1}(\mathrm{x}) \tag{14.16}$$

$$= ((1 + \gamma\lambda_2)Id + \gamma f')^{-1}(x) \tag{14.17}$$

where the second equality is true only for differentiable f. If f is real analytic, we can apply the analytic inverse function theorem [148] and verify analyticity in λ_2 of the proximal.

Finally, we remind a result from [25] describing the limiting behavior of regularized estimators for vanishing regularization.

Proposition 12. *(Theorem 26.20 from [25]) Let f and h be proper, lower semi-continuous, convex functions defined on $\mathcal{X}$. Suppose that $\arg\min f \cap dom(h) \neq \emptyset$ and that h is coercive and strictly convex. Then h admits a unique minimizer $\mathbf{x}_0$ over $\arg\min f$ and , for every $\epsilon \in]0,1[$, the regularized problem*

$$\arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \epsilon h(\mathbf{x}) \tag{14.18}$$

admits a unique solution $\mathbf{x}_\epsilon$. If we assume further that h is uniformly convex on any closed ball of the input space, then $\lim_{\epsilon\to 0} \mathbf{x}_\epsilon = \mathbf{x}_0$.

14.3 From replica potentials to Moreau envelopes

Here we show how the potentials defined for the replica free energy of corollary 6 can be mapped to Moreau envelopes in the zero temperature limit, i.e. $\beta \to \infty$ where β is the inverse temperature. We consider the scalar case since the replica expressions are scalar. All functions are separable here, so any needed generalization to the multidimensional case is immediate. We start by reminding the definition of the Moreau envelope [25, 224] $\mathcal{M}_{\gamma f}$ of a proper, closed and convex function f for a given $\gamma \in \mathbb{R}_+^*$ and any $z \in \mathbb{R}$:

$$\mathcal{M}_{\gamma f}(z) = \inf_{x\in\mathbb{R}} \left\{ f(x) + (1/2\gamma)\|x - z\|_2^2 \right\} \tag{14.19}$$

The Moreau envelope can be interpreted as a smoothed version of a given objective function with the same minimizer. For ℓ_1 minimization for example, it allows to work with a differentiable objective. By definition of the proximal operator we have the following identity:

$$\text{Prox}_{\gamma f}(z) = \arg\min_{x\in\mathbb{R}} \left\{ f(x) + (1/2\gamma)\|x - z\|_2^2 \right\} \tag{14.20}$$

$$\mathcal{M}_{\gamma f}(z) = f(\text{Prox}_{\gamma f}(z)) + \frac{1}{2}\|\text{Prox}_{\gamma f}(z) - z\|_2^2 \tag{14.21}$$

We can now match the replica potentials with the Moreau envelope. We start from the definition of said potentials, to which we apply Laplace's approximation:

$$\begin{aligned}
&\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) = \lim_{\beta\to\infty} ... \\
&\frac{1}{\beta}\log \int e^{-\frac{\beta\hat{Q}_{1x}}{2}x^2 + \beta(\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - \beta f(x)} dx \qquad (14.22)\\
&= -\frac{\hat{Q}_{1x}}{2}(x^*)^2 + (\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x^* - f(x^*) \qquad (14.23)
\end{aligned}$$

where

$$\begin{aligned}
x^* = \arg\min_x \Big\{ &-\frac{\hat{Q}_{1x}}{2}x^2 + ... \\
&(\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - f(x) \Big\}
\end{aligned} \tag{14.24}$$

This is an unconstraint convex optimization problem, thus its optimality condition is enough to characterize its set of minimizers:

$$-\hat{Q}_{1x}x^* + (\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}) - \partial f(x^*) = 0 \tag{14.25}$$

$$\iff x^* = (Id + \frac{1}{\hat{Q}_{1x}}\partial f)^{-1}\left(\frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}\right) \tag{14.26}$$

$$\iff x^* = \text{Prox}_{\frac{f}{\hat{Q}_{1x}}}\left(\frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}\right) \tag{14.27}$$

Replacing this in the replica potential and completing the square, we get:

$$\begin{aligned}\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) &= -f(\text{Prox}_{\gamma f}(X))... \\ &\quad - \frac{\hat{Q}_{1x}}{2}\|X - \text{Prox}_{\gamma f}(X)\|_2^2 + \frac{X^2}{2}\hat{Q}_{1x} \qquad (14.28)\\ &= \hat{Q}_{1x}\frac{X^2}{2} - \mathcal{M}_{\frac{1}{\hat{Q}_{1x}}f}(X) \qquad (14.29)\end{aligned}$$

where we used the shorthand $X = \frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}$.

14.4 Fixed point of multilayer vector approximate message passing

Here we show that the fixed point of 2-layer MLVAMP coincides with the optimality condition of the convex problem 13.2, proving Lemma 53. Writing the fixed point of the scalar parameters of algorithm (1), we get the following prescriptions on the scalar quantities:

$$\frac{1}{\chi_x} \equiv \frac{1}{\chi_{1x}} = \frac{1}{\chi_{2x}} = \hat{Q}_{1x} + \hat{Q}_{2x} \tag{14.30}$$

$$\frac{1}{\chi_z} \equiv \frac{1}{\chi_{1z}} = \frac{1}{\chi_{2z}} = \hat{Q}_{1z} + \hat{Q}_{2z} \tag{14.31}$$

$$\hat{Q}_{1x}\chi_{1x} + \hat{Q}_{2x}\chi_{2x} = 1 \tag{14.32}$$

$$\hat{Q}_{1z}\chi_{1z} + \hat{Q}_{2z}\chi_{2z} = 1 \tag{14.33}$$

and the following ones on the estimates, as proved in [222] section III:

$$\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2 \qquad \hat{\mathbf{z}}_1 = \hat{\mathbf{z}}_2 \tag{14.34}$$

$$\hat{\mathbf{z}}_1 = \mathbf{F}\hat{\mathbf{x}}_1 \qquad \hat{\mathbf{z}}_2 = \mathbf{F}\hat{\mathbf{x}}_2 \tag{14.35}$$

We would like the fixed point of MLVAMP to satisfy the following first-order optimality condition

$$\partial f(\hat{\mathbf{x}}) + \mathbf{F}^T\partial g(\mathbf{F}\hat{\mathbf{x}}) = 0, \tag{14.36}$$

which characterizes the unique minimizer of the unconstraint convex problem (13.2). Replacing $\mathbf{h}_{1x}$'s expression inside $\mathbf{h}_{2x}$ reads

$$\begin{aligned}\mathbf{h}_{2x} &= \left(\frac{\hat{\mathbf{x}}_1}{\chi_x} - \hat{Q}_{1x}\mathbf{h}_{1x}\right)/\hat{Q}_{2x} \qquad (14.37)\\ &= \left(\frac{\hat{\mathbf{x}}_1}{\chi_x} - \left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - \hat{Q}_{2x}\mathbf{h}_{2x}\right)\right)/\hat{Q}_{2x} \qquad (14.38)\end{aligned}$$

and using (14.31) we get $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2$, and a similar reasoning gives $\hat{\mathbf{z}}_2 = \hat{\mathbf{z}}_1$. From (13.8) and (13.9), we clearly find $\hat{\mathbf{z}}_2 = \mathbf{F}\hat{\mathbf{x}}_\mathbf{2}$. Inverting the proximal operators in (13.5) and (13.7) yields

$$\hat{\mathbf{x}}_1 + \frac{1}{\hat{Q}_{1x}} \partial g(\hat{\mathbf{x}}_1) = \mathbf{h}_{1x} \tag{14.39}$$

$$\hat{\mathbf{z}}_1 + \frac{1}{\hat{Q}_{1z}} \partial g(\hat{\mathbf{z}}_1) = \mathbf{h}_{1z}. \tag{14.40}$$

Starting from the MLVAMP equation on $\mathbf{h}_{1x}$, we write

$$\mathbf{h}_{1x} = \left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - \hat{Q}_{2x} \mathbf{h}_{2x} \right) / \hat{Q}_{1x} \tag{14.41}$$

$$= \frac{\left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - (\hat{Q}_{2z} \mathbf{F}^T \mathbf{F} + \hat{Q}_{2x} \mathrm{Id}) \hat{\mathbf{x}}_2 + \hat{Q}_{2z} \mathbf{F}^T \mathbf{h}_{2z} \right)}{\hat{Q}_{1x}} \tag{14.42}$$

$$= - \frac{\left(\hat{Q}_{2z} \mathbf{F}^T \mathbf{F} + \hat{Q}_{2x} \left(1 - \frac{1}{\chi_x \hat{Q}_{2x}} \right) \mathrm{Id} \right) \hat{\mathbf{x}}_2}{\hat{Q}_{2x}} \tag{14.43}$$

$$+ \mathbf{F}^T \left(\hat{Q}_{1z} \left(\frac{1}{\chi_z \hat{Q}_{1z}} - 1 \right) \hat{\mathbf{z}}_1 - \partial \mathbf{g}(\hat{\mathbf{z}}_1) \right) \tag{14.44}$$

which is equal to the left-hand term in (14.39). Using this equality, as well as $\hat{\mathbf{z}}_1 = \mathbf{F}\hat{\mathbf{x}}_\mathbf{2}$ and relations (14.31) and (14.33) yields

$$\partial f(\hat{\mathbf{x}}_2) + \mathbf{F}^T \partial g(\mathbf{F}\hat{\mathbf{x}}_\mathbf{2}) = 0. \tag{14.45}$$

Hence, the fixed point of MLVAMP satisfies the optimality condition (14.36) and is indeed the desired estimator: $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2 = \hat{\mathbf{x}}$.

14.5 State evolution equations

This appendix is intended mainly for completeness, to show that the fixed point equations from Theorem 22, stemming from the heuristic state evolution written in [277] are indeed made rigorous by the results presented in [97].

14.5.1 Heuristic state evolution equations

The state evolution equations track the evolution of MLVAMP (1) and provide statistical properties of its iterates. They are derived in [277] taking the heuristic assumption that $\mathbf{h_{1x}}, \mathbf{h_{1z}}, \mathbf{h_{2x}}, \mathbf{h_{2z}}$ behave as Gaussian estimates, which comes from the physics cavity approach:

$$\hat{Q}_{1x}^{(t)} \mathbf{h}_{1x}^{(t)} - \hat{m}_{1x}^{(t)} \mathbf{x_0} \overset{PL2}{=} \sqrt{\hat{\chi}_{1x}^{(t)}} \boldsymbol{\xi}_{1x}^{(t)} \tag{14.46a}$$

$$\mathbf{V}^T (\hat{Q}_{2x}^{(t)} \mathbf{h}_{2x}^{(t)} - \hat{m}_{2x}^{(t)} \mathbf{x_0}) \overset{PL2}{=} \sqrt{\hat{\chi}_{2x}^{(t)}} \boldsymbol{\xi}_{2x}^{(t)} \tag{14.46b}$$

$$\mathbf{U}^T (\hat{Q}_{1z}^{(t)} \mathbf{h}_{1z}^{(t)} - \hat{m}_{1z}^{(t)} \mathbf{z_0}) \overset{PL2}{=} \sqrt{\hat{\chi}_{1z}^{(t)}} \boldsymbol{\xi}_{1z}^{(t)} \tag{14.46c}$$

$$\hat{Q}_{2z}^{(t)} \mathbf{h}_{2z}^{(t)} - \hat{m}_{2z}^{(t)} \mathbf{z_0} \overset{PL2}{=} \sqrt{\hat{\chi}_{2z}^{(t)}} \boldsymbol{\xi}_{2z}^{(t)} \tag{14.46d}$$

where $\overset{PL2}{=}$ denotes $PL2$ convergence. $\mathbf{U}$ and $\mathbf{V}$ come from the singular value decomposition $\mathbf{F} = \mathbf{UDV}^T$ and are Haar-sampled; $\xi_{1x}^{(t)}, \xi_{2x}^{(t)}, \xi_{1z}^{(t)}, \xi_{2z}^{(t)}$ are normal Gaussian vectors, independent from $\mathbf{x_0}, \mathbf{z_0}, \mathbf{V}^T\mathbf{x_0}$ and $\mathbf{U}^T\mathbf{z_0}$. Parameters $\hat{Q}_{1x}^{(t)}, \hat{Q}_{1z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)}$ are defined through MLVAMP's iterations (1); while parameters $\hat{m}_{1x}^{(t)}, \hat{m}_{1z}^{(t)}, \hat{m}_{2x}^{(t)}, \hat{m}_{2z}^{(t)}$ and $\hat{\chi}_{1x}^{(t)}, \hat{\chi}_{1z}^{(t)}, \hat{\chi}_{2x}^{(t)}, \hat{\chi}_{2z}^{(t)}$ are prescribed through SE equations. Other useful variables are the overlaps and squared norms of estimators, for $k \in \{1, 2\}$:

$$m_{kx}^{(t)} = \frac{\mathbf{x}_0^\top \hat{\mathbf{x}}_k^{(t)}}{N} \qquad q_{kx}^{(t)} = \frac{\|\hat{\mathbf{x}}_k^{(t)}\|_2^2}{N}$$
$$m_{kz}^{(t)} = \frac{\mathbf{z}_0^\top \hat{\mathbf{z}}_k^{(t)}}{M} \qquad q_{kz}^{(t)} = \frac{\|\hat{\mathbf{z}}_k^{(t)}\|_2^2}{M}.$$

Starting from assumptions (14.46), and following the derivation of [277] adapted to the iteration order from (1), the heuristic state evolution equations read:

$$\text{Initialize} \quad \hat{Q}_{1x}^{(0)}, \hat{Q}_{2z}^{(0)}, \hat{m}_{1x}^{(0)}, \hat{m}_{2z}^{(0)}, \hat{\chi}_{1x}^{(0)}, \hat{\chi}_{2z}^{(0)} > 0.$$

$$m_{1x}^{(t)} = \mathbb{E}\left[x_0 \eta_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)} x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{14.47a}$$

$$\chi_{1x}^{(t)} = \frac{1}{\hat{Q}_{1x}^{(t)}}\mathbb{E}\left[\eta'_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)} x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{14.47b}$$

$$q_{1x}^{(t)} = \mathbb{E}\left[\eta^2_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)} x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{14.47c}$$

$$\hat{Q}_{2x}^{(t)} = \frac{1}{\chi_{1x}^{(t)}} - \hat{Q}_{1x}^{(t)} \tag{14.47d}$$

$$\hat{m}_{2x}^{(t)} = \frac{m_{1x}^{(t)}}{\rho_x \chi_{1x}^{(t)}} - \hat{m}_{1x}^{(t)} \tag{14.47e}$$

$$\hat{\chi}_{2x}^{(t)} = \frac{q_{1x}^{(t)}}{(\chi_{1x}^{(t)})^2} - \frac{(m_{1x}^{(t)})^2}{\rho_x (\chi_{1x}^{(t)})^2} - \hat{\chi}_{1x}^{(t)} \tag{14.47f}$$

$$m_{2z}^{(t)} = \frac{\rho_x}{\alpha}\mathbb{E}\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t)})}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)}}\right] \tag{14.47g}$$

$$\chi_{2z}^{(t)} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)}}\right] \tag{14.47h}$$

$$q_{2z}^{(t)} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda(\hat{\chi}_{2x}^{(t)} + \lambda \hat{\chi}_{2z}^{(t)})}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)})^2}\right] \tag{14.47i}$$

$$+ \frac{\rho_x}{\alpha}\mathbb{E}\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)})^2}\right] \tag{14.47j}$$

$$\hat{Q}_{1z}^{(t)} = \frac{1}{\chi_{2z}^{(t)}} - \hat{Q}_{2z}^{(t)} \tag{14.47k}$$

$$\hat{m}_{1z}^{(t)} = \frac{m_{2z}^{(t)}}{\rho_z \chi_{2z}^{(t)}} - \hat{m}_{2z}^{(t)} \tag{14.47l}$$

$$\hat{\chi}_{1z}^{(t)} = \frac{q_{2z}^{(t)}}{(\chi_{2z}^{(t)})^2} - \frac{(m_{2z}^{(t)})^2}{\rho_z(\chi_{2z}^{(t)})^2} - \hat{\chi}_{2z}^{(t)} \tag{14.47m}$$

$$m_{1z}^{(t)} = \mathbb{E}\left[z_0 \eta_{g(y,.)/\hat{Q}_{1z}^{(t)}}\left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\right)\right] \tag{14.47n}$$

$$\chi_{1z}^{(t)} = \frac{1}{\hat{Q}_{1z}^{(t)}}\mathbb{E}\left[\eta'_{g(y,.)/\hat{Q}_{1z}^{(t)}}\left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\right)\right] \tag{14.47o}$$

$$q_{1z}^{(t)} = \mathbb{E}\left[\eta^2_{g(y,.)/\hat{Q}_{1z}^{(t)}}\left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\right)\right] \tag{14.47p}$$

$$\hat{Q}_{2z}^{(t+1)} = \frac{1}{\chi_{1z}^{(t)}} - \hat{Q}_{1z}^{(t)} \tag{14.47q}$$

$$\hat{m}_{2z}^{(t+1)} = \frac{m_{1z}^{(t)}}{\rho_z \chi_{1z}^{(t)}} - \hat{m}_{1z}^{(t)} \tag{14.47r}$$

$$\hat{\chi}_{2z}^{(t+1)} = \frac{q_{1z}^{(t)}}{(\chi_{1z}^{(t)})^2} - \frac{(m_{1z}^{(t)})^2}{\rho_z(\chi_{1z}^{(t)})^2} - \hat{\chi}_{1z}^{(t)} \tag{14.47s}$$

$$m_{2x}^{(t+1)} = \rho_x \mathbb{E}\left[\frac{\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t+1)}}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)}}\right] \tag{14.47t}$$

$$\chi_{2x}^{(t+1)} = \mathbb{E}\left[\frac{1}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)}}\right] \tag{14.47u}$$

$$q_{2x}^{(t+1)} = \mathbb{E}\left[\frac{\hat{\chi}_{2x}^{(t)} + \lambda \hat{\chi}_{2z}^{(t+1)}}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)})^2}\right] \tag{14.47v}$$

$$+ \rho_x \mathbb{E}\left[\frac{(\hat{m}_{2x}^{(t+1)} + \lambda \hat{m}_{2z}^{(t+1)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)})^2}\right] \tag{14.47w}$$

$$\hat{Q}_{1x}^{(t+1)} = \frac{1}{\chi_{2x}^{(t+1)}} - \hat{Q}_{2x}^{(t)} \tag{14.47x}$$

$$\hat{m}_{1x}^{(t+1)} = \frac{m_{2x}^{(t+1)}}{\rho_x \chi_{2x}^{(t+1)}} - \hat{m}_{2x}^{(t)} \tag{14.47y}$$

$$\hat{\chi}_{1x}^{(t+1)} = \frac{q_{2x}^{(t+1)}}{(\chi_{2x}^{(t+1)})^2} - \frac{(m_{2x}^{(t+1)})^2}{\rho_x(\chi_{2x}^{(t+1)})^2} - \hat{\chi}_{2x}^{(t)}. \tag{14.47z}$$

We are interested in the fixed point of these state evolution equations, where $\chi_{1x}^{(t)} = \chi_{2x}^{(t)} = \chi_x$, $q_{1x}^{(t)} = q_{2x}^{(t)} = q_x$, $m_{1x}^{(t)} = m_{2x}^{(t)} = m_x$, $\chi_{1z}^{(t)} = \chi_{2z}^{(t)} = \chi_z$, $q_{1z}^{(t)} = q_{2z}^{(t)} = q_z$, and $m_{1z}^{(t)} = m_{2z}^{(t)} = m_z$ are achieved. From there we easily recover eq. (13.13). However, these equations are not rigorous since the starting assumptions are not proven. Therefore, we will turn to a rigorous formalism to consolidate those results.

14.5.2 Necessary assumptions for the rigorous state evolution equations

Here we remind the main assumptions needed for the rigorous state evolution equations to hold, as they are listed for Theorem 1 of [97], and show they are verified in our setting.

Assumption 4.

- *the empirical distributions of the underlying truth* $\mathbf{x}_0$*, eigenvalues of* $\mathbf{F}^T\mathbf{F}$*, and noise vector* w_0*, respectively converge with second order moments, as defined in appendix 14.1, to independent scalar random variables* x_0, w_0, λ *with distributions* p_{x_0}*,* p_λ*,* p_{w_0}*. We assume that the distribution* p_λ *is not all-zero and has compact support.*
- *the design matrix* $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{M\times N}$ *is rotationally invariant, as defined in the introduction, where the elements of the Haar distributed matrices* $\mathbf{U}, \mathbf{V}$ *are independent of the random variables* x_0, w_0, λ
- *assume that* $M, N \to \infty$ *with fixed ratio* $\alpha = M/N$ *independent of* M, N*.*
- *the activation function* $\phi(., \mathbf{w}_0)$ *from Eq.(13.1) is pseudo-Lipschitz of order 2.*
- *the constants* $\left\langle \partial_{\mathbf{h}_{1x}^{(t)}} g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) \right\rangle, \left\langle \partial_{\mathbf{h}_{1z}^{(t)}} g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}) \right\rangle, \left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)}) \right\rangle$ $\left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)}) \right\rangle$ *from algorithm (1) are all in* $[0, 1]$*.*
- *the component estimation functions* $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}), g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}), g_{2x}(\mathbf{h}_{2x}^{(t)},$ $\mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)}), g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})$ *from algorithm (1) are uniformly Lipschitz continuous, at all time steps* t*, respectively in* $\mathbf{h}_{1x}^{(t)}$ *at* $\hat{Q}_{1x}^{(t)}$*, in* $\mathbf{h}_{1z}^{(t)}$ *at* $\hat{Q}_{1z}^{(t)}$*,* $\mathbf{h}_{2x}^{(t)}$ *at* $\hat{Q}_{2x}^{(t)}$ *and in* $\mathbf{h}_{2z}^{(t)}$ *at* $\hat{Q}_{2z}^{(t)}$*.*

The first four points are included in the set of assumptions 2 and are therefore verified. We need to check the last two points, starting with the function $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) = \text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)})$. Since proximal operators are firmly nonexpansive, they are 1-Lipschitz and we thus have, using the separability of the function f:

$$\left\langle \partial_{\mathbf{h}_{1x}^{(t)}} g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) \right\rangle = \frac{1}{N}\sum_{i=1}^{N} \text{Prox}'_{f_i/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x,i}^{(t)}) \in [0, 1] \tag{14.48}$$

where each $f_i : \mathbb{R} \to \mathbb{R}$ is the same function applied to each coordinates. Now consider the restriction of $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$ to its second argument. Its gradient w.r.t. $\hat{Q}_{1x}^{(t)}$ at a given point $\mathbf{h}_{1x}^{(t)}$ verifies, assuming the function f is differentiable:

$$\begin{aligned}
\left\| \nabla_{\hat{Q}_{1x}^{(t)}} \text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)}) \right\|_2 &= \left\| (Id + \frac{1}{\hat{Q}_{1x}^{(t)}} \mathcal{H}_f(\text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)})))^{-1} \nabla f(\mathbf{h}_{\mathbf{1x}}^{(\mathbf{t})}) \right\|_2 \\
&\leqslant \left\| \nabla f(\mathbf{h}_{1x}^{(t)}) \right\|_2 \\
&\leqslant C(1 + \left\| \mathbf{h}_{1x}^{(t)} \right\|_2)
\end{aligned} \tag{14.49}$$

where the last line is obtained using the scaling conditions on the subdifferential of f from assumption 2. Then, for any $\hat{Q}_{1x}^{(t)}, \hat{Q}_{1x}^{(t')}$, $\left\|\text{Prox}_{f/\hat{Q}_{1x}^{(t)}} - \text{Prox}_{f/\hat{Q}_{1x}^{(t')}}\right\|_2 \leqslant C(1+\left\|\mathbf{h}_{1x}^{(t)}\right\|_2)\left|\hat{Q}_{1x}^{(t)} - \hat{Q}_{1x}^{(t')}\right|$ and $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$ is uniformly Lipschitz in $\mathbf{h}_{1x}^{(t)}$ at $\hat{Q}_{1x}^{(t)}$, at any time index t. The argument is identical for $g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}) = \text{Prox}_{f/\hat{Q}_{1z}^{(t)}}(\mathbf{h}_{1z}^{(t)})$. The functions $g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)}), g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})$ have explicit expressions and it is straightforward to check the last two points using linear algebra and the assumptions on the spectrum of $\mathbf{F}^\top\mathbf{F}$.

14.5.3 Rigorous state evolution formalism

We now look into the state evolution equations derived for MLVAMP in [256]. Those equations are proven to be exact in the asymptotic limit, and follow the same algorithm as (1). In particular, they provide statistical properties of vectors $\mathbf{h}_{1x}, \mathbf{h}_{2x}, \mathbf{h}_{1z}, \mathbf{h}_{2z}$. We can read relations from [97] using the following dictionary between our notations and theirs, valid at each iteration of the algorithm:

$$\hat{Q}_{1x}, \hat{Q}_{2x}, \hat{Q}_{1z}, \hat{Q}_{2z} \longleftrightarrow \gamma_0^-, \gamma_0^+, \gamma_1^+, \gamma_1^- \tag{14.50a}$$
$$\chi_{1x}\hat{Q}_{1x}, \chi_{2x}\hat{Q}_{2x} \longleftrightarrow \alpha_0^-, \alpha_0^+ \tag{14.50b}$$
$$\chi_{1z}\hat{Q}_{1z}, \chi_{2z}\hat{Q}_{2z} \longleftrightarrow \alpha_1^-, \alpha_1^+ \tag{14.50c}$$
$$\mathbf{x_0}, \mathbf{z_0}, \rho_x, \rho_z \longleftrightarrow \mathbf{Q}_0^0, \mathbf{Q}_1^0, \tau_0^0, \tau_1^0 \tag{14.50d}$$
$$\mathbf{h}_{1x}, \mathbf{h}_{2x}, \mathbf{h}_{1z}, \mathbf{h}_{2z} \longleftrightarrow \mathbf{r}_0^-, \mathbf{r}_0^+, \mathbf{r}_1^+, \mathbf{r}_1^-. \tag{14.50e}$$

Placing ourselves in the asymptotic limit, [97] shows the following equalities:

$$\mathbf{r}_0^- = \mathbf{Q}_0^0 + \mathbf{Q}_0^- \tag{14.51a}$$
$$\mathbf{r}_0^+ = \mathbf{Q}_0^0 + \mathbf{Q}_0^+ \tag{14.51b}$$
$$\mathbf{r}_1^- = \mathbf{Q}_1^0 + \mathbf{Q}_1^- \tag{14.51c}$$
$$\mathbf{r}_1^+ = \mathbf{Q}_1^0 + \mathbf{Q}_1^+ \tag{14.51d}$$

where $\mathbf{Q}_0^- \sim \mathcal{N}(0, \tau_0^-)^N$ and $\mathbf{Q}_1^- \sim \mathcal{N}(0, \tau_1^-)^N$ are i.i.d. Gaussian vectors. $\mathbf{Q}_0^+$, $\mathbf{Q}_1^+$ have the following norms and non-zero correlations with ground-truth vectors $\mathbf{Q}_0^0, \mathbf{Q}_1^0$:

$$\tau_0^+ \equiv \frac{\left\|\mathbf{Q}_0^+\right\|_2^2}{N} \qquad c_0^+ \equiv \frac{\mathbf{Q}_0^{0T}\mathbf{Q}_0^+}{N} \tag{14.52}$$
$$\tau_1^+ \equiv \frac{\left\|\mathbf{Q}_1^+\right\|_2^2}{M} \qquad c_1^+ \equiv \frac{\mathbf{Q}_1^{0T}\mathbf{Q}_1^+}{M}. \tag{14.53}$$

With simple manipulations, we can rewrite (14.51) as:

$$\mathbf{r}_0^- \overset{\text{d}}{=} \mathbf{Q}_0 + \mathbf{Q}_0^- \tag{14.54a}$$
$$\mathbf{V}^T\mathbf{r}_0^+ \overset{\text{d}}{=} \left(1 + \frac{c_0^+}{\tau_0^0}\right)\mathbf{V}^T\mathbf{Q}_0^0 + \mathbf{V}^T\tilde{\mathbf{Q}}_0^+ \tag{14.54b}$$
$$\mathbf{r}_1^- \overset{\text{d}}{=} \mathbf{Q}_1^0 + \mathbf{Q}_1^- \tag{14.54c}$$
$$\mathbf{U}^T\mathbf{r}_1^+ \overset{\text{d}}{=} \left(1 + \frac{c_1^+}{\tau_1^0}\right)\mathbf{U}^T\mathbf{Q}_1^0 + \mathbf{U}^T\tilde{\mathbf{Q}}_1^+ \tag{14.54d}$$

where for $k \in \{1, 2\}$ vectors

$$\tilde{\mathbf{Q}}_k^+ = -\frac{c_k^+}{\tau_k^0}\mathbf{Q}_k^0 + \mathbf{Q}_k^+ \tag{14.55}$$

and $\mathbf{Q}_0^-, \mathbf{Q}_1^-$ have no correlation with ground-truth vectors $\mathbf{Q}_0^0$, $\mathbf{Q}_1^0$, $\mathbf{U}^T\mathbf{Q}_0^0$, $\mathbf{V}^T\mathbf{Q}_1^0$. Besides, Lemma 5 from [242] states that $\mathbf{V}^T\tilde{\mathbf{Q}}_0^+$ and $\mathbf{U}^T\tilde{\mathbf{Q}}_1^+$ have components that converge empirically to Gaussian variables, respectively $\mathcal{N}(0, \tau_0^+)$ and $\mathcal{N}(0, \tau_1^+)$. Let us now translate this in our own terms, using the following relations that complete our dictionary with state evolution parameters:

$$\frac{\hat{m}_{1x}}{\hat{Q}_{1x}} \longleftrightarrow 1 \qquad \frac{\hat{m}_{2z}}{\hat{Q}_{2z}} \longleftrightarrow 1 \tag{14.56a}$$

$$\frac{\hat{m}_{2x}}{\hat{Q}_{2x}} \longleftrightarrow 1 + \frac{c_0^+}{\tau_0^0} \qquad \frac{\hat{m}_{1z}}{\hat{Q}_{1z}} \longleftrightarrow 1 + \frac{c_1^+}{\tau_1^0} \tag{14.56b}$$

$$\frac{\hat{\chi}_{1x}}{\hat{Q}_{1x}^2} \longleftrightarrow \tau_0^- \qquad \frac{\hat{\chi}_{2z}}{\hat{Q}_{2z}^2} \longleftrightarrow \tau_1^- \tag{14.56c}$$

$$\frac{\hat{\chi}_{2x}}{\hat{Q}_{2x}^2} \longleftrightarrow \tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0} \qquad \frac{\hat{\chi}_{1z}}{\hat{Q}_{1z}^2} \longleftrightarrow \tau_1^+ - \frac{(c_1^+)^2}{\tau_1^0}. \tag{14.56d}$$

Simple bookkeeping transforms equations (14.54) into a rigorous statement of starting assumptions (14.51) from [277]. Since those assumptions are now rigorously established in the asymptotic limit, the remaining derivation of state evolution equations (14.47) holds and provides a mathematically exact statement.

14.5.4 Scalar equivalent model of state evolution

For the sake of completeness, we will provide an overview of the explicit matching between the state evolution formalism from [97] which was developed in a series of papers, and the replica formulation from [277] which relies on statistical physics methods. Although not necessary to our proof, it is interesting to develop an intuition about the correspondence between those two faces of the same coin. We have seen in the previous subsection that [97] introduces ground-truth vectors $\mathbf{Q}_0^0, \mathbf{Q}_1^0$, estimates $\mathbf{r}_0^\pm, \mathbf{r}_1^\pm$ which are related to vectors $\mathbf{Q}_0^\pm, \mathbf{Q}_1^\pm$. Let us introduce a few more vectors using matrices from the singular value decomposition $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Let $\mathbf{s}_\nu \in \mathbb{R}^N$ be the vector containing all square roots of eigenvalues of $\mathbf{F}^T\mathbf{F}$ with p_ν its element-wise distribution; and $\mathbf{s}_\mu \in \mathbb{R}^M$ the vector containing all square roots of eigenvalues of $\mathbf{F}\mathbf{F}^T$ with p_μ its element-wise distribution. Note that those two vectors contain the singular values of $\mathbf{F}$, but one of them also contains $\max(M, N) - \min(M, N)$ zero values. p_μ and p_ν are both well-defined since p_λ is properly defined in Assumptions 2. We also define

$$\begin{aligned} &\mathbf{P}_0^0 = \mathbf{V}^T\mathbf{Q}_0^0 \quad \mathbf{P}_0^+ = \mathbf{V}^T\mathbf{Q}_0^+ \quad \mathbf{P}_0^- = \mathbf{V}^T\mathbf{Q}_0^- \\ &\mathbf{P}_1^0 = \mathbf{U}\mathbf{Q}_1^0 \quad \mathbf{P}_1^+ = \mathbf{U}\mathbf{Q}_1^+ \quad \mathbf{P}_1^- = \mathbf{U}\mathbf{Q}_1^-. \end{aligned}$$

By virtue of Lemma 5 from [242], the six previous vectors have elements that converge empirically to a Gaussian variable. Hence, all defined vectors have an element-wise separable distribution, and we can write the state evolution as a scalar model on random variables sampled from those distributions. To do so, we will simply write the variables without the bold font: for instance $Z_0^0 \sim p_{x_0}$, $s_\nu \sim p_\nu$, and Q_0^- refers to the random variable distributed according to the element-wise

distribution of vector $\mathbf{Q}_0^-$. The scalar random variable state evolution from [97] now reads:

$$\text{Initialize } \gamma_1^{-(0)}, \gamma_0^{-(0)}, \tau_0^{-(0)}, \tau_1^{-(0)}, \tag{14.57a}$$
$$Q_0^{-(0)} \sim \mathcal{N}(0, \tau_0^{-(0)}), Q_1^{-(0)} \sim \mathcal{N}(0, \tau_1^{-(0)}), \alpha_0^{-(0)}, \alpha_1^{-(0)}$$
Initial pass (ground truth only)
$$s_\nu \sim p_\nu, \quad s_\mu \sim p_\mu, \quad Q_0^0 \sim p_{x_0} \tag{14.57b}$$
$$\tau_0^0 = \mathbb{E}[(Q_0^0)^2] \quad P_0^0 \sim \mathcal{N}(0, \tau_0^0) \tag{14.57c}$$
$$Q_1^0 = s_\mu P_0^0 \quad \tau_1^0 = \mathbb{E}[(s_\mu P_0^0)^2] = \mathbb{E}[(s_\mu)^2]\tau_0^0 \tag{14.57d}$$
$$P_1^0 \sim \mathcal{N}(0, \tau_1^0) \tag{14.57e}$$

Forward Pass (estimation):
$$\alpha_0^{+(t)} = \mathbb{E}\left[\eta'_{f/\gamma_0^{-(t)}}(Q_0^0 + Q_0^{-(t)})\right] \tag{14.57f}$$
$$\gamma_0^{+(t)} = \frac{\gamma_0^{(t)}}{\alpha_0^{+(t)}} - \gamma_0^{-(t)} \tag{14.57g}$$
$$Q_0^{+(t)} = \frac{1}{1-\alpha_0^{+(t)}}\Big\{\eta_{f/\gamma_0^{-(t)}}(Q_0^0 + Q_0^{-(t)}) - ... \\ Q_0^0 - \alpha_0^+ Q_0^{-(t)}\Big\} \tag{14.57h}$$
$$\mathbf{K}_0^{+(t)} = Cov\left(Q_0^0, Q_0^{+(t)}\right) \tag{14.57i}$$
$$\left(P_0^0, P_0^{+(t)}\right) \sim \mathcal{N}\left(0, \mathbf{K}_0^{+(t)}\right) \tag{14.57j}$$

$$\alpha_1^{+(t)} = \mathbb{E}\left[\frac{s_\mu^2 \gamma_1^{-(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}\right] \tag{14.57k}$$

$$\gamma_1^{+(t)} = \frac{\gamma_1^{-(t)}}{\alpha_1^{+(t)}} - \gamma_1^{-(t)} \tag{14.57l}$$

$$Q_1^{+(t)} = \frac{1}{1-\alpha_1^{+(t)}}\Bigg\{\frac{s_\mu^2 \gamma_1^{-(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}(Q_1^{-(t)} + Q_1^0) + ... \\ \frac{s_\mu \gamma_0^{+(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}(P_0^{+(t)} + P_0^0) - Q_1^0 - \alpha_1^{+(t)} Q_1^{-(t)}\Bigg\} \tag{14.57m}$$

$$\mathbf{K}_1^{+(t)} = Cov\left(Q_1^0, Q_1^{+(t)}\right) \tag{14.57n}$$

$$\left(P_1^0, P_1^{+(t)}\right) \sim \mathcal{N}\left(0, \mathbf{K}_1^{+(t)}\right) \tag{14.57o}$$

Backward Pass (estimation):

$$\alpha_1^{-(t+1)} = \mathbb{E}\left[\eta_{g(y,.)/\gamma_1^{+(t)}}(P_1^0 + P_1^{+(t)})\right] \tag{14.57p}$$

$$\gamma_1^{-(t+1)} = \frac{\gamma_1^{+(t)}}{\alpha_1^{-(t+1)}} - \gamma_1^{+(t)} \tag{14.57q}$$

$$P_1^{-(t+1)} = \frac{1}{1-\alpha_1^{-(t+1)}}\Bigg\{\eta_{g(y,.)/\gamma_1^{+(t)}}(P_1^0 + P_1^{+(t)}) \\ - P_1^0 - \alpha_1^{-(t+1)} P_1^{+(t)}\Bigg\} \tag{14.57r}$$

$$\tau_1^{-(t+1)} = \mathbb{E}\left[(P_1^{-(t+1)})^2\right] \qquad Q_1^{-(t+1)} \sim \mathcal{N}(0, \tau_1^{-(t+1)}) \tag{14.57s}$$

$$\alpha_0^{-(t+1)} = \mathbb{E}\left[\frac{\gamma_0^{+(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}\right] \tag{14.57t}$$

$$\gamma_0^{-(t+1)} = \frac{\gamma_0^{+(t)}}{\alpha_0^{-(t+1)}} - \gamma_0^{+(t)} \tag{14.57u}$$

$$P_0^{-(t+1)} = \frac{1}{1-\alpha_0^{-(t+1)}}\Bigg\{\frac{s_\nu \gamma_1^{-(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}(Q_1^{-(t+1)} + Q_1^0) \\ + \frac{\gamma_0^{+(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}(P_0^{+(t)} + P_0^0) - P_0^0 - \alpha_0^{-(t+1)} P_0^{+(t)}\Bigg\} \tag{14.57v}$$

$$\tau_0^{-(t+1)} = \mathbb{E}\left[(P_0^{-(t+1)})^2\right] \qquad Q_0^{-(t+1)} \sim \mathcal{N}(0, \tau_0^{-(t+1)}). \tag{14.57w}$$

14.5.5 Direct matching of the state evolution fixed point equations

To be consistent, we should be able to show that equations (14.57) allow us to recover equations (14.47) at their fixed point. Although somewhat tedious, this task is facilitated using dictionaries (14.50) and (14.56). We shall give here an overview of this matching through a few examples.

- Recovering equation (14.47e)

Let us start from the rigorous scalar state evolution, in particular equation (14.57h) that defines variable Q_0^+. We get rid of time indices here since we focus on the fixed point. We first compute the correlation

$$c_0^+ = \mathbb{E}\left[Q_0^0 Q_0^+\right] \tag{14.58}$$

$$= \frac{1}{1-\alpha_0^+}\left\{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] - \tau_0^0\right\} \tag{14.59}$$

where we have used $\mathbb{E}[(Q_0^0)^2] = \tau_0^0$. At the fixed point, we know from MLVAMP or simply translating equations (14.31), (14.33) that

$$1-\alpha_0^+ = \alpha_0^-, \qquad \frac{1}{\alpha_0^-} = \frac{\gamma_0^- + \gamma_0^+}{\gamma_0^+}, \qquad \gamma_0^+\alpha_0^+ = \gamma_0^-\alpha_0^-.$$

Simple manipulations take us to

$$c_0^+ = \frac{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]}{\alpha_0^-} - \tau_0^0\left(1+\frac{\gamma_0^-}{\gamma_0^+}\right) \tag{14.60}$$

$$\left(1+\frac{c_0^+}{\tau_0^0}\right)\gamma_0^+ = \frac{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\gamma_0^+}{\tau_0^0\alpha_0^-} - \gamma_0^-. \tag{14.61}$$

Now let us translate this back into our notations. The term $\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]$ simply translates into m_{1x}, and the rest of the terms can all be changed according to our dictionary. (14.61) exactly becomes

$$\hat{m}_{2x} = \frac{m_{1x}}{\rho_x \chi_x} - \hat{m}_{1x}, \tag{14.62}$$

hence we perfectly recover equations (14.47e) at the fixed point.

- Recovering equation (14.47f)

We start again from (14.57h) and square it:

$$\begin{aligned}\mathbb{E}\left[(Q_0^+)^2\right] = \frac{1}{(1-\alpha_0^+)^2}\Big\{&\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] + \dots \\ &(\alpha_0^+)^2\mathbb{E}\left[(Q_0^-)^2\right] - 2\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] \\ &- 2\alpha_0^+\mathbb{E}\left[Q_0^-\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-) + \mathbb{E}\left[(Q_0^0)^2\right]\right]\Big\}\end{aligned} \tag{14.63}$$

$$\begin{aligned}\tau_0^+ = \frac{1}{(1-\alpha_0^+)^2}\Big\{&\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] + \tau_0^0 + \dots \\ &(\alpha_0^+)^2\tau_0^- - 2\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] - \dots \\ &2\alpha_0^+\mathbb{E}\left[Q_0^-\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\Big\}.\end{aligned} \tag{14.64}$$

Since Q_0^- is a Gaussian variable, independent from Q_0^0, we can use Stein's lemma and use equation (14.57f) to get

$$\mathbb{E}\left[Q_0^-\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] = \alpha_0^+\tau_0^-. \tag{14.65}$$

Moreover, from (14.59) we have

$$(c_0^+)^2(\alpha_0^-)^2 = \left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]-\tau_0^0\right)^2 \tag{14.66}$$

$$\frac{(c_0^+)^2(\alpha_0^-)^2}{\tau_0^0} - \frac{(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right])^2}{\tau_0^0} = \ldots$$

$$-2\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]+\tau_0^0. \tag{14.67}$$

Replacing (14.65) and (14.67) into (14.64), we reach

$$\left(\tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0}\right)(\alpha_0^-)^2 = \mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]$$

$$-\frac{\left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\right)^2}{\tau_0^0} - (\alpha_0^+)^2\tau_0^- \tag{14.68}$$

$$\left(\tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0}\right)(\gamma_0^+)^2 = \frac{\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right](\gamma_0^+)^2}{(\alpha_0^-)^2}$$

$$-\frac{\left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\right)^2(\gamma_0^+)^2}{\tau_0^0(\alpha_0^-)^2} - (\gamma_0^-)^2\tau_0^-. \tag{14.69}$$

Notice that $\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]$ simply translates into our variable q_{1x} from its definition (14.47c), and our dictionary directly transforms (14.68) into equation (14.47f):

$$\hat{\chi}_{2x} = \frac{q_{1x}}{\chi_{1x}^2} - \frac{m_{1x}^2}{\rho_x\chi_{1x}^2} - \hat{\chi}_{1x}. \tag{14.70}$$

- Recovering equation (14.47t)

We first note that for any function h,

$$\mathbb{E}[h(s_\nu)] = \min(1,\alpha)\mathbb{E}[h(s_\mu)] + \max(0,1-\alpha)h(0). \tag{14.71}$$

and $s_\nu^2 \sim p_\lambda$. Applying this to $h(s) = \dfrac{\gamma_1^- s^2}{\gamma_1^- s^2+\gamma_0^+}$ and starting from (14.57m), we rewrite

$$\alpha_1^+ = \mathbb{E}\left[\frac{\gamma_1^- s_\mu^2}{\gamma_1^- s_\mu^2+\gamma_0^+}\right] \tag{14.72}$$

$$= \frac{1}{\alpha}\mathbb{E}\left[\frac{\gamma_1^-\lambda}{\gamma_1^-\lambda+\gamma_0^+}\right] \tag{14.73}$$

with $\lambda \sim p_\lambda$, which translates into equation (14.47t):

$$\chi_{2z} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right]. \tag{14.74}$$

In a similar fashion, we can recover all equations (14.47) by writing variances and correlations between scalar random variables defined in (14.57), and using the independence properties established in [97]; thus directly showing the matching between the two state evolution formalisms at their fixed point.

14.6 Numerical implementation details

The plots were generated using the toolbox available at https://github.com/cgerbelo/Replica_GLM_orth.inv.git

Here we give a few derivation details for implementation of the equations presented in Theorem 22. We provide the Python script used to produce the figures in the main body of the paper as an example. The experimental points were obtained using the convex optimization tools of [229], with a data matrix of dimension $N = 200, M = \alpha N$, for $\alpha \in [0.1, 3]$. Each point is averaged 100 times to get smoother curves. The theoretical prediction was simply obtained by iterating the equations from Theorem 22. This can lead to unstable numerical schemes, and we include a few comments about stability in the code provided with this version of the paper. For Gaussian data, the design matrices were simply obtained by sampling a normal distribution $\mathcal{N}(0, \sqrt{1/M})$, effectively yielding the Marchenko-Pastur distribution [285] for averaging on the eigenvalues of $\mathbf{F}^T\mathbf{F}$ in the state evolution equations :

$$\lambda_{\mathbf{F}^T\mathbf{F}} \sim \max(0, 1-\alpha)\delta(\lambda - 0) + \alpha \frac{\sqrt{(0, \lambda - a)^+ (0, b - \lambda)^+}}{2\pi\lambda} \tag{14.75}$$

where $a = \sqrt{1 - \left(\frac{1}{\alpha}\right)^2}, b = \sqrt{1 + \left(\frac{1}{\alpha}\right)^2}$, and $(0, x)^+ = \max(0, x)$. For the example of orthogonally invariant matrix with arbitrary spectrum, we chose to sample the singular values of $\mathbf{F}$ from the uniform distribution $\mathcal{U}([(1-\alpha)^2, (1+\alpha)^2])$. This leads to the following distribution for the eigenvalues of $\mathbf{F}^T\mathbf{F}$:

$$\lambda_{\mathbf{F}^T\mathbf{F}} \sim \max(0, 1-\alpha)\delta(0) + \min(1, \alpha) d(\lambda, \alpha) \tag{14.76}$$

where $d(\lambda, \alpha) = \left(\frac{1}{2((1+\alpha)^2 - (1-\alpha)^2)} \mathbb{I}_{\{\sqrt{\lambda} \in [(1-\alpha)^2, (1+\alpha)^2]\}} \frac{1}{\sqrt{\lambda}}\right)$, and $\mathbb{I}$ is the indicator function.

The only quantities that need additional calculus are the averages of proximals, squared proximals and derivatives of proximals. Here we give the corresponding expressions for the losses and regularizations that were used to make the figures. Note that the stability and convergence of the state evolution equations closely follow the result of Lemma 54. For example, a ridge regularized logistic regression, which is a strongly convex objective in both the loss (on compact spaces) and regularization will lead to more stable iterations than a LASSO SVC.

14.6.1 Regularization : elastic net

For the elastic net regularization, we can obtain an exact expression, avoiding any numerical integration. The proximal of the elastic net reads:

$$\text{Prox}_{\frac{1}{\hat{Q}_{1x}}(\lambda_1 |\mathbf{x}|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2)}(.) = \frac{1}{1 + \frac{\lambda_2}{\hat{Q}_{1x}}} s\left(., \frac{\lambda_1}{\hat{Q}_{1x}}\right) \tag{14.77}$$

where $s\left(., \frac{\lambda_1}{\hat{Q}_{1x}}\right)$ is the soft-thresholding function:

$$s\left(r_{1k}, \frac{\lambda_1}{\hat{Q}_{1x}}\right) = \begin{cases} r_{1k} + \frac{\lambda_1}{\hat{Q}_{1x}} & \text{if} \quad r_{1k} < -\frac{\lambda_1}{\hat{Q}_{1x}} \\ 0 & \text{if} \quad -\frac{\lambda_1}{\hat{Q}_{1x}} < r_{1k} < \frac{\lambda_1}{\hat{Q}_{1x}} \\ r_{1k} - \frac{\lambda_1}{\hat{Q}_{1x}} & \text{if} \quad r_{1k} > \frac{\lambda_1}{\hat{Q}_{1x}}. \end{cases} \tag{14.78}$$

$$\mathbb{E}[\mathrm{Prox}^2_{\mathbf{f}/\hat{Q}_{1x}}(X)]$$

$$= \left(\frac{1}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\right)^2 \Bigg[(1-\rho)\left(\frac{\lambda_1^2+\hat{\chi}_{1x}}{(\hat{Q}_{1x})^2}\,\mathrm{erfc}\left(\frac{\lambda_1}{\sqrt{2\hat{\chi}_{1x}}}\right) - \frac{\lambda_1\sqrt{2\hat{\chi}_{1x}}\exp\left(-\frac{\lambda_1^2}{2(\hat{\chi}_{1x})}\right)}{\sqrt{\pi}}\right) \tag{14.80}$$

$$+\rho\Bigg(\frac{\lambda_1^2+\hat{\chi}_{1x}+\sigma^2\hat{m}_{1x}^2}{(\hat{Q}_{1x})^2}\,\mathrm{erfc}\left(\frac{\lambda_1}{\sqrt{2(\hat{\chi}_{1x}+\sigma^2\hat{m}_{1x}^2)}}\right)$$

$$- \frac{\lambda_1\sqrt{2(\hat{\chi}_{1x}+\sigma^2\hat{m}_{1x}^2)}\exp\left(-\frac{\lambda_1^2}{2(\hat{Q}_{1x})^2(\hat{\chi}_{1x}+\sigma^2\hat{m}_{1x}^2)}\right)}{\sqrt{\pi}}\Bigg)\Bigg] \tag{14.81}$$

Similarly, we have

$$\mathbb{E}[\mathrm{Prox}'_{\mathbf{f}/\hat{Q}_{1x}}(X)] = \frac{1}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\left[(1-\rho)\,\mathrm{erfc}\left(\frac{\lambda_1}{\sqrt{2\hat{\chi}_{1x}}}\right) + \rho\,\mathrm{erfc}\left(\frac{\lambda_1}{\sqrt{2(\hat{\chi}_{1x}+\sigma^2\hat{m}_{1x}^2)}}\right)\right] \tag{14.82}$$

and

$$\mathbb{E}[x_0\mathrm{Prox}_{\mathbf{f}/\hat{Q}_{1x}}(X)] = \frac{\rho|\sigma\hat{m}_{1x}|}{\hat{Q}_{1x}+\lambda_2}\,\mathrm{erfc}\left(\frac{\lambda_1}{\sqrt{2(\hat{\chi}_{1x}+\sigma^2\hat{m}_{1x}^2)}}\right) \tag{14.83}$$

We assume that the ground-truth x_0 is pulled from a Gauss-Bernoulli law of the form:

$$\phi(x_0) = (1-\rho)\delta(0) + \rho\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big\{(-x_0^2/(2\sigma^2))\Big\}. \tag{14.79}$$

Note that we did our plots with $\rho = 1$, but this form can be used to study the effect of sparsity in the model. Writing $X = \frac{\hat{m}_{1x}x_0+\sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}$, and remembering that $\xi_{1x} \sim \mathcal{N}(0,1)$, some calculus then shows that: We now turn to the loss functions.

14.6.2 Loss functions

The loss functions sometimes have no closed form, as is the case for the logistic loss. In that case, numerical integration cannot be avoided, and we recommend marginalizing all the possible variables that can be averaged out. In the present model, if the teacher y is chosen as a sign, one-dimensional integrals can be reached, leading to stable and reasonably fast implementation (a few minutes to generate a curve comparable to those of Figure 13.1 for the non-linear models, the ridge regression being very fast). The interested reader can find the corresponding marginalized prefactors in the code jointly provided with this paper.

Square loss The square loss is defined as:

$$f(x,y) = \frac{1}{2}(x-y)^2, \tag{14.84}$$

its proximal and partial derivative then read:

$$\text{Prox}_{\frac{1}{\gamma}f}(p) = \frac{\gamma}{1+\gamma}p + \frac{1}{1+\gamma}y \tag{14.85}$$

$$\frac{\partial}{\partial p}\text{Prox}_{\frac{1}{\gamma}f}(p) = \frac{\gamma}{1+\gamma}. \tag{14.86}$$

Using this form with a plain ridge penalty (elastic net with $\ell_1 = 0$) leads to great simplification in the equations of Theorem 22 and we recover the classical expressions obtained for ridge regression in papers such as [125, 109].

Hinge loss The hinge loss reads:

$$f(x, y) = \max(0, 1 - yx). \tag{14.87}$$

Assuming $y \in \{-1, +1\}$, its proximal and partial derivative then read:

$$\text{Prox}_{\frac{1}{\gamma}f}(p) = \begin{cases} p + \frac{y}{\gamma} & \text{if} \quad \gamma(1-yp) \geqslant 1 \\ y & \text{if} \quad 0 \leqslant \gamma(1-yp) \leqslant 1 \\ p & \text{if} \quad \gamma(1-yp) \leqslant 0 \end{cases} \tag{14.88}$$

$$\frac{\partial}{\partial p}\text{Prox}_{\frac{1}{\gamma}f}(p) = \begin{cases} 1 & \text{if} \quad \gamma(1-yp) \geqslant 1 \\ 0 & \text{if} \quad 0 \leqslant \gamma(1-yp) \leqslant 1 \\ 1 & \text{if} \quad \gamma(1-yp) \leqslant 0. \end{cases} \tag{14.89}$$

Logistic loss

$$f(x, y) = \log(1 + \exp(-yx)) \tag{14.90}$$

Its proximal (at point p) is the solution to the fixed point problem:

$$x = p + \frac{y}{\gamma(1 + \exp(yx))}, \tag{14.91}$$

and its derivative, given that the logistic loss is twice differentiable, reads:

$$\frac{\partial}{\partial p}\text{Prox}_{\frac{1}{\gamma}f}(p) = \frac{1}{1 + \frac{1}{\gamma}\frac{\partial^2}{\partial p^2}f(\text{Prox}_{\frac{1}{\gamma}f}(p))} \tag{14.92}$$

$$= \frac{1}{1 + \frac{1}{\gamma}\frac{1}{(2+2\cosh(\text{Prox}_{\frac{1}{\gamma}f}(p))}}. \tag{14.93}$$

14.7 Proof of Lemma 54: Convergence analysis of 2-layer ML-VAMP

In this section, we give the detail of the convergence proof of 2-layer MLVAMP.

14.7.1 Proof of Proposition 9

This proof is quite straightforward and close to the one of Theorem 4 from [166].

Multiplying Eq.(13.55) on the left and right by $[(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})^\top \quad (\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)})^\top]$ and its transpose respectively, we get

$$
\begin{aligned}
&(\mathbf{A}^{(t)}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})+\mathbf{B}^{(t)}(\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}))^\top\mathbf{P}(\mathbf{A}^{(t)}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})+\mathbf{B}^{(t)}(\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}))\\
&-(\tau_{(t)})^2(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})^\top\mathbf{P}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})\\
&+\beta_1^{(t)}(\mathbf{C}_1^{(t)}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})+\mathbf{D}_1(\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}))^\top\mathbf{M}_1^{(t)}(\mathbf{C}_1^{(t)}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})+\mathbf{D}_1(\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}))\\
&+\beta_2^{(t)}(\mathbf{C}_2^{(t)}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})+\mathbf{D}_2(\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}))^\top\mathbf{M}_2^{(t)}(\mathbf{C}_2^{(t)}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})+\mathbf{D}_2(\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}))\leqslant 0
\end{aligned}
$$

Using the definition of the iteration (13.46)-(13.48), this simplifies to

$$
\begin{aligned}
&(\mathbf{h}^{(t+1)}-\mathbf{h}^{(t)})^\top\mathbf{P}(\mathbf{h}^{(t+1)}-\mathbf{h}^{(t)})-(\tau_{(t)})^2(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})^\top\mathbf{P}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})\\
&+\beta_1(\mathbf{w}_1^{(t)}-\mathbf{w}_1^{(t-1)})^\top\mathbf{M}_1^{(t)}(\mathbf{w}_1^{(t)}-\mathbf{w}_1^{(t-1)})+\beta_2(\mathbf{w}_2^{(t)}-\mathbf{w}_2^{(t-1)})^\top\mathbf{M}_2^{(t)}(\mathbf{w}_2^{(t)}-\mathbf{w}_2^{(t-1)})\leqslant 0
\end{aligned}
$$

Owing to the Lipschitz properties of $\tilde{\mathcal{O}}_1^{(t)},\tilde{\mathcal{O}}_2^{(t)}$ and the definitions of $\mathbf{w}_1^{(t)},\mathbf{w}_2^{(t)}$, the terms factoring β_1,β_2 are both non-negative. We thus have, at each time step t:

$$
(\mathbf{h}^{(t+1)}-\mathbf{h}^{(t)})^\top\mathbf{P}(\mathbf{h}^{(t+1)}-\mathbf{h}^{(t)})\leqslant\tau_{(t)}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)})^\top\mathbf{P}(\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)}) \tag{14.94}
$$

Letting $\tau^*=\sup_t\tau_{(t)}$, an immediate induction concludes the proof.

14.7.2 Bounds on $\hat{Q}_{1x}^{(t+1)},\hat{Q}_{1z}^{(t)},\hat{Q}_{2x}^{(t)},\hat{Q}_{2z}^{(t+1)}$

We remind that, since the functions f and g are separable, their Hessians are diagonal matrices. For any time index t, the following bounds hold:

$\hat{Q}_{2x}^{(t)}$:

$$
\hat{Q}_{2x}^{(t)}=1/\chi_{1x}^{(t)}-\hat{Q}_{1x}^{(t)}\qquad\text{where}\qquad\chi_{1x}^{(t)}=\left\langle\partial_{\mathbf{h}_{1x}^{(t)}}g_{1x}(...)\right\rangle/\hat{Q}_{1x}^{(t)}, \tag{14.95}
$$

$$
\text{then}\qquad\frac{1}{\hat{Q}_{2x}^{(t)}+\hat{Q}_{1x}^{(t)}}=\frac{1}{N}\left(\mathrm{Tr}\left[(\hat{Q}_{1x}^{(t)}Id+\mathcal{H}_f(\mathrm{prox}))^{-1}\right]\right), \tag{14.96}
$$

$$
\hat{Q}_{1x}^{(t)}+\lambda_{min}(\mathcal{H}_f)\leqslant\hat{Q}_{1x}^{(t)}+\hat{Q}_{2x}^{(t)}\leqslant\hat{Q}_{1x}^{(t)}+\lambda_{max}(\mathcal{H}_f). \tag{14.97}
$$

$\hat{Q}_{2z}^{(t+1)}$:

$$
\hat{Q}_{2z}^{(t+1)}=1/\chi_{1z}^{(t)}-\hat{Q}_{1z}^{(t)}\qquad\text{where}\qquad\chi_{1z}^{(t)}=\left\langle\partial_{\mathbf{h}_{1z}^{(t)}}g_{1z}(...)\right\rangle/\hat{Q}_{1z}^{(t)}, \tag{14.98}
$$

$$
\text{then}\qquad\frac{1}{\hat{Q}_{2z}^{(t+1)}+\hat{Q}_{1z}^{(t)}}=\frac{1}{M}\left(\mathrm{Tr}\left[(\hat{Q}_{1z}^{(t)}Id+\mathcal{H}_g(\mathrm{prox}))^{-1}\right]\right), \tag{14.99}
$$

$$
\hat{Q}_{1z}^{(t)}+\lambda_{min}(\mathcal{H}_g)\leqslant\hat{Q}_{1z}^{(t)}+\hat{Q}_{2z}^{(t+1)}\leqslant\hat{Q}_{1z}^{(t)}+\lambda_{max}(\mathcal{H}_g). \tag{14.100}
$$

$\hat{Q}_{1z}^{(t)}$:

$$\hat{Q}_{1z}^{(t)} = 1/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)} \qquad \chi_{2z}^{(t)} = \left\langle \partial_{\mathbf{h}_{2z}^{(t)}} g_{2z}(...) \right\rangle / \hat{Q}_{2z}^{(t)} \tag{14.101}$$

$$\text{then} \qquad \frac{1}{\hat{Q}_{1z}^{(t)} + \hat{Q}_{2z}^{(t)}} = \frac{1}{M} \text{Tr} \left[\mathbf{F}\mathbf{F}^\top \left(\hat{Q}_{2z}^{(t)} \mathbf{F}\mathbf{F}^\top + \hat{Q}_{2x}^{(t)} Id \right)^{-1} \right] \tag{14.102}$$

The matrices on the r.h.s. of the previous equation are all diagonalisable in the same basis. Then each eigenvalue has the form

$$\frac{\lambda_{min}(\mathbf{F}\mathbf{F}^\top)}{\hat{Q}_{2z}^{(t)} \lambda_{min}(\mathbf{F}\mathbf{F}^\top) + \hat{Q}_{2x}^{(t)}} \leqslant \frac{\lambda_k(\mathbf{F}\mathbf{F}^\top)}{\hat{Q}_{2z}^{(t)} \lambda_k(\mathbf{F}\mathbf{F}^\top) + \hat{Q}_{2x}^{(t)}} \leqslant \frac{\lambda_{max}(\mathbf{F}\mathbf{F}^\top)}{\hat{Q}_{2z}^{(t)} \lambda_{max}(\mathbf{F}\mathbf{F}^\top) + \hat{Q}_{2x}^{(t)}}, \tag{14.103}$$

which leads to the bound

$$\hat{Q}_{2z}^{(t)} + \frac{\hat{Q}_{2x}^{(t)}}{\lambda_{max}(\mathbf{F}\mathbf{F}^\top)} \leqslant \hat{Q}_{1z}^{(t)} + \hat{Q}_{2z}^{(t)} \leqslant \hat{Q}_{2z}^{(t)} + \frac{\hat{Q}_{2x}^{(t)}}{\lambda_{min}(\mathbf{F}\mathbf{F}^\top)}. \tag{14.104}$$

$\hat{Q}_{1x}^{(t+1)}$:

$$\hat{Q}_{1x}^{(t+1)} = 1/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)} \qquad \chi_{2x}^{(t+1)} = \left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2x}(...) \right\rangle / \hat{Q}_{2x}^{(t)}, \tag{14.105}$$

$$\text{then} \qquad \frac{1}{\hat{Q}_{1x}^{(t+1)} + \hat{Q}_{2x}^{(t)}} = \frac{1}{N} \text{Tr} \left[\left(\hat{Q}_{2z}^{(t+1)} \mathbf{F}^\top \mathbf{F} + \hat{Q}_{2x}^{(t)} Id \right)^{-1} \right], \tag{14.106}$$

which leads to

$$\hat{Q}_{2x}^{(t)} + \lambda_{min}(\mathbf{F}^\top \mathbf{F}) \hat{Q}_{2z}^{(t+1)} \leqslant \hat{Q}_{1x}^{(t+1)} + \hat{Q}_{2x}^{(t)} \leqslant \hat{Q}_{2x}^{(t)} + \lambda_{max}(\mathbf{F}^\top \mathbf{F}) \hat{Q}_{2z}^{(t+1)}. \tag{14.107}$$

14.7.3 Operator norms and Lipschitz constants

Operator norms of matrices $\mathbf{W_1}^{(t)}, \mathbf{W_2}^{(t)}, \mathbf{W_3}^{(t)}, \mathbf{W_4}^{(t)}$

The norms of the linear operators $\mathbf{W_1}^{(t)}, \mathbf{W_2}^{(t)}, \mathbf{W_3}^{(t)}, \mathbf{W_4}^{(t)}$ can be computed or bounded with respect to the singular values of the matrix $\mathbf{F}$. The derivations are straightforward and do not require any specific mathematical result. Denoting $\|\mathbf{W}\|$ the operator norm of a given matrix $\mathbf{W}$,

we have the following:

$$\left\|\mathbf{W_1}^{(t)}\right\| = \frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t+1)}} \max\left(\frac{|\hat{Q}_{1x}^{(t+1)} - \hat{Q}_{2z}^{(t+1)}\lambda_{min}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\lambda_{min}(\mathbf{F}^T\mathbf{F})}, \right. \tag{14.108}$$

$$\left. \frac{|\hat{Q}_{1x}^{(t+1)} - \hat{Q}_{2z}^{(t+1)}\lambda_{max}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\lambda_{max}(\mathbf{F}^T\mathbf{F})}\right) \tag{14.109}$$

$$\left\|\mathbf{W_2}^{(t)}\right\| = \frac{\hat{Q}_{2z}^{(t+1)}}{\chi_{2x}^{(t+1)}\hat{Q}_{1x}^{(t+1)}} \frac{\sqrt{\lambda_{max}(\mathbf{F}^T\mathbf{F})}}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\lambda_{min}(\mathbf{F}^T\mathbf{F})} \tag{14.110}$$

$$\left\|\mathbf{W_3}^{(t)}\right\| = \frac{\hat{Q}_{2z}^{(t)}}{\hat{Q}_{1z}^{(t)}} \max\left(\frac{|\hat{Q}_{2x}^{(t)} - \hat{Q}_{1z}^{(t)}\lambda_{min}(\mathbf{F}\mathbf{F}^T)|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\lambda_{min}(\mathbf{F}\mathbf{F}^T)}, \right. \tag{14.111}$$

$$\left. \frac{|\hat{Q}_{2x}^{(t)} - \hat{Q}_{1z}^{(t)}\lambda_{max}(\mathbf{F}\mathbf{F}^T)|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\lambda_{max}(\mathbf{F}\mathbf{F}^T)}\right) \tag{14.112}$$

$$\left\|\mathbf{W_4}^{(t)}\right\| = \frac{\hat{Q}_{2x}^{(t)}}{\chi_{2z}^{(t)}\hat{Q}_{1z}^{(t)}} \frac{\sqrt{\lambda_{max}(\mathbf{F}^T\mathbf{F})}}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\lambda_{min}(\mathbf{F}^T\mathbf{F})}. \tag{14.113}$$

Lispchitz constants of $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$

We now derive upper bounds of the Lipschitz constants of $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$ using the convex analysis reminder in appendix 14.2. We give detail for $\tilde{\mathcal{O}}_1^{(t)}$, the derivation is identical for $\tilde{\mathcal{O}}_2^{(t)}$. Let $(\sigma_1, \beta_1) \in \mathbb{R}_+^{*2}$ be the strong-convexity and smoothness constants of f, if they exist. If f has no strong convexity constant, we set $\sigma_1 = 0$, and if it holds no smoothness assumption, we set $\beta_1 = +\infty$. Note that, from the upper and lower bounds obtained in appendix 14.7.2, we have $\sigma_1 \leqslant \hat{Q}_{2x}^{(t)} \leqslant \beta_1$.

Case 1: $0 < \sigma_1 < \beta_1$ Proposition 11 gives the following expression:

$$\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f} = \frac{1}{2}\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}} + \frac{1}{1+\beta_1/\hat{Q}_{1x}^{(t)}}\right)\mathrm{Id} \tag{14.114}$$

$$+ \frac{1}{2}\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}} - \frac{1}{1+\beta_1/\hat{Q}_{1x}^{(t)}}\right)S_1 \tag{14.115}$$

where S_1 is a non-expansive operator. Replacing in the expression of $\tilde{\mathcal{O}}_1$ leads to:

$$\tilde{\mathcal{O}}_1^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\left(\left(\frac{1}{2\chi_{1x}^{(t)}}\left(\frac{1}{\hat{Q}_{1x}^{(t)}+\sigma_1} + \frac{1}{\hat{Q}_{1x}^{(t)}+\beta_1}\right) - 1\right)\mathrm{Id}\right. \tag{14.116}$$

$$\left. + \frac{1}{2\chi_{1x}^{(t)}}\left(\frac{1}{\hat{Q}_{1x}^{(t)}+\sigma_1} - \frac{1}{\hat{Q}_{1x}^{(t)}+\beta_1}\right)S_1\right) \tag{14.117}$$

$$\left\|\tilde{\mathcal{O}}_1^{(t)}(x)-\tilde{\mathcal{O}}_1^{(t)}(y)\right\|_2^2=\left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\left(\frac{1}{(\hat{Q}_{1x}^{(t)})^2(\chi_{1x}^{(t)})^2}\left\|\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(x)-\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(y)\right\|_2^2\right. \tag{14.120}$$

$$\left.-2\frac{1}{\hat{Q}_{1x}^{(t)}\chi_{1x}^{(t)}}\left\langle x-y,\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(x)-\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(y)\right\rangle+\|x-y\|_2^2\right) \tag{14.121}$$

$$\leqslant\left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\left(\left(\frac{1}{(\hat{Q}_{1x}^{(t)})^2(\chi_{1x}^{(t)})^2}-2\frac{1}{\hat{Q}_{1x}^{(t)}\chi_{1x}^{(t)}}\right)\left\|\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}}f}(x)-\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(y)\right\|_2^2+\|x-y\|_2^2\right) \tag{14.122}$$

$$=\left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\left(\left(\frac{1}{(\hat{Q}_{1x}^{(t)})^2(\chi_{1x}^{(t)})^2}-2\frac{1}{\hat{Q}_{1x}^{(t)}\chi_{1x}^{(t)}}\right)\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}}\right)^2+1\right)\|x-y\|_2^2 \tag{14.123}$$

$$=\left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\left(\frac{(\hat{Q}_{2x}^{(t)})^2-(\hat{Q}_{1x}^{(t)})^2}{(\hat{Q}_{1x}^{(t)}+\sigma_1)^2}+1\right)\|x-y\|_2^2. \tag{14.124}$$

---

which, knowing that $\hat{Q}_{1x}^{(t)}+\hat{Q}_{2x}^{(t)}=\frac{1}{\chi_{1x}^{(t)}}$, and separating the case where the first term of the sum in Eq.(14.116) is negative or positive, $\tilde{\mathcal{O}}_1$ has Lipschitz constant:

$$\omega_1^{(t)}=\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\max\left(\frac{\hat{Q}_{2x}^{(t)}-\sigma_1}{\hat{Q}_{1x}^{(t)}+\sigma_1},\frac{\beta_1-\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t)}+\beta_1}\right). \tag{14.118}$$

Case 2: $0<\sigma_1=\beta_1$ In this case, we have from Proposition 11:

$$\left\|\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(x)-\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(y)\right\|_2^2=\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}}\right)^2\|x-y\|_2^2 \tag{14.119}$$

which, with the firm non-expansiveness of the proximal operator gives, for any $x,y\in\mathbb{R}$: The upper bound on the Lipschitz constant is therefore:

$$\omega_1=\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\sqrt{1+\frac{((\hat{Q}_{2x}^{(t)})^2-(\hat{Q}_{1x}^{(t)})^2)}{(\hat{Q}_{1x}^{(t)}+\sigma_1)^2}}. \tag{14.125}$$

Case 3: no strong convexity or smoothness assumption This setting is not necessary for our proof, because we only handle penalty functions which have a strictly positive strong convexity constant, by adding a ridge term. However, we list it for completeness. In this case, the only information we have is the firm nonexpansiveness of the proximal operator, which leads us to the same derivation as the previous one up to (14.122), where the first term in the sum can be positive or negative. This yields the Lipschitz constant:

$$\omega_1^{(t)}=\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\max\left(1,\frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right). \tag{14.126}$$

Recovering (13.52) In our proof, we make no assumption on the strong-convexity or smoothness of the function, but adding the ridge penalties $\lambda_2, \tilde{\lambda}_2$ brings us for both $\tilde{\mathcal{O}}_1^{(t)}$ and $\tilde{\mathcal{O}}_2^{(t)}$ to either the first of the second case above. It is straightforward to see that the Lipschitz constant (14.125) is an upper bound of (14.118). We thus use (14.125) for generality, and recover the expressions (13.52) shown in the main body of the paper.

$$\omega_1^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}} \sqrt{1 + \frac{(\hat{Q}_{2x}^{(t)})^2 - (\hat{Q}_{1x}^{(t)})^2}{(\hat{Q}_{1x}^{(t)} + \lambda_2)^2}} \tag{14.127}$$

$$\omega_2^{(t)} = \frac{\hat{Q}_{1z}^{(t)}}{\hat{Q}_{2z}^{(t)}} \sqrt{1 + \frac{(\hat{Q}_{2z}^{(t)})^2 - (\hat{Q}_{1z}^{(t)})^2}{(\hat{Q}_{1z}^{(t)} + \tilde{\lambda}_2)^2}}. \tag{14.128}$$

14.7.4 Dynamical system convergence analysis

We are now ready to prove Lemma 54.

We will use the bounds derived above to prove the convergence lemma. Since we have proved the required bounds at any time step, we drop the time indices in the remainder of this proof for simplicity. The choice of additional regularization is λ_2 arbitrarily large, and $\tilde{\lambda}_2$ fixed but finite and non-zero. $\hat{Q}_{2x}, \hat{Q}_{1z}$ can thus be made arbitrarily large, and $\hat{Q}_{2z}, \hat{Q}_{1x}$ remain finite. We write the corresponding linear matrix inequality (13.55) and expand the constraint term. Some algebra shows that:

$$\mathbf{C}_1^T \mathbf{M}_1 \mathbf{C}_1 = \begin{bmatrix} \mathbf{0}_{M\times M} & 0_{M\times N} \\ \mathbf{0}_{N\times M} & \omega_1^2 I_{N\times N} \end{bmatrix} \tag{14.129}$$

$$\mathbf{C}_2^T \mathbf{M}_2 \mathbf{C}_2 = \begin{bmatrix} \omega_2^2 \mathbf{W}_3^T \mathbf{W}_3 & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N} \end{bmatrix} \tag{14.130}$$

$$\mathbf{C}_1^T \mathbf{M}_1 \mathbf{D}_1 = \mathbf{0}_{(M+N)\times(M+N)} \tag{14.131}$$

$$\mathbf{D}_1^T \mathbf{M}_1 \mathbf{C}_1 = \mathbf{0}_{(M+N)\times(M+N)} \tag{14.132}$$

$$\mathbf{C}_2^T \mathbf{M}_2 \mathbf{D}_2 = \begin{bmatrix} \mathbf{0}_{M\times M} & \omega_2^2 \mathbf{W}_3^T \mathbf{W}_4 \\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N} \end{bmatrix} \tag{14.133}$$

$$\mathbf{D}_2^T \mathbf{M}_2 \mathbf{C}_2 = \begin{bmatrix} \mathbf{0}_{M\times M} & \mathbf{0}_{M\times N} \\ \omega_2^2 \mathbf{W}_4^T \mathbf{W}_3 & \mathbf{0}_{N\times N} \end{bmatrix} \tag{14.134}$$

$$\mathbf{D}_1^T \mathbf{M}_1 \mathbf{D}_1 = \begin{bmatrix} \mathbf{0}_{M\times M} & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & -\mathbf{I}_{N\times N} \end{bmatrix} \tag{14.135}$$

$$\mathbf{D}_2^T \mathbf{M}_2 \mathbf{D}_2 = \begin{bmatrix} -\mathbf{I}_{M\times M} & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & \omega_2^2 \mathbf{W}_4^T \mathbf{W}_4 \end{bmatrix} \tag{14.136}$$

where all the matrices constituting the blocks have been defined in section 13.6. This gives the following form for the constraint matrix:

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_2^T & \mathbf{H}_3 \end{bmatrix} \tag{14.137}$$

where

$$\mathbf{H}_1 = \begin{bmatrix} \beta_1\omega_2^2\mathbf{W}_3^T\mathbf{W}_3 & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & \beta_0\omega_1^2\mathbf{I}_{N\times N} \end{bmatrix} \tag{14.138}$$

$$\mathbf{H}_2 = \begin{bmatrix} \mathbf{0}_{M\times M} & \beta_1\omega_2^2\mathbf{W}_3^T\mathbf{W}_4 \\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N} \end{bmatrix} \tag{14.139}$$

$$\mathbf{H}_3 = \begin{bmatrix} -\beta_1\mathbf{I}_{M\times M} & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & -\beta_0\mathbf{I}_{N\times N} + \beta_1\omega_2^2\mathbf{W}_4^T\mathbf{W}_4 \end{bmatrix} \tag{14.140}$$

thus the LMI (13.55) becomes:

$$0 \succeq \begin{bmatrix} -\tau^2\mathbf{P} + \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_2^T & \mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3 \end{bmatrix}. \tag{14.141}$$

We take $\mathbf{P}$ as block diagonal:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & \mathbf{P}_2 \end{bmatrix} \tag{14.142}$$

where $\mathbf{P}_1 \in \mathbb{R}^{M\times M}$ and $\mathbf{P}_2 \in \mathbb{R}^{N\times N}$ are positive definite (no zero eigenvalues) and diagonalizable in the same basis as $\mathbf{F}^T\mathbf{F}$, which is also the eigenbasis of $\mathbf{W}_1, \mathbf{W}_3, \mathbf{W}_2^T\mathbf{W}_2, \mathbf{W}_4^T\mathbf{W}_4$. We then have:

$$\mathbf{B}^T\mathbf{P}\mathbf{B} = \begin{bmatrix} \mathbf{P}_1 + \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2 & \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1 \\ \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2 & \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 \end{bmatrix}. \tag{14.143}$$

We are then trying to find the conditions for the following problem to be feasible with $0 < \tau < 1$:

$$\begin{bmatrix} \tau^2\mathbf{P} - \mathbf{H}_1 & -\mathbf{H}_2 \\ -\mathbf{H}_2^T & -(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \end{bmatrix} \succeq 0 \tag{14.144}$$

Schur's lemma then gives that the strict version of (14.144), which we will consider, is equivalent [127] to:

$$-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \succ 0 \quad \text{and} \tag{14.145}$$

$$\tau^2\mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T \succ 0 \tag{14.146}$$

We start with $-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)$.

Conditions for $-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \succ 0$

Expanding $-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \succ 0$ and applying Schur's lemma again gives the equivalent problem:

$$\beta_1\mathbf{I}_{N\times N} - \beta_2\omega_2^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 \succ 0 \quad \text{and} \tag{14.147}$$

$$\beta_2\mathbf{I}_{M\times M} - \mathbf{P}_1 - \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2 - \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1\mathbf{K}_1\mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2 \succ 0. \tag{14.148}$$

where $\mathbf{K}_1 = (\beta_1 \mathbf{I}_{N\times N} - \beta_2 \omega_2^2 \mathbf{W}_4^T \mathbf{W}_4 - \mathbf{W}_1^T \mathbf{P}_2 \mathbf{W}_1)^{-1}$. We start with (14.147). A sufficient condition for it to hold true is:

$$\beta_1 > \beta_2 \omega_2^2 \lambda_{max}(\mathbf{W}_4^T \mathbf{W}_4) + \lambda_{max}(\mathbf{P}_2)\lambda_{max}(\mathbf{W}_1^T \mathbf{W}_1). \tag{14.149}$$

Using the bounds from appendix 14.7.3, we have:

$$\begin{aligned}\lambda_{max}(\mathbf{W}_1^T \mathbf{W}_1) &\leqslant \left(\frac{\hat{Q}_{2x}}{\hat{Q}_{1x}}\right)^2 \max\Bigg(... \\ &\frac{|\hat{Q}_{1x} - \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\hat{Q}_{1x} - \hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})}\Bigg)^2 \qquad (14.150) \\ &\leqslant \max\Bigg(\left(1 - \frac{\hat{Q}_{2z}}{\hat{Q}_{1x}}\lambda_{min}(\mathbf{F}^T\mathbf{F})\right)^2, \\ &\qquad \left(1 - \frac{\hat{Q}_{2z}}{\hat{Q}_{1x}}\lambda_{max}(\mathbf{F}^T\mathbf{F})\right)^2\Bigg) = b_1 \qquad (14.151)\end{aligned}$$

and

$$\begin{aligned}\omega_2^2 \lambda_{max}(\mathbf{W}_4^T \mathbf{W}_4) &\leqslant \left(\frac{\hat{Q}_{1z}}{\hat{Q}_{2z}}\right)^2 \left(\frac{\hat{Q}_{2x}}{\chi_{2z}\hat{Q}_{1z}}\right)^2 \times ... \\ &\left(1 + \frac{(\hat{Q}_{2z})^2 - (\hat{Q}_{1z})^2}{(\hat{Q}_{1z} + \tilde{\lambda}_2)^2}\right) \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F})}{(\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F}))^2} \qquad (14.152) \\ &\leqslant \hat{Q}_{1z}\left(2\tilde{\lambda}_2 + \frac{\tilde{\lambda}_2^2}{\hat{Q}_{1z}} + \frac{(\hat{Q}_{2z})^2}{\hat{Q}_{1z}}\right) \times ... \\ &\qquad \left(\frac{\hat{Q}_{1z} + \hat{Q}_{2z}}{\hat{Q}_{2z}(\hat{Q}_{1z} + \tilde{\lambda}_2)}\right)^2 \lambda_{max}(\mathbf{F}^T\mathbf{F}). \qquad (14.153)\end{aligned}$$

For arbitrarily large $\hat{Q}_{1z}$, the quantity $\left(2\tilde{\lambda}_2 + \frac{\tilde{\lambda}_2^2}{\hat{Q}_{1z}} + \frac{(\hat{Q}_{2z})^2}{\hat{Q}_{1z}}\right)\left(\frac{\hat{Q}_{1z}+\hat{Q}_{2z}}{\hat{Q}_{2z}(\hat{Q}_{1z}+\tilde{\lambda}_2)}\right)^2 \lambda_{max}(\mathbf{F}^T\mathbf{F})$ is trivially bounded above whatever the value of $\tilde{\lambda}_2, \hat{Q}_{2z}$. Let b_2 be such an upper bound independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$. The sufficient condition for (14.147) to hold thus becomes:

$$\beta_1 > \beta_2 \hat{Q}_{1z} b_2 + \lambda_{max}(\mathbf{P}_2) b_1 \tag{14.154}$$

where b_1, b_2 are constants independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$.

We now turn to (14.148). A sufficient condition for it to hold is:

$$\begin{aligned}\beta_2 &> \lambda_{max}(\mathbf{P}_1) + \lambda_{max}(\mathbf{W}_2^T \mathbf{W}_2)\lambda_{max}(\mathbf{P}_2) \\ &+ \frac{(\lambda_{max}(\mathbf{P}_2))^2 \lambda_{max}(\mathbf{W}_2^T \mathbf{W}_2)\lambda_{max}(\mathbf{W}_1^T \mathbf{W}_1)}{\beta_1 - \beta_2 \omega_2^2 \lambda_{max}(\mathbf{W}_4^T \mathbf{W}_4) - \lambda_{max}(\mathbf{P}_2)\lambda_{max}(\mathbf{W}_1^T \mathbf{W}_1)} \qquad (14.155)\end{aligned}$$

Note that condition (14.147) ensures that the denominator in (14.155) is non-zero. We then have:

$$\lambda_{max}(\mathbf{W}_2^T\mathbf{W}_2) \leqslant \left(\frac{\hat{Q}_{2z}}{\chi_{2x}\hat{Q}_{1x}}\right)^2 \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F})}{(\hat{Q}_{2x}+\hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F}))^2} \tag{14.156}$$

$$\leqslant \left(\frac{\hat{Q}_{2z}(1+\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}})}{\hat{Q}_{1x}}\right)^2 \lambda_{max}(\mathbf{F}^T\mathbf{F}) \tag{14.157}$$

This quantity can be bounded above by a constant independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$ for arbitrarily large $\hat{Q}_{2x}$. Let b_3 be such a constant . Then a sufficient condition for condition (14.148) to hold is:

$$\beta_2 > \lambda_{max}(\mathbf{P}_1) + b_3\lambda_{max}(\mathbf{P}_2) + \frac{b_1b_3(\lambda_{max}(\mathbf{P}_2))^2}{\beta_1 - \beta_2\hat{Q}_{1z}b_2 - \lambda_{max}(\mathbf{P}_2)b_1} \tag{14.158}$$

we see that β_1 must scale linearly with $\hat{Q}_{1z}$ which is one of the parameters that is made arbitrarily large. Then β_1 also needs to become arbitrarily large for the conditions to hold. We choose $\beta_1 = 2\beta_2\hat{Q}_{1z}b_2 + \lambda_{max}(\mathbf{P}_2)b_1$ for the rest of the proof. Condition (14.154) is then verified, and β_2 needs to be chosen according to condition (14.158), which becomes:

$$\beta_2 > \lambda_{max}(\mathbf{P}_1) + b_3\lambda_{max}(\mathbf{P}_2) + \frac{b_1b_3\lambda_{max}^2(\mathbf{P}_2)}{\beta_2\hat{Q}_{1z}b_2} \tag{14.159}$$

This has a bounded solution for large values of $\hat{Q}_{1z}$. We now turn to the second part of (14.145).

Conditions for $\tau^2\mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T \succ 0$

We need to study the term $-\mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T$ (we study it with the $-$ sign since the middle matrix is negative definite from conditions (14.147,14.148) which are now verified). As we will see, because of the form of $\mathbf{H}_2$, we don't need to explicitly compute the whole inverse. Let $\mathbf{Z} = -(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)^{-1} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2 \\ \mathbf{Z}_2^T & \mathbf{Z}_3\end{bmatrix}$ ($\mathbf{Z}$ has the same block dimensions as $(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)$). We then have:

$$-\mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T = \mathbf{H}_2\mathbf{Z}\mathbf{H}_2^T \tag{14.160}$$

$$= \begin{bmatrix}\beta_2^2\omega_2^4\mathbf{W}_3^T\mathbf{W}_4\mathbf{Z}_3\mathbf{W}_4^T\mathbf{W}_3 & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N}\end{bmatrix}. \tag{14.161}$$

We thus only need to characterize the lower right block of $\mathbf{Z}$. It is easy to see that conditions (14.147) and (14.148) also enforce that both the Schur complements associated with the upper left and lower right blocks of $-(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)$ are invertible, thus giving the following form for $\mathbf{Z}_3$ using the block matrix inversion lemma [127]:

$$\mathbf{Z}_3 = (\beta_1\mathbf{I}_N - \beta_2\omega_2^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2\mathbf{K}_2\mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1)^{-1}. \tag{14.162}$$

where $\mathbf{K}_2 = (\beta_1 \mathbf{I}_M - \mathbf{P}_1 - \mathbf{W}_2^T \mathbf{P}_2 \mathbf{W}_2)^{-1}$. We thus have the following upper bound on the largest eigenvalue of $\mathbf{Z}_3$:

$$\lambda_{max}(\mathbf{Z}_3) \leqslant \frac{1}{\beta_1 - \beta_2 \hat{Q}_{1z} b_2 - \lambda_{max}(\mathbf{P}_2) b_1 - k}, \tag{14.163}$$

where $k = \frac{b_1 b_3 \lambda_{max}^2(\mathbf{P}_2)}{\beta_2 - \lambda_{max}(\mathbf{P}_1) - b_2 \lambda_{max}(\mathbf{P}_2)}$. Using the prescription $\beta_1 = 2\beta_2 \hat{Q}_{1z} b_2 + \lambda_{max}(\mathbf{P}_1) b_1$, we get:

$$\lambda_{max}(\mathbf{Z}_3) = \frac{1}{\beta_1 \hat{Q}_{1z} b_2 - \frac{b_1 b_3 \lambda_{max}^2(\mathbf{P}_2)}{\beta_1 - \lambda_{max}(\mathbf{P}_1) - b_2 \lambda_{max}(\mathbf{P}_2)}} \leqslant \frac{b_4}{\hat{Q}_{1z}} \tag{14.164}$$

where b_4 is a constant independent of the arbitrarily large parameters $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$. Thus $\lambda_{max}(\mathbf{Z}_3)$ can be made arbitrarily small by making λ_2 arbitrarily large.

We now want to find conditions for $\tau^2 \mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2 (\mathbf{B}^T \mathbf{P} \mathbf{B} + \mathbf{H}_3)^{-1} \mathbf{H}_2^T \succ 0$ which is equivalent to:

$$\begin{aligned} &\tau^2 \mathbf{P}_1 - \beta_2 \omega_2^2 \mathbf{W}_3^T \mathbf{W}_3 - \beta_2^2 \omega_2^4 \mathbf{W}_3^T \mathbf{W}_4 \mathbf{Z}_3 \mathbf{W}_4^T \mathbf{W}_3 \succeq 0 \\ &\tau^2 \mathbf{P}_2 - \beta_1 \omega_1^2 \mathbf{I}_N \succeq 0 \end{aligned} \tag{14.165}$$

We start with the upper matrix inequality, for which a sufficient condition is:

$$\begin{aligned} &\tau^2 \lambda_{min}(\mathbf{P}_1) - \beta_2 \omega_2^2 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \\ &- \beta_2^2 \omega_2^4 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \lambda_{max}(\mathbf{W}_4^T \mathbf{W}_4) \lambda_{max}(\mathbf{Z}_3) > 0 \end{aligned} \tag{14.166}$$

Using the bounds from appendix 14.7.3, we have:

$$\omega_2^2 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \leqslant ...$$

$$\left(\frac{\hat{Q}_{1z}}{\hat{Q}_{2z}}\right)^2 \left(1 + \frac{(\hat{Q}_{2z})^2 - (\hat{Q}_{1z})^2}{(\hat{Q}_{1z} + \tilde{\lambda}_2)^2}\right) \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \tag{14.167}$$

$$\leqslant \frac{2\tilde{\lambda}_2 \hat{Q}_{1z} + \tilde{\lambda}_2^2 + (\hat{Q}_{2z})^2}{(\hat{Q}_{1z} + \tilde{\lambda}_2)^2} \times ...$$

$$\max((1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{min}(\mathbf{F}^T \mathbf{F}))^2, (1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{max}(\mathbf{F}^T \mathbf{F}))^2) \tag{14.168}$$

$$\leqslant \frac{1}{\hat{Q}_{1z}} (2\tilde{\lambda}_2 + \frac{(\tilde{\lambda}_2^2 + (\hat{Q}_{2z})^2)}{\hat{Q}_{1z}}) \times ...$$

$$\max((1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{min}(\mathbf{F}^T \mathbf{F}))^2, (1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{max}(\mathbf{F}^T \mathbf{F}))^2) \tag{14.169}$$

Thus there exists a constant b_5, independent of $\lambda_2, \hat{Q}_{1z}, \hat{Q}_{2x}$ such that, for sufficiently large $\hat{Q}_{1z}$:

$$\omega_2^2 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \leqslant \frac{b_5}{\hat{Q}_{1z}}. \tag{14.170}$$

Remember that we had:

$$\omega_2^2 \lambda_{max}(\mathbf{W}_4^T \mathbf{W}_4) \leqslant \hat{Q}_{1z} b_2, \tag{14.171}$$

which gives the following sufficient condition for the upper left block in (14.165):

$$\tau^2 \lambda_{min}(\mathbf{P}_1) - \beta_2 \frac{b_5}{\hat{Q}_{1z}} - \beta_2^2 \frac{b_2 b_5 b_4}{\hat{Q}_{1z}} > 0. \tag{14.172}$$

A sufficient condition for the lower right block in (14.165) then reads:

$$\tau^2 \lambda_{min}(\mathbf{P}_2) - \beta_1 \omega_1^2 > 0, \tag{14.173}$$

where we have:

$$\begin{aligned}\beta_1 \omega_1^2 &= \left(\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\right)^2 \left(1 + \frac{(\hat{Q}_{2x})^2 - (\hat{Q}_{1x})^2}{(\hat{Q}_{1x} + \lambda_2)^2}\right) \times ... \\ &\quad (2\beta_1 \hat{Q}_{1z} b_2 + \lambda_{max}(\mathbf{P}_2) b_1) \qquad (14.174) \\ &= \frac{1}{\hat{Q}_{2x}} (\hat{Q}_{1x})^2 \left(1 + \frac{(\hat{Q}_{2x})^2 - (\hat{Q}_{1x})^2}{(\hat{Q}_{1x} + \lambda_2)^2}\right) \times ... \\ &\quad \left(2\beta_1 \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} b_2 + \lambda_{max}(\mathbf{P}_2) \frac{b_1}{\hat{Q}_{2x}}\right) \qquad (14.175)\end{aligned}$$

We remind the reader that $\hat{Q}_{1z}, \hat{Q}_{2x}$ grow linearly with λ_2. Thus the dominant scaling at large λ_2 is (exchanging $\hat{Q}_{2x}$ with $\hat{Q}_{1z}$ up to a constant):

$$\beta_1 \omega_1^2 \leqslant \frac{b_6}{\hat{Q}_{1z}}, \tag{14.176}$$

where b_6 is a constant independent of the arbitrarily large quantities. The final condition becomes:

$$\tau^2 \lambda_{min}(\mathbf{P}_1) - \beta_2 \frac{b_5}{\hat{Q}_{1z}} - \beta_2^2 \frac{b_2 b_5 b_4}{\hat{Q}_{1z}} > 0 \tag{14.177}$$

$$\tau^2 \lambda_{min}(\mathbf{P}_2) - \frac{b_6}{\hat{Q}_{1z}} > 0 \tag{14.178}$$

where we want $\tau < 1$. We now choose $\tau^2 = \tilde{c}/\hat{Q}_{1z}$ with a constant $\tilde{c}$ independent of $\lambda_2, \hat{Q}_{1z}, \hat{Q}_{2x}$ that verifies $\tilde{c} > \max\left(\frac{\beta_2 b_5 + \beta_2^2 b_2 b_5 b_4}{\lambda_{min}(\mathbf{P}_1)}, \frac{b_6}{\lambda_{min}(\mathbf{P}_2)}\right)$, such that:

$$\frac{\tilde{c}}{\hat{Q}_{1z}} \lambda_{min}(\mathbf{P}_1) - \beta_2 \frac{b_5}{\hat{Q}_{1z}} - \beta_2^2 \frac{b_2 b_5 b_4}{\hat{Q}_{1z}} > 0 \tag{14.179}$$

$$\frac{\tilde{c}}{\hat{Q}_{1z}} \lambda_{min}(\mathbf{P}_2) - \frac{b_6}{\hat{Q}_{1z}} > 0. \tag{14.180}$$

Since β_2 is bounded for large values of $\hat{Q}_{1z}$, and the b_i and c are constants independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$, we can then enforce $\tilde{c} < \hat{Q}_{1z}$ using the additional ridge penalty parametrized by λ_2 on the regularization to obtain $\tau < 1$ and a linear convergence rate proportional to $\sqrt{\frac{\tilde{c}}{\lambda_2}}$. We see that the eigenvalues of the matrix $\mathbf{P}$ are of little importance as long as they are non-vanishing. We choose $\mathbf{P}$ as the identity. In the statement of Lemma 54, we write c the exact constant which comes linking $\hat{Q}_{1z}$ to λ_2.
This proves Lemma 54.

14.8 Analytic continuation

In this section, we prove the validity of the analytic continuation and approximation argument used to prove Theorem 22, under the required set of assumptions 2. According to Lemma 4, for any

$\tilde{\lambda}_2 > 0$ and $\lambda_2 > \lambda_2^*$, any scalar pseudo-Lipschitz observable of order 2 ϕ, we have almost surely

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \phi(x_{0,i}, \hat{x}_i(\lambda_2)) = \mathbb{E}[\phi(x_0, \text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(H_x))] \tag{14.181}$$

where $H_x = \frac{\hat{m}_{1x}^* x_0 + \sqrt{\hat{\chi}_{1x}^*}\xi_{1x}}{\hat{Q}_{1x}}$ is defined in Theorem 22. We would like to show that this equality still holds for any $\lambda_2 > 0$. To do so we will show that, for a real analytic approximation of problem Eq.(13.2), both sides of Eq.(14.181) are real analytic in λ_2. We may then use the real analytic continuation theorem, as given in [148] to extend to any $\lambda_2 > 0$. We will treat the case $\lambda_2 = 0$ separately. In what follows, we will write the dependency in λ_2 of the estimator explicitly, i.e., $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\lambda_2)$.

14.8.1 Real analyticity of the left hand side of Eq.(14.181)

We remind a useful characterization of real analytic functions from [148]:

Proposition 13 (Proposition 1.2.10 from [148])**.** *Let $f \in \mathcal{C}^\infty(I)$ for some open interval I. The function f is in fact real analytic on I if and only if, for each $\alpha \in I$, there are an open interval J, with $\alpha \in J \subset I$, and finite constants $C > 0$ and $R > 0$ such that the derivatives of f satisfy :*

$$\left|f^{(j)}(\alpha)\right| \leqslant C\frac{j!}{R^j}, \quad \forall \alpha \in J \tag{14.182}$$

We also remind the formula for the higher order derivatives of a composition of two infinitely differentiable functions:

Proposition 14. *(Faa di Bruno's formula, [148] Theorem 1.3.2.) Consider two scalar functions f and g defined on an open interval $I \in \mathbb{R}$. Assume that both functions are infinitely differentiable on I and taking value in I. Then the derivatives of $h = g \circ f$ are given by*

$$h^{(n)}(t) = \sum \frac{n!}{k_1!k_2!...k_n!} g^{(k)}\left(f(t)\right)\left(\frac{f^{(1)}(t)}{1!}\right)^{k_1}\left(\frac{f^{(2)}(t)}{2!}\right)^{k_2}...\left(\frac{f^{(n)}(t)}{n!}\right)^{k_n} \tag{14.183}$$

where $k = k_1+k_2+...+k_n$ and the sum is taken over all $k_1, k_2, ..., k_n$ for which $k_1+2k_2+...+nk_n = n$.

The following lemma establishes bounds on the higher order derivatives of $\hat{\mathbf{x}}(\lambda_2)$ with respect to λ_2.

Lemma 59. *$\hat{\mathbf{x}}(\lambda_2)$ is infinitely differentiable w.r.t. λ_2 and, for any integer p, there exists a constant K' such that its elementwise p-th derivative, denoted $D_{\lambda_2}^{(p)}\hat{\mathbf{x}}(\lambda_2)$ verifies, almost surely*

$$\frac{1}{N}\left\|D_{\lambda_2}^{(p)}\hat{\mathbf{x}}(\lambda_2)\right\|_2^2 \leqslant K' \tag{14.184}$$

Furthermore, $D_{\lambda_2}^{(p)}\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of $\hat{\mathbf{x}}(\lambda_2)$.

Proof. Recall the strongly convex problem, for any finite N,

$$\hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2) = \arg\min_{\mathbf{x}\in\mathcal{X}} \tilde{g}(\mathbf{F}\mathbf{x}, \mathbf{y}) + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \tag{14.185}$$

where we absorbed $\tilde{\lambda}_2$ in $\tilde{g}$ as we are only interested in prolonging on λ_2.
The optimality condition then uniquely defines $\hat{\mathbf{x}}(\lambda_2)$ of each value of λ_2 and reads :

$$\mathbf{F}^\top \nabla \tilde{g}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2), \mathbf{y}) + \nabla f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2 \hat{\mathbf{x}}(\lambda_2) = 0 \tag{14.186}$$

The function $\mathbf{F}^\top \nabla \tilde{g}(\mathbf{F}\cdot, \mathbf{y}) + \nabla f(\cdot) + \lambda_2 \cdot$ is real analytic in $\mathbb{R}^N$ and its Jacobian $\mathbf{F}^\top \mathcal{H}_{\tilde{g}} \mathbf{F} + \mathcal{H}_f + \lambda_2 \mathbb{I}_N$ is non singular since f and $\tilde{g}$ are convex. The implicit function theorem [148] then ensures that, at any finite $N > 0$, the function $\hat{\mathbf{x}}(\lambda_2)$ is elementwise real analytic in λ_2. We can now prove the lemma with an induction.

Initialization Owing to assumption 2, we have almost surely

$$\lim_{N\to\infty} \frac{1}{N} \|\hat{\mathbf{x}}(\lambda_2)\|_2^2 \leqslant K' \tag{14.187}$$

and the identity is a Lipshchitz function of $\hat{\mathbf{x}}(\lambda_2)$ The function of λ_2 defined by :

$$\lambda_2 \mapsto \nabla \tilde{g}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2), \mathbf{y}) + \nabla f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2 \hat{\mathbf{x}}(\lambda_2) \tag{14.188}$$

is always zero valued from the definition of $\hat{\mathbf{x}}(\lambda_2)$, thus all its derivatives are zero. Taking the first derivative with respect to λ_2 yields:

$$\begin{aligned}(\mathbf{F}^T \mathcal{H}_{\tilde{g}}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2), \mathbf{y})\mathbf{F} + \mathcal{H}_f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2 \mathbf{I}_N) D\hat{\mathbf{x}}(\lambda_2) \\ + \hat{\mathbf{x}}(\lambda_2) = 0\end{aligned} \tag{14.189}$$

where D^p is the $(N \times 1)$ dimensional element-wise p-th differential of $\hat{\mathbf{x}}(\lambda_2)$. We then define the operator

$$\mathcal{O} : \begin{cases} \mathbb{R} \to \mathbb{R}^{N\times N} \\ \lambda_2 \mapsto \mathbf{F}^T \mathcal{H}_{\tilde{g}}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2), \mathbf{y})\mathbf{F} + \mathcal{H}_f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2 \mathbf{I}_N. \end{cases}$$

We obtain a simple expression for $D\hat{\mathbf{x}}(\lambda_2)$

$$D\hat{\mathbf{x}}(\lambda_2) = -\mathcal{O}^{-1}(\lambda_2)\hat{\mathbf{x}}(\lambda_2) \tag{14.190}$$

Since f and g are convex, the operator norm of $\mathcal{O}^{-1}(\lambda_2)$ is bounded with probability one, and $D\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of $\hat{\mathbf{x}}(\lambda_2)$ where $\frac{1}{N}\|D\hat{\mathbf{x}}(\lambda_2)\|_2^2$ is almost surely bounded.

Induction step Assume the property is verified up to $p-1$. For higher order derivatives, applying Leibniz's rule on Eq.(14.189) gives, denoting $\mathcal{O}^{(i)}(\lambda_2)$ the i-th derivative of $\mathcal{O}(\lambda_2)$, for the (p-1)-th derivative of (14.189) :

$$\sum_{i=0}^{p-1} \binom{p-1}{i} \mathcal{O}^{(i)}(\lambda_2) D^{(p-i)}\hat{\mathbf{x}}(\lambda_2) + D^{(p-1)}\hat{\mathbf{x}}(\lambda_2) = 0, \tag{14.191}$$

such that

$$\begin{aligned}\sum_{i=1}^{p-1} \binom{p-1}{i} \mathcal{O}^{(i)}(\lambda_2) D^{(p-i)}\hat{\mathbf{x}}(\lambda_2) + \mathcal{O}(\lambda_2) D^{(p)}\hat{\mathbf{x}}(\lambda_2) \\ + D^{(p-1)}\hat{\mathbf{x}}(\lambda_2) = 0\end{aligned} \tag{14.192}$$

We obtain the recursion on the differentials of $\hat{\mathbf{x}}(\lambda_2)$:

$$D^p\hat{\mathbf{x}}(\lambda_2) = -\mathcal{O}^{-1}(\lambda_2)\Bigg(\sum_{i=1}^{p-1}\binom{p-1}{i}\mathcal{O}^{(i)}(\lambda_2)D^{(p-i)}\hat{\mathbf{x}}(\lambda_2) + D^{(p-1)}\hat{\mathbf{x}}(\lambda_2)\Bigg). \tag{14.193}$$

where the matrix inverse $\mathcal{O}^{-1}(\lambda_2)$ is well defined for any $\lambda_2 > 0$ since f and g are convex. Using proposition 14, the assumption on the fast decay of the higher-order (larger than 2) derivatives of f and g, the bounded spectrum of the matrix $\mathbf{F}$, and the induction hypothesis, the operator norm of $\mathcal{O}^{(p)}(\lambda_2)$ is bounded with probability one for any $p \in \mathbb{N}$, $D^{(p)}\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of $\hat{\mathbf{x}}(\lambda_2)$ as a finite sum of Lipschitz functions of $\hat{\mathbf{x}}(\lambda_2)$, and its averaged squared norm is bounded almost surely. This concludes the induction. □

Lemma 60. *Under assumption 2, the function $\psi(\lambda_2)$ defined as*

$$\psi : \mathbb{R} \to \mathbb{R} \tag{14.194}$$

$$\lambda_2 \to \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi(x_{0,i}, \hat{x}_i(\lambda_2)) \tag{14.195}$$

is real analytic for $\lambda_2 > 0$.

Proof. Since ϕ is pseudo Lipschitz of order 2, there exists a constant C_ϕ such that, for any $x \in \mathbb{R}$, $\phi(x) \leqslant C_\phi(1 + x^2)$. Thus :

$$\lim_{N\to\infty}|\psi(\lambda_2)| \leqslant \lim_{N\to\infty}\frac{C_\phi}{N}(1 + \|\hat{\mathbf{x}}(\lambda_2)\|_2^2) \tag{14.196}$$

which is almost surely bounded. By assumption, the boundedness of ψ is enough to obtain its convergence. For the first derivative, the pseudo-Lipschitz property ensures that there exists a constant C'_ϕ such that, for any $x \in \mathbb{R}$, $\left|\frac{d\phi}{dx}(x)\right| \leqslant C'_\phi(1 + |x|)$. Then

$$\left|\frac{d}{d\lambda_2}\phi(\hat{x}(\lambda_2))\right| \leqslant C'_\phi\left|\frac{d}{d\lambda_2}\hat{x}(\lambda_2)\right|(1 + |\hat{x}|(\lambda_2)) \tag{14.197}$$

so there exists a constant C'_ψ such that

$$\lim_{N\to\infty} D\psi(\lambda_2) \leqslant \lim_{N\to\infty}\frac{1}{N}C'_\psi\left(\|D\hat{\mathbf{x}}(\lambda_2)\|_2 + \|D\hat{\mathbf{x}}(\lambda_2)\|_2\|\hat{\mathbf{x}}(\lambda_2)\|_2\right) \tag{14.198}$$

which is almost surely bounded. We have also proved in the previous lemma that $D\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of λ_2, thus $D\psi(\lambda_2)$ is a PL2 function of $\hat{\mathbf{x}}(\lambda_2)$ and its limit exists according to Assumption 2 (c). For the higher order derivatives, we use proposition 14 to obtain, for any coordinate $1 \leqslant i \leqslant n$:

$$\left|\frac{d^{(p)}}{d\lambda_2^{(p)}}\phi(\hat{x}_i(\lambda_2))\right| = \sum\frac{p!}{k_1!k_2!...k_p!}\phi^{(k)}\left(\hat{x}_i(\lambda_2)\right)\left(\frac{\hat{x}_i^{(1)}(\lambda_2)}{1!}\right)^{k_1}\left(\frac{\hat{x}_i^{(2)}(\lambda_2)}{2!}\right)^{k_2}...\left(\frac{\hat{x}_i^{(p)}(\lambda_2)}{p!}\right)^{k_p}$$

The assumption on the higher order derivatives of ϕ from Theorem 22 and Lemma 59 implies that the term
$\phi^{(k)}\left(\hat{x}_i(\lambda_2)\right)\left(\frac{\hat{x}_i^{(1)}(\lambda_2)}{1!}\right)^{k_1}\left(\frac{\hat{x}_i^{(2)}(\lambda_2)}{2!}\right)^{k_2}...\left(\frac{\hat{x}_i^{(p)}(\lambda_2)}{p!}\right)^{k_p}$ has bounded absolute value with probability one, for all coordinates i. Using the characterization of real analytic functions and assumption 2 (c) from proposition 13, this concludes the proof. □

14.8.2 Analytic continuation to $(\tilde{\lambda}_2, \lambda_2) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$

From assumption 2, the set of fixed point equations from Theorem 22 admit a unique solution for any $\lambda_2, \tilde{\lambda}_2$. Additionally, the implicit function theorem [148] can also be applied to the set of fixed point equations from Theorem 22 regarding the dependencies in $\lambda_2, \tilde{\lambda}_2$ to show that each quantity involved is real analytic in $\lambda_2, \tilde{\lambda}_2$. At this point, we have two analytic functions, the observable and the one defined by the fixed point of the state evolution equations, that coincide for any $\lambda_2 \in [\lambda_2^*, +\infty[$ and any $\tilde{\lambda}_2 > 0$. We can now use the analytic continuation theorem [148] to show that these functions remain equal for any $\lambda_2 > 0$ and for $\tilde{\lambda}_2 > 0$. This concludes the proof of Lemma 56.

14.8.3 Real analytic approximation of strongly convex problems

Consider

$$\hat{\mathbf{x}}_\epsilon(\lambda_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ \tilde{g}_\epsilon(\mathbf{F}\mathbf{x}, \mathbf{y}) + f_\epsilon(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \tag{14.199}$$

$$\hat{\mathbf{x}}(\lambda_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ \tilde{g}(\mathbf{F}\mathbf{x}, \mathbf{y}) + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \tag{14.200}$$

where g_ϵ, f_ϵ are real analytic approximations of the loss g and regularizer f verifying assumption 2(e). To relax the analytic approximation, we need to prove the following equality.

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi(\hat{x}_{\epsilon,i}(\lambda_2)) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi(\hat{x}_i(\lambda_2)) \tag{14.201}$$

Under assumption 2 (c) and owing to the definition of PL2 functions, it is sufficient to prove

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 = 0 \tag{14.202}$$

Denote $\mathcal{C}$ the cost function $\tilde{g}(\mathbf{F}., \mathbf{y}) + f(.)$ and its real analytic counterpart $\mathcal{C}_\epsilon$ the cost function $\tilde{g}_\epsilon(\mathbf{F}., \mathbf{y}) + f_\epsilon(.)$.

$$\forall \mathbf{x} \in \mathbb{R}^d \quad \lim_{\epsilon\to 0}\mathcal{C}_\epsilon(\mathbf{x}) = \mathcal{C}(\mathbf{x}) \tag{14.203}$$

Since minimizers of convex functions are fixed points of the corresponding proximity operators, it holds that

$$\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 = \frac{1}{N}\left\|\text{prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}_\epsilon(\lambda_2)) - \text{prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\right\|_2^2 \tag{14.204}$$

$$\leqslant \frac{1}{N}\left\|\text{prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}_\epsilon(\lambda_2)) - \text{prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\right\|_2^2 + \frac{1}{N}\left\|\text{prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2)) - \text{prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\right\|_2^2 \tag{14.205}$$

The results from appendix 14.7.3 show that proximity operators of strongly convex functions are contractions, thus their exists a positive constant $L_{\lambda_2} < 1$ such that for any realisation of $\mathbf{F}, \mathbf{x}^0, \boldsymbol{\omega}_0$

$$\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 \leqslant \frac{1}{N}L_{\lambda_2}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 + \frac{1}{N}\left\|\text{prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2)) - \text{prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\right\|_2^2 \tag{14.206}$$

Furthermore, the function $\text{prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(.)$ converges uniformly to $\text{prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(.)$ when $\epsilon \to 0$, and thus

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\left\|\text{prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))-\text{prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\right\|_2^2=0 \tag{14.207}$$

which gives

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2)-\hat{\mathbf{x}}(\lambda_2)\|_2^2 \leqslant L_{\lambda_2}\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2)-\hat{\mathbf{x}}(\lambda_2)\|_2^2. \tag{14.208}$$

Since $L_{\lambda_2}<1$, this implies

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2)-\hat{\mathbf{x}}(\lambda_2)\|_2^2=0 \tag{14.209}$$

14.8.4 Continuous extension to $\tilde{\lambda}_2 = 0$

Making the dependence on $\tilde{\lambda}_2$ explicit, define

$$\hat{\mathbf{x}}(\tilde{\lambda}_2,\lambda_2)=\underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ g(\mathbf{F}\mathbf{x},\mathbf{y})+f(\mathbf{x})+\frac{\lambda_2}{2}\|\mathbf{x}\|_2^2+\frac{\tilde{\lambda}_2}{2}\|\mathbf{F}\mathbf{x}\|_2^2 \tag{14.210}$$

$$\hat{\mathbf{x}}(0,\lambda_2)=\underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ g(\mathbf{F}\mathbf{x},\mathbf{y})+f(\mathbf{x})+\frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \tag{14.211}$$

Both cost functions defining $\hat{\mathbf{x}}(\tilde{\lambda}_2,\lambda_2),\hat{\mathbf{x}}(0,\lambda_2)$ are strongly convex for any $\lambda_2>0$. We can then use the same argument as in the previous subsection C to conclude

$$\lim_{\tilde{\lambda}_2\to 0}\lim_{N\to\infty}\frac{1}{N}\left\|\hat{\mathbf{x}}(\tilde{\lambda}_2,\lambda_2)-\hat{\mathbf{x}}(0,\lambda_2)\right\|_2^2=0 \tag{14.212}$$

14.8.5 Continuous extension to $\lambda_2 = 0$

For $\tilde{\lambda}_2=0$, the estimator $\hat{\mathbf{x}}(\lambda_2)$ is still unique for any $\lambda_2>0$. We now need to study the limiting ridgeless estimator

$$\lim_{\lambda_2\to 0}\underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\, g(\mathbf{F}\mathbf{x},\mathbf{y})+f(\mathbf{x})+\frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \tag{14.213}$$

for functions f,g that may not be strictly convex. To do so we will use Theorem 26.20 from [25], which is reminded in appendix 14.2, proposition 12. Under assumption 2 and since the l_2 norm is strongly convex thus uniformly convex, we have, denoting $\hat{\mathbf{x}}_0$ the unique least l_2 norm element in $\arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{F}\mathbf{x},\mathbf{y})+f(\mathbf{x})$,

$$\lim_{\lambda_2\to 0}\hat{\mathbf{x}}(\lambda_2)=\hat{\mathbf{x}}_0 \tag{14.214}$$

We can therefore uniquely define the continuous extension of any continuous observable ϕ of $\hat{\mathbf{x}}(\lambda_2)$ such that $\phi(\lambda_2=0)=\phi(\hat{\mathbf{x}}_0)$. Then this observable and the corresponding function implicitly defined by the set of fixed point equations are continuous on $[0,+\infty[$ and equal for any $\lambda_2\in]0,+\infty[$, and thus also equal at $\lambda_2=0$ using the definition of continuity and the fact that $]0,+\infty[$ is dense in $[0,+\infty[$.

14.8.6 Real analytic approximation of usual cost functions with fast decaying higher-order derivatives

In this section, we show that any combination of the square, logistic and hinge loss with ℓ_1 or ℓ_2 verifies Assumption 2 (e), i.e. they can be approximated with real analytic functions whose second derivatives have higher-order derivatives that decrease faster than any polynomial. The square loss and ℓ_2 immediately verify these assumptions. Assuming $y = 1$ without loss of generality, the second derivative of the logistic loss is given by

$$g''(x) = \frac{\exp(x)}{(1 + \exp(x))}. \tag{14.215}$$

All higher order derivatives will be a polynomial in $\exp(x)$ divided by a higher order polynomial in $\exp(x)$ plus one. Thus, for any sign of x, higher-order derivatives of the logistic loss will decrease exponentially fast when the absolute value of x goes to infinity. We now turn to the ℓ_1 penalty. Real analytic approximations of functions may be constructed by considering their convolution with a Gaussian kernel, which is also known as the Weierstrass transform. Denoting $\mathcal{W}_\epsilon[f]$ the Weierstrass transform of a function f with parameter $\epsilon > 0$, we obtain for the ℓ_1 penalty

$$\mathcal{W}_\epsilon[|.|](x) = \frac{1}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{+\infty} |u| \exp\left(-\frac{1}{2\epsilon}(u-x)^2\right) du \tag{14.216}$$

$$= \frac{1}{\sqrt{2\pi\epsilon}} \left(2\epsilon \exp\left(-\frac{1}{2\epsilon}x^2\right) + 2x \int_0^x \exp\left(-\frac{1}{2\epsilon}u^2\right) du\right) \tag{14.217}$$

whose second derivative reads

$$\frac{d^2}{dx^2}\mathcal{W}_\epsilon[|.|](x) = \frac{\sqrt{2}}{\sqrt{\pi\epsilon}} \exp\left(-\frac{1}{2\epsilon}x^2\right) \tag{14.218}$$

thus $\mathcal{W}_\epsilon[|.|)]$ is strongly convex and its higher order derivatives all decay faster than any finite order polynomial. A similar computation shows that, for the hinge loss,

$$\mathcal{W}_\epsilon[\max(0, 1-.)](x) = \frac{1}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{+\infty} \max(0, 1-u) \exp\left(-\frac{1}{2\epsilon}(u-x)^2\right) du$$

$$= \frac{1}{\sqrt{2\pi\epsilon}} \left((1-x)\sqrt{\frac{\pi\epsilon}{2}} + \epsilon \exp\left(-\frac{1}{2\epsilon}(1-x)^2\right)\right. \tag{14.219}$$

$$\left. + (1-x) \int_0^x \exp\left(-\frac{1}{2\epsilon}(1-x)^2\right) du\right) \tag{14.220}$$

whose second derivative reads

$$\frac{d^2}{dx^2}\mathcal{W}_\epsilon[\max(0, 1-.)](x) = \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{1}{2\epsilon}(1-x)^2\right) \tag{14.221}$$

Thus the hinge loss and ℓ_1 penalty verify Assumption 2 (e).

Part IV

Future directions and bibliography

14.9 Future directions

Universality and finite size rate analysis As mentioned in Chapters 2, 9, 11, state evolution proofs are amenable to both universality proofs [27, 62] and finite-size rates analysis [251, 250]. We therefore expect all our results to hold when the design matrix has independently (but not necessary identically) distributed subGaussian entries. We also expect that all the asymptotic statements given for square dominated observables to present exponentially decreasing rates in the problem dimensions, as proven in [250] or [204]. Such rates are prized in the statistics community to perform hypothesis testing and confidence interval computations.

Further realistic models We have shown that exactly solvable models that capture realistic learning curves can be defined by using Gaussian mixtures for the data, and block covariate models for the features. Exploring further results in Gaussian equivalence, as was done in [116, 128, 261, 209], notably for multilayer models, is a proximsing avenue of research to better describe feature maps. For data models, any distribution can be approximated by a Gaussian mixture, provided enough centroids are considered. One of the main limitations of our results is that this number of centroids should remain finite, while Gaussian kernel dennsity estiamtors [293] would systematically lead to an extensive number of order parameters, since we a priori don't know the tail behaviour of realistic data. It is thus interesting to pursue the deisgn of models that may capture geometrical properties of probability distributions in ways that are more appropriate than correlated Gaussian mixtures, and still give exactly solvable models. The problem of dealing with order parameters of extensive sizes leads us to similar issues as the recently investigated matrix factorization with extensive rank [19, 183], which is the subject of the next paragraph. We note that the rigorous tools developed in Chapter 2 allow to obtain Bayes-optimal recovery guarantees for multilayer networks with dense or convolutional random matrices, which could be combined with a form of convex regression to model learning of the last layer of a multilayer feature map with random weights.

Extensive rank problems Throughout this thesis, all estimators were low-rank matrices, in that a finite number of vectors of extensive dimensions were considered to be learned. We have seen that the convex Gaussian comparison inequalities are only interesting for vector-valued estimators, while AMP proofs work for matrix-valued estimators with low rank with respect to the extensive problem dimension d. Indeed, the Gaussian iterative conditoning scheme relies strongly on the fact that projectors are low-rank, which simplifies error terms as shown in the introduction, section 1.7 and lemma 21 from chapter 3. Equivalent of lemma 2 decomposing random variables into independent ones are not known in random matrix theory, although a rich litterature now exists for equivalents of non-linear transforms of products of random matrices with extensive ranks, see e.g.[92, 231, 175, 96, 36]. An interesting example is the note of Sandrine Péché [228] which proposes a decomposition of such non-linear transforms into linear information plus independent noise matrices. Unfortunately, all those approaches are based on linearisation arguments, which is not the case for lemma 2 and the related proofs. We note that recent results using the replica method have led to solutions of matrix factorization problems with extensive ranks, where the order parameters are spectral densities [19, 183], which suggests corresponding mathematical tools could be designed. Learning in neural networks also requires extensive rank asymptotic tools, since the matrices of a generic model of deep networks going beyond the committee machine [15] or our ensembling models of Chapter 11 immediately lead to extensive rank weight matrices.

Beyond convexity We have seen that convexity is crucial and quite convenient to study the exact asymptotics of landscapes : it enables to characterize in an intuitive, stable, and algorithmically reachable way the solution of optimization problems, using Moreau envelopes and proximal operators. It would be interesting to attempt to study more complex landscapes by finding equivalents of the proximity operators for multi-convex functions [122], where functions are assumed to be block-convex in their arguments, i.e. convex in one set of variables when the other variables are fixed. A possible idea would be to consider the corresponding block-proximal operators, which would solve subproblems defining metastable states for one set of variable at a time while the others are fixed. We note that defining simpler landscapes by freezing variables and/or order parameters is reminescent of Franz-Parisi potentials [99].

Bibliography

[1] B. Adlam and J. Pennington, *The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization*, in International Conference on Machine Learning, PMLR, 2020, pp. 74–84.

[2] ——, *Understanding double descent requires a fine-grained bias-variance decomposition*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 11022–11032.

[3] M. Advani and S. Ganguli, *An equivalence between high dimensional bayes optimal inference and m-estimation*, in Advances in Neural Information Processing Systems, 2016, pp. 3378–3386.

[4] M. S. Advani and A. M. Saxe, *High-dimensional dynamics of generalization error in neural networks*, 2017.

[5] E. Agoritsas, G. Biroli, P. Urbani, and F. Zamponi, *Out-of-equilibrium dynamical mean-field equations for the perceptron model*, Journal of Physics A: Mathematical and Theoretical, 51 (2018), p. 085002.

[6] P. K. Andersen and R. D. Gill, *Cox's regression model for counting processes: a large sample study*, The annals of statistics, (1982), pp. 1100–1120.

[7] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Cambridge university press, 2010.

[8] P. W. Anderson, *Spin glass i: A scaling law rescued*, Physics Today, 41 (1988), pp. 9–11.

[9] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, et al., *Kymatio: Scattering transforms in python.*, Journal of Machine Learning Research, 21 (2020), pp. 1–6.

[10] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American mathematical society, 68 (1950), pp. 337–404.

[11] G. B. Arous, A. Dembo, and A. Guionnet, *Aging of spherical spin glasses*, Probability theory and related fields, 120 (2001), pp. 1–67.

[12] B. Aubin, F. Krzakala, Y. Lu, and L. Zdeborová, *Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization*, Advances in Neural Information Processing Systems, 33 (2020).

[13] B. Aubin, B. Loureiro, A. Maillard, F. Krzakala, and L. Zdeborová, *The spiked matrix model with generative priors*, Advances in Neural Information Processing Systems, 32 (2019).

[14] ——, *The spiked matrix model with generative priors*, IEEE Transactions on Information Theory, (2020).

[15] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová, *The committee machine: Computational to statistical gaps in learning a two-layers neural network*, Journal of Statistical Mechanics: Theory and Experiment, 2019 (2019), p. 124023.

[16] F. Bach, *Breaking the curse of dimensionality with convex neural networks*, The Journal of Machine Learning Research, 18 (2017), pp. 629–681.

[17] Z. Bai and W. Zhou, *Large sample covariance matrices without independence structures in columns*, Statistica Sinica, (2008), pp. 425–442.

[18] J. Barbier and N. Macris, *The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference*, Probability theory and related fields, 174 (2019), pp. 1133–1185.

[19] ——, *Statistical limits of dictionary learning: random matrix theory and the spectral replica method*, arXiv preprint arXiv:2109.06610, (2021).

[20] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information theory, 39 (1993), pp. 930–945.

[21] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, *Benign overfitting in linear regression*, Proceedings of the National Academy of Sciences, 117 (2020), pp. 30063–30070.

[22] P. L. Bartlett and S. Mendelson, *Rademacher and gaussian complexities: Risk bounds and structural results*, Journal of Machine Learning Research, 3 (2002), pp. 463–482.

[23] H. Bauschke, P. Combettes, and D. Noll, *Joint minimization with alternating bregman proximity operators*, Pacific Journal of Optimization, (2006).

[24] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Bregman monotone optimization algorithms*, SIAM Journal on control and optimization, 42 (2003), pp. 596–636.

[25] H. H. Bauschke, P. L. Combettes, et al., *Convex analysis and monotone operator theory in Hilbert spaces*, vol. 408, Springer, 2011.

[26] H. H. Bauschke, M. N. Dao, and S. B. Lindstrom, *Regularizing with bregman–moreau envelopes*, SIAM Journal on Optimization, 28 (2018), pp. 3208–3228.

[27] M. Bayati, M. Lelarge, and A. Montanari, *Universality in polytope phase transitions and message passing algorithms*, Annals of Applied Probability, 25 (2015), pp. 753–822.

[28] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Transactions on Information Theory, 57 (2011), pp. 764–785.

[29] ——, *The lasso risk for gaussian matrices*, IEEE Transactions on Information Theory, 58 (2011), pp. 1997–2017.

[30] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 15849–15854.

[31] ——, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 15849–15854.

[32] ——, *Reply to loog et al.: Looking beyond the peaking phenomenon*, Proceedings of the National Academy of Sciences, 117 (2020), pp. 10627–10627.

[33] M. Belkin, D. Hsu, and J. Xu, *Two models of double descent for weak features*, SIAM Journal on Mathematics of Data Science, 2 (2020), pp. 1167–1180.

[34] M. Belkin, S. Ma, and S. Mandal, *To understand deep learning we need to understand kernel learning*, in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, eds., vol. 80 of Proceedings of Machine Learning Research, PMLR, 10–15 Jul 2018, pp. 541–549.

[35] G. Ben Arous, A. Dembo, and A. Guionnet, *Cugliandolo-kurchan equations for dynamics of spin-glasses*, Probability theory and related fields, 136 (2006), pp. 619–660.

[36] L. Benigni and S. Péché, *Eigenvalue distribution of some nonlinear models of random matrices*, Electronic Journal of Probability, 26 (2021), pp. 1–37.

[37] R. Berthier, A. Montanari, and P.-M. Nguyen, *State evolution for approximate message passing with non-separable functions*, Information and Inference: A Journal of the IMA, 9 (2020), pp. 33–79.

[38] H. A. Bethe, *Statistical theory of superlattices*, Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences, 150 (1935), pp. 552–575.

[39] T. P. Biehl, Michael; Caticha, Nestor; Opper, Manfred; Villmann, *Statistical Physics of Learning and Generalization*, Adaptivity and Learning, (2003), pp. 77–88.

[40] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.

[41] E. Bolthausen, *On the high-temperature phase of the sherrington-kirkpatrick model*, Sep 2009.

[42] ——, *An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model*, Communications in Mathematical Physics, 325 (2014), pp. 333–366.

[43] ——, *The thouless-anderson-palmer equation in spin glass theory*, https://anr-malin.sciencesconf.org/data/pages/Aussois_2.pdf, (2019).

[44] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, *Compressed sensing using generative models*, in International Conference on Machine Learning, PMLR, 2017, pp. 537–546.

[45] B. Bordelon, A. Canatar, and C. Pehlevan, *Spectrum dependent learning curves in kernel regression and wide neural networks*, in International Conference on Machine Learning, PMLR, 2020, pp. 1024–1034.

[46] L. Bottou, *Stochastic gradient descent tricks*, in Neural networks: Tricks of the trade, Springer, 2012, pp. 421–436.

[47] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013.

[48] O. Bousquet, S. Boucheron, and G. Lugosi, *Introduction to statistical learning theory*, in Summer school on machine learning, Springer, 2003, pp. 169–207.

[49] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[50] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends® in Machine learning, 3 (2011), pp. 1–122.

[51] L. Breiman, *Bagging predictors*, Machine Learning, 24 (1996), pp. 123–140.

[52] J. Bruna and S. Mallat, *Invariant scattering convolution networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 1872–1886.

[53] E. J. Candès, P. Sur, et al., *The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression*, The Annals of Statistics, 48 (2020), pp. 27–42.

[54] Y. Cao, Q. Gu, and M. Belkin, *Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures*, Preprint arXiv:2104.13628, (2021).

[55] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Foundations of Computational Mathematics, 7 (2007), pp. 331–368.

[56] M. Celentano, C. Cheng, and A. Montanari, *The high-dimensional asymptotics of first order methods with random data*, arXiv preprint arXiv:2112.07572, (2021).

[57] M. Celentano, A. Montanari, and Y. Wei, *The lasso with general gaussian designs with applications to hypothesis testing*, arXiv preprint arXiv:2007.13716, (2020).

[58] K. A. Chandrasekher, A. Pananjady, and C. Thrampoulidis, *Sharp global convergence guarantees for iterative nonconvex optimization: A gaussian process perspective*, arXiv preprint arXiv:2109.09859, (2021).

[59] N. S. Chatterji and P. M. Long, *Finite-sample analysis of interpolating linear classifiers in the overparameterized regime*, Preprint arXiv:2004.12019, (2021).

[60] L. Chen, Y. Min, M. Belkin, and A. Karbasi, *Multiple descent: Design your own generalization curve*, arXiv preprint arXiv:2008.01036, (2020).

[61] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61.

[62] W.-K. Chen and W.-K. Lam, *Universality of approximate message passing algorithms*, Electronic Journal of Probability, 26 (2021), pp. 1–44.

[63] X. Cheng and A. Singer, *The spectrum of random inner-product kernel matrices*, Random Matrices: Theory and Applications, 2 (2013), p. 1350010.

[64] L. Chizat, E. Oyallon, and F. Bach, *On lazy training in differentiable programming*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[65] E. Cornacchia, F. Mignacco, R. Veiga, C. Gerbelot, B. Loureiro, and L. Zdeborová, *Learning curves for the multi-class teacher-student perceptron*, arXiv preprint, (2022).

[66] T. M. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE transactions on electronic computers, EC-14 (1965), pp. 326–334.

[67] A. Crisanti, H. Horner, and H.-J. Sommers, *The sphericalp-spin interaction spin-glass model*, Zeitschrift für Physik B Condensed Matter, 92 (1993), pp. 257–271.

[68] L. F. Cugliandolo and J. Kurchan, *Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model*, Physical Review Letters, 71 (1993), p. 173.

[69] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of control, signals and systems, 2 (1989), pp. 303–314.

[70] M. Daniels, C. Gerbelot, F. Krzakala, and L. Zdeborová, *Multi-layer state evolution under random convolutional design*, arXiv preprint arXiv:2205.13503, (2022).

[71] S. D'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala, *Double trouble in double descent: Bias and variance(s) in the lazy regime*, in Proceedings of the 37th International Conference on Machine Learning, H. Daumé III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18 Jul 2020, pp. 2280–2290.

[72] S. d'Ascoli, L. Sagun, and G. Biroli, *Triple descent and the two kinds of overfitting: where and why do they appear?*, Journal of Statistical Mechanics: Theory and Experiment, 2021 (2021), p. 124002.

[73] Z. Deng, A. Kammoun, and C. Thrampoulidis, *A model of double descent for high-dimensional binary linear classification*, Preprint arXiv:1911.05822, (2020).

[74] Y. Deshpande, E. Abbe, and A. Montanari, *Asymptotic mutual information for the balanced binary stochastic block model*, Information and Inference: A Journal of the IMA, 6 (2017), pp. 125–170.

[75] Y. Deshpande and A. Montanari, *Information-theoretically optimal sparse pca*, in 2014 IEEE International Symposium on Information Theory, IEEE, 2014, pp. 2197–2201.

[76] O. Dhifallah and Y. M. Lu, *A precise performance analysis of learning with random features*, arXiv preprint arXiv:2008.11904, (2020).

[77] R. Dietrich, M. Opper, and H. Sompolinsky, *Statistical mechanics of support vector networks*, Phys. Rev. Lett., 82 (1999), pp. 2975–2978.

[78] E. Dobriban, S. Wager, et al., *High-dimensional asymptotics of prediction: Ridge regression and classification*, The Annals of Statistics, 46 (2018), pp. 247–279.

[79] D. Donoho, *For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 59 (2006), pp. 797–829.

[80] D. Donoho, A. Javanmard, and A. Montanari, *Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing*, IEEE transactions on information theory, 59 (2013), pp. 7434–7464.

[81] D. Donoho and J. Jin, *Higher criticism thresholding: Optimal feature selection when useful features are rare and weak*, Proceedings of the National Academy of Sciences, 105 (2008), pp. 14790–14795.

[82] D. Donoho and A. Montanari, *High dimensional robust m-estimation: Asymptotic variance via approximate message passing*, Probability Theory and Related Fields, 166 (2016), pp. 935–969.

[83] D. L. Donoho, A. Maleki, and A. Montanari, *Message-passing algorithms for compressed sensing*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 18914–18919.

[84] ——, *Message passing algorithms for compressed sensing: I. motivation and construction*, in 2010 IEEE information theory workshop on information theory (ITW 2010, Cairo), IEEE, 2010, pp. 1–5.

[85] ——, *Message passing algorithms for compressed sensing: Ii. analysis and validation*, in 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo), IEEE, 2010, pp. 1–5.

[86] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, *Boosting and other ensemble methods*, Neural computation, 6 (1994), pp. 1289–1301.

[87] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization.*, Journal of machine learning research, 12 (2011).

[88] R. Durrett, *Probability: theory and examples*, vol. 49, Cambridge university press, 2019.

[89] N. El Karoui, *On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators*, Probability Theory and Related Fields, 170 (2018), pp. 95–175.

[90] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, *On robust regression with high-dimensional predictors*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 14557–14562.

[91] N. El Karoui et al., *Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond*, Annals of Applied Probability, 19 (2009), pp. 2362–2405.

[92] ——, *The spectrum of kernel random matrices*, Annals of statistics, 38 (2010), pp. 1–50.

[93] M. Emami, M. Sahraee-Ardakan, P. Pandit, S. Rangan, and A. Fletcher, *Generalization error of generalized linear models in high dimensions*, in International Conference on Machine Learning, PMLR, 2020, pp. 2892–2901.

[94] A. Engel and C. Van den Broeck, *Statistical mechanics of learning*, Cambridge University Press, 2001.

[95] Z. Fan, *Approximate message passing algorithms for rotationally invariant matrices*, arXiv preprint arXiv:2008.11892, (2020).

[96] Z. Fan and A. Montanari, *The spectral norm of random inner-product kernel matrices*, Probability Theory and Related Fields, 173 (2019), pp. 27–85.

[97] A. Fletcher, S. Rangan, and P. Schniter, *Inference in deep networks in high dimensions*, in 2018 IEEE International Symposium on Information Theory (ISIT), IEEE, 2018, pp. 1884–1888.

[98] A. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, *Expectation consistent approximate inference: Generalizations and convergence*, in 2016 IEEE International Symposium on Information Theory (ISIT), IEEE, 2016, pp. 190–194.

[99] S. Franz and G. Parisi, *Recipes for metastable states in spin glasses*, Journal de Physique I, 5 (1995), pp. 1401–1415.

[100] R. Gallager, *Low-density parity-check codes*, IRE Transactions on information theory, 8 (1962), pp. 21–28.

[101] S. Ganguli and H. Sompolinsky, *Statistical mechanics of compressed sensing*, Physical review letters, 104 (2010), p. 188701.

[102] E. Gardner, *The space of interactions in neural network models*, Journal of physics A: Mathematical and general, 21 (1988), p. 257.

[103] E. Gardner and B. Derrida, *Three unfinished works on the optimal storage capacity of networks*, Journal of Physics A: Mathematical and General, 22 (1989), p. 1983.

[104] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart, *Scaling description of generalization with number of parameters in deep learning*, Journal of Statistical Mechanics: Theory and Experiment, 2020 (2020), p. 023401.

[105] S. Geman, E. Bienenstock, and R. Doursat, *Neural networks and the bias/variance dilemma*, Neural computation, 4 (1992), pp. 1–58.

[106] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, *Generalisation error in learning with random features and the hidden manifold model*, in 37th International Conference on Machine Learning, 2020.

[107] F. Gerace, B. Loureiro, F. Krzakala, M. Mezard, and L. Zdeborova, *Generalisation error in learning with random features and the hidden manifold model*, in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18 Jul 2020, pp. 3452–3462.

[108] C. Gerbelot, A. Abbara, and F. Krzakala, *Asymptotic errors for high-dimensional convex penalized linear regression beyond gaussian matrices*, in Conference on Learning Theory, PMLR, 2020, pp. 1682–1713.

[109] C. Gerbelot, A. Abbara, and F. Krzakala, *Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula)*, arXiv preprint arXiv:2006.06581, (2020).

[110] C. Gerbelot and R. Berthier, *Graph-based approximate message passing iterations*, arXiv preprint arXiv:2109.11905, (2021).

[111] C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborova, *Rigorous dynamical mean field theory for stochastic gradient descent methods*, arXiv preprint arXiv:2210.06591, (2022).

[112] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, *Limitations of lazy training of two-layers neural network*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, 2019, pp. 9111–9121.

[113] ——, *When do neural networks outperform kernel methods?*, in Advances in Neural Information Processing Systems, vol. 33, 2020.

[114] P. Giselsson and S. Boyd, *Linear convergence and metric selection for douglas-rachford splitting and admm*, IEEE Transactions on Automatic Control, 62 (2016), pp. 532–544.

[115] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, *The gaussian equivalence of generative models for learning with shallow neural networks*, Proceedings of Machine Learning Research, 145 (2021), pp. 1–46.

[116] S. Goldt, B. Loureiro, G. Reeves, M. Mézard, F. Krzakala, and L. Zdeborová, *The gaussian equivalence of generative models for learning with two-layer neural networks*, in Mathematical and Scientific Machine Learning, 2021.

[117] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, *Modeling the influence of data structure on learning in neural networks: The hidden manifold model*, Phys. Rev. X, 10 (2020), p. 041044.

[118] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, *Modeling the influence of data structure on learning in neural networks: The hidden manifold model*, Physical Review X, 10 (2020), p. 041044.

[119] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT press, 2016.

[120] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial networks*, 2014.

[121] Y. Gordon, *Some inequalities for gaussian processes and applications*, Israel Journal of Mathematics, 50 (1985), pp. 265–289.

[122] J. Gorski, F. Pfeuffer, and K. Klamroth, *Biconvex sets and optimization with biconvex functions: a survey and extensions*, Mathematical methods of operations research, 66 (2007), pp. 373–407.

[123] W. Hachem, P. Loubaton, J. Najim, et al., *Deterministic equivalents for certain functionals of large random matrices*, The Annals of Applied Probability, 17 (2007), pp. 875–930.

[124] L. K. Hansen and P. Salamon, *Neural network ensembles*, IEEE transactions on pattern analysis and machine intelligence, 12 (1990), pp. 993–1001.

[125] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, The Annals of Statistics, 50 (2022), pp. 949–986.

[126] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical learning: data mining, inference and prediction*, The Mathematical Intelligencer, 27 (2005), pp. 83–85.

[127] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge university press, 2012.

[128] H. Hu and Y. M. Lu, *Universality laws for high-dimensional learning with random features*, arXiv preprint arXiv:2009.07669, (2020).

[129] H. Huang and Q. Yang, *Large scale analysis of generalization error in learning using margin based classification methods*, Journal of Statistical Mechanics: Theory and Experiment, 2020 (2020), p. 103407.

[130] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, in Advances in neural information processing systems, 2018, pp. 8571–8580.

[131] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., vol. 31, Curran Associates, Inc., 2018.

[132] A. Jacot, B. Simsek, F. Spadaro, C. Hongler, and F. Gabriel, *Implicit regularization of random feature models*, in International Conference on Machine Learning, PMLR, 2020, pp. 4631–4640.

[133] A. Jacot, B. Şimşek, F. Spadaro, C. Hongler, and F. Gabriel, *Kernel alignment risk estimator: Risk prediction from training data*, arXiv preprint arXiv:2006.09796, (2020).

[134] P. Jain, P. Kar, et al., *Non-convex optimization for machine learning*, Foundations and Trends® in Machine Learning, 10 (2017), pp. 142–363.

[135] A. Javanmard and A. Montanari, *State evolution for general approximate message passing algorithms, with applications to spatial coupling*, Information and Inference: A Journal of the IMA, 2 (2013), pp. 115–144.

[136] J. Jin, *Impossibility of successful classification when useful features are rare and weak*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 8859–8864.

[137] Y. Kabashima, *A cdma multiuser detection algorithm on the basis of belief propagation*, Journal of Physics A: Mathematical and General, 36 (2003), p. 11111.

[138] ——, *Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels*, in Journal of Physics: Conference Series, vol. 95, IOP Publishing, 2008, p. 012001.

[139] Y. Kabashima and S. Uda, *A bp-based algorithm for performing bayesian inference in large perceptron-type networks*, in International Conference on Algorithmic Learning Theory, Springer, 2004, pp. 479–493.

[140] Y. Kabashima, T. Wadayama, and T. Tanaka, *A typical reconstruction limit for compressed sensing based on lp-norm minimization*, Journal of Statistical Mechanics: Theory and Experiment, 2009 (2009), p. L09003.

[141] N. E. Karoui, *The spectrum of kernel random matrices*, The Annals of Statistics, 38 (2010), pp. 1 – 50.

[142] T. Karras, T. Aila, S. Laine, and J. Lehtinen, *Progressive growing of GANs for improved quality, stability, and variation*, in International Conference on Learning Representations, 2018.

[143] T. Karras, S. Laine, and T. Aila, *A style-based generator architecture for generative adversarial networks*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[144] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).

[145] G. Kini and C. Thrampoulidis, *Analytic study of double descent in binary classification: The impact of loss*, Preprint arXiv:2001.11572, (2020).

[146] D. E. Kirk, *Optimal control theory: an introduction*, Courier Corporation, 2004.

[147] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, *Optimization by simulated annealing*, science, 220 (1983), pp. 671–680.

[148] S. G. Krantz and H. R. Parks, *A primer of real analytic functions*, Springer Science & Business Media, 2002.

[149] W. Krauth, *Statistical mechanics: algorithms and computations*, vol. 13, OUP Oxford, 2006.

[150] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 25 (2012).

[151] A. Krogh and P. Sollich, *Statistical mechanics of ensemble learning*, Phys. Rev. E, 55 (1997), pp. 811–825.

[152] A. Krogh and J. Vedelsby, *Neural network ensembles, cross validation, and active learning*, in Advances in Neural Information Processing Systems, G. Tesauro, D. Touretzky, and T. Leen, eds., vol. 7, MIT Press, 1995.

[153] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, *Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices*, Journal of Statistical Mechanics: Theory and Experiment, 2012 (2012), p. P08009.

[154] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, *Statistical-physics-based reconstruction in compressed sensing*, Physical Review X, 2 (2012), p. 021005.

[155] F. Krzakala and L. Zdeborová, *Statistical physics methods in optimization and machine learning*, 2021.

[156] B. Lakshminarayanan, A. Pritzel, and C. Blundell, *Simple and scalable predictive uncertainty estimation using deep ensembles*, 2017.

[157] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[158] Y. LeCun and C. Cortes, *Mnist database*, ATT Labs [Online], (2010). Database released under CC BY-SA 3.0 license at http://yann.lecun.com/exdb/mnist/.

[159] O. Ledoit and S. Péché, *Eigenvectors of some large sample covariance matrix ensembles*, Probability Theory and Related Fields, 151 (2011), pp. 233–264.

[160] M. Ledoux, *The concentration of measure phenomenon*, American Mathematical Soc., 2001.

[161] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*, vol. 23, Springer Science & Business Media, 1991.

[162] D. LeJeune, H. Javadi, and R. Baraniuk, *The implicit regularization of ordinary least squares ensembles*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3525–3535.

[163] M. Lelarge and L. Miolane, *Fundamental limits of symmetric low-rank matrix estimation*, Probability Theory and Related Fields, 173 (2019), pp. 859–929.

[164] T. Lesieur, F. Krzakala, and L. Zdeborová, *Phase transitions in sparse pca*, in 2015 IEEE International Symposium on Information Theory (ISIT), IEEE, 2015, pp. 1635–1639.

[165] ——, *Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications*, Journal of Statistical Mechanics: Theory and Experiment, 2017 (2017), p. 073403.

[166] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM Journal on Optimization, 26 (2016), pp. 57–95.

[167] Y. Li and J. Jia, *L1 least squares for sparse high-dimensional LDA*, Electronic Journal of Statistics, 11 (2017), pp. 2499 – 2518.

[168] T. Liang, S. Sen, and P. Sur, *High-dimensional asymptotics of langevin dynamics in spiked matrix models*, arXiv preprint arXiv:2204.04476, (2022).

[169] T. Liang and P. Sur, *A precise high-dimensional asymptotic theory for Boosting and minimum-ℓ_1-norm interpolated classifiers*, Preprint arXiv:2002.01586, (2020).

[170] Z. Liao, R. Couillet, and M. W. Mahoney, *A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent*, in Advances in Neural Information Processing Systems, vol. 33, 2020.

[171] L. Lin and E. Dobriban, *What Causes the Test Error? Going Beyond Bias-Variance via ANOVA*, Journal of Machine Learning Research, 22 (2021), pp. 1–82.

[172] C. Liu, G. Biroli, D. R. Reichman, and G. Szamel, *Dynamics of liquids in the large-dimensional limit*, Physical Review E, 104 (2021), p. 054606.

[173] F. Liu, Z. Liao, and J. A. Suykens, *Kernel regression in high dimension: Refined analysis beyond double descent*, arXiv preprint arXiv:2010.02681, (2020).

[174] C. Louart and R. Couillet, *Concentration of measure and large random matrices with an application to sample covariance matrices*, arXiv preprint arXiv:1805.08295, (2018).

[175] C. Louart, Z. Liao, and R. Couillet, *A random matrix approach to neural networks*, The Annals of Applied Probability, 28 (2018), pp. 1190–1248.

[176] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová, *Learning curves of generic features maps for realistic datasets with a teacher-student model*, Advances in Neural Information Processing Systems, 34 (2021), pp. 18137–18151.

[177] B. Loureiro, C. Gerbelot, M. Refinetti, G. Sicuro, and F. Krzakala, *Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension*, International Conference on Machine Learning (ICML), (2022).

[178] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, *Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions*, Advances in Neural Information Processing Systems, 34 (2021), pp. 10144–10157.

[179] C. Ma, L. Wu, et al., *The barron space and the flow-induced function spaces for neural network models*, Constructive Approximation, 55 (2022), pp. 369–406.

[180] Y. Ma, C. Rush, and D. Baron, *Analysis of approximate message passing with a class of non-separable denoisers*, in 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 231–235.

[181] Q. Mai, H. Zou, and M. Yuan, *A direct approach to sparse discriminant analysis in ultra-high dimensions*, Biometrika, 99 (2012), pp. 29–42.

[182] X. Mai, Z. Liao, and R. Couillet, *A large scale analysis of logistic regression: Asymptotic performance and new insights*, in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3357–3361.

[183] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, *Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising*, arXiv preprint arXiv:2110.08775, (2021).

[184] T. Maimbourg, J. Kurchan, and F. Zamponi, *Solution of the dynamics of liquids in the large-dimensional limit*, Physical review letters, 116 (2016), p. 015902.

[185] S. Mallat, *A wavelet tour of signal processing*, Elsevier, 1999.

[186] A. Manacorda, G. Schehr, and F. Zamponi, *Numerical solution of the dynamical mean field theory of infinite-dimensional equilibrium liquids*, The Journal of Chemical Physics, 152 (2020), p. 164506.

[187] S. S. Mannelli and P. Urbani, *Analytical study of momentum-based acceleration methods in paradigmatic high-dimensional non-convex problems*, 2021.

[188] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová, *Multi-layer generalized linear estimation*, in 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 2098–2102.

[189] V. A. Marchenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Matematicheskii Sbornik, 114 (1967), pp. 507–536.

[190] S. Mei and A. Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv preprint arXiv:1908.05355, (2019).

[191] ——, *The generalization error of random features regression: Precise asymptotics and the double descent curve*, Communications on Pure and Applied Mathematics, (2021).

[192] C. A. Metzler, A. Maleki, and R. G. Baraniuk, *Bm3d-amp: A new image recovery algorithm based on bm3d denoising*, in 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 3116–3120.

[193] M. Mézard, *The space of interactions in neural networks: Gardner's computation with the cavity method*, Journal of Physics A: Mathematical and General, 22 (1989), p. 2181.

[194] ——, *Mean-field message-passing equations in the hopfield model and its generalizations*, Physical Review E, 95 (2017), p. 022117.

[195] M. Mezard and A. Montanari, *Information, physics, and computation*, Oxford University Press, 2009.

[196] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, vol. 9, World Scientific Publishing Company, 1987.

[197] F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborova, *The role of regularization in classification of high-dimensional noisy Gaussian mixture*, in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18 Jul 2020, pp. 6874–6883.

[198] F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová, *Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification*, Advances in Neural Information Processing Systems, 33 (2020), pp. 9540–9550.

[199] F. Mignacco and P. Urbani, *The effective noise of stochastic gradient descent*, Journal of Statistical Mechanics: Theory and Experiment, 2022 (2022), p. 083405.

[200] F. Mignacco, P. Urbani, and L. Zdeborová, *Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem*, Machine Learning: Science and Technology, 2 (2021), p. 035029.

[201] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, *Recurrent neural network based language model.*, in Interspeech, vol. 2, Makuhari, 2010, pp. 1045–1048.

[202] T. Minka, *Expectation propagation for approximate bayesian inference*, arXiv preprint arXiv:1301.2294, (2013).

[203] T. P. Minka, *A family of algorithms for approximate Bayesian inference*, PhD thesis, Massachusetts Institute of Technology, 2001.

[204] L. Miolane and A. Montanari, *The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning*, The Annals of Statistics, 49 (2021), pp. 2313–2335.

[205] P. P. Mitra, *Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation*, arXiv preprint arXiv:1906.03667, (2019).

[206] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*, MIT press, 2018.

[207] A. Montanari, *Optimization of the sherrington–kirkpatrick hamiltonian*, SIAM Journal on Computing, (2021), pp. FOCS19–1.

[208] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, *The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime*, arXiv preprint arXiv:1911.01544, (2019).

[209] A. Montanari and B. Saeed, *Universality of empirical risk minimization*, arXiv preprint arXiv:2202.08832, (2022).

[210] M. V. Narkhede, P. P. Bartakke, and M. S. Sutaone, *A review on weight initialization strategies for neural networks*, Artificial intelligence review, (2021), pp. 1–32.

[211] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, *A modern take on the bias-variance tradeoff in neural networks*, arXiv preprint arXiv:1810.08591, (2018).

[212] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2003.

[213] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate o (1/kˆ 2)*, in Dokl. akad. nauk Sssr, vol. 269, 1983, pp. 543–547.

[214] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan, *A general analysis of the convergence of admm*, in International Conference on Machine Learning, PMLR, 2015, pp. 343–352.

[215] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*, Clarendon Press, 2001.

[216] D. Opitz and R. Maclin, *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research, 11 (1999), pp. 169–198.

[217] M. Opper and W. Kinzel, *Statistical mechanics of generalization*, in Models of neural networks III, Springer, 1996, pp. 151–209.

[218] M. Opper and R. Urbanczik, *Universal learning curves of support vector machines*, Phys. Rev. Lett., 86 (2001), pp. 4410–4413.

[219] M. Opper, O. Winther, and M. Jordan, *Expectation consistent approximate inference.*, Journal of Machine Learning Research, 6 (2005).

[220] S. Oymak, C. Thrampoulidis, and B. Hassibi, *The squared-error of generalized lasso: A precise analysis*, in 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2013, pp. 1002–1009.

[221] D. Panchenko, *The sherrington-kirkpatrick model*, Springer Science & Business Media, 2013.

[222] P. Pandit, M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. Fletcher, *Inference in multi-layer networks with matrix-valued unknowns*, arXiv preprint arXiv:2001.09396, (2020).

[223] V. Papyan, X. Y. Han, and D. L. Donoho, *Prevalence of neural collapse during the terminal phase of deep learning training*, Proceedings of the National Academy of Sciences, 117 (2020), pp. 24652–24663.

[224] N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in optimization, 1 (2014), pp. 127–239.

[225] G. Parisi, *Infinite number of order parameters for spin-glasses*, Physical Review Letters, 43 (1979), p. 1754.

[226] ——, *The order parameter for spin glasses: a function on the interval 0-1*, Journal of Physics A: Mathematical and General, 13 (1980), p. 1101.

[227] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan kaufmann, 1988.

[228] S. Péché, *A note on the pennington-worah distribution*, Electronic Communications in Probability, 24 (2019), pp. 1–7.

[229] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *Scikit-learn: Machine learning in python*, the Journal of machine Learning research, 12 (2011), pp. 2825–2830.

[230] J. Pennington and P. Worah, *Nonlinear random matrix theory for deep learning*, in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 2637–2646.

[231] ——, *Nonlinear random matrix theory for deep learning*, in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017.

[232] M. Perrone, *Putting it all together: Methods for combining neural networks*, in Advances in Neural Information Processing Systems, J. Cowan, G. Tesauro, and J. Alspector, eds., vol. 6, Morgan-Kaufmann, 1994.

[233] M. P. Perrone and L. N. Cooper, *When networks disagree: Ensemble methods for hybrid neural networks*, in Neurips, Chapman and Hall, 1993, pp. 126–142.

[234] P. P. Petrushev, *Approximation by ridge functions and neural networks*, SIAM Journal on Mathematical Analysis, 30 (1998), pp. 155–189.

[235] L. Pillaud-Vivien, A. Rudi, and F. Bach, *Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes*, in Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 8114–8124.

[236] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, Ussr computational mathematics and mathematical physics, 4 (1964), pp. 1–17.

[237] A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, arXiv preprint arXiv:1511.06434, (2015).

[238] A. Rahimi and B. Recht, *Random Features for Large-Scale Kernel Machines*, in NIPS, 2007, pp. 1177–1184.

[239] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in Advances in neural information processing systems, 2008, pp. 1177–1184.

[240] S. Rangan, *Generalized approximate message passing for estimation with random linear mixing*, in 2011 IEEE International Symposium on Information Theory Proceedings, IEEE, 2011, pp. 2168–2172.

[241] S. Rangan and A. Fletcher, *Iterative estimation of constrained rank-one matrices in noise*, in 2012 IEEE International Symposium on Information Theory Proceedings, IEEE, 2012, pp. 1246–1250.

[242] S. Rangan, P. Schniter, and A. K. Fletcher, *Vector approximate message passing*, IEEE Transactions on Information Theory, (2019).

[243] T. Richardson and R. Urbanke, *Modern coding theory*, Cambridge university press, 2008.

[244] R. T. Rockafellar, *Convex analysis*, vol. 18, Princeton university press, 1970.

[245] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain.*, Psychological review, 65 (1958), p. 386.

[246] S. Rosset, J. Zhu, and T. Hastie, *Margin maximizing loss functions.*, in NIPS, 2003, pp. 1237–1244.

[247] ——, *Margin maximizing loss functions*, in Advances in Neural Information Processing Systems, S. Thrun, L. Saul, and B. Schölkopf, eds., vol. 16, MIT Press, 2004.

[248] F. Roy, G. Biroli, G. Bunin, and C. Cammarota, *Numerical implementation of dynamical mean field theory for disordered systems: application to the lotka–volterra model of ecosystems*, Journal of Physics A: Mathematical and Theoretical, 52 (2019), p. 484001.

[249] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors*, nature, 323 (1986), pp. 533–536.

[250] C. Rush, *An asymptotic rate for the lasso loss*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3664–3673.

[251] C. Rush and R. Venkataramanan, *Finite sample analysis of approximate message passing algorithms*, IEEE Transactions on Information Theory, 64 (2018), pp. 7264–7286.

[252] F. Salehi, E. Abbasi, and B. Hassibi, *The impact of regularization on high-dimensional logistic regression*, Advances in Neural Information Processing Systems, 32 (2019).

[253] ——, *The performance analysis of generalized margin maximizers on separable data*, in International Conference on Machine Learning, PMLR, 2020, pp. 8417–8426.

[254] R. E. Schapire, *The strength of weak learnability*, Machine learning, 5 (1990), pp. 197–227.

[255] P. Schniter and S. Rangan, *Compressive phase retrieval via generalized approximate message passing*, IEEE Transactions on Signal Processing, 63 (2014), pp. 1043–1055.

[256] P. Schniter, S. Rangan, and A. K. Fletcher, *Vector approximate message passing for the generalized linear model*, in 2016 50th Asilomar Conference on Signals, Systems and Computers, IEEE, 2016, pp. 1525–1529.

[257] B. Scholkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Adaptive Computation and Machine Learning, MIT Press, 2018.

[258] B. Schölkopf, A. J. Smola, F. Bach, et al., *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.

[259] H. Schwarze and J. Hertz, *Generalization in a large committee machine*, Europhysics Letters (EPL), 20 (1992), pp. 375–380.

[260] A. Sclocchi and P. Urbani, *High-dimensional optimization under nonconvex excluded volume constraints*, Physical Review E, 105 (2022), p. 024134.

[261] M. E. A. Seddik, C. Louart, M. Tamaazousti, and R. Couillet, *Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures*, in International Conference on Machine Learning, PMLR, 2020, pp. 8573–8582.

[262] ——, *Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures*, in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18 Jul 2020, pp. 8573–8582.

[263] H. S. Seung, H. Sompolinsky, and N. Tishby, *Statistical mechanics of learning from examples*, Physical review A, 45 (1992), p. 6056.

[264] J. Shao, Y. Wang, X. Deng, and S. Wang, *Sparse linear discriminant analysis by thresholding for high dimensional data*, The Annals of Statistics, 39 (2011), pp. 1241 – 1265.

[265] J. Shawe-Taylor, N. Cristianini, et al., *Kernel methods for pattern analysis*, Cambridge university press, 2004.

[266] D. Sherrington and S. Kirkpatrick, *Solvable model of a spin-glass*, Physical review letters, 35 (1975), p. 1792.

[267] H. Sifaou, A. Kammoun, and M.-S. Alouini, *Phase transition in the hard-margin support vector machines*, in 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2019, pp. 415–419.

[268] H. Sompolinsky and A. Zippelius, *Dynamic theory of the spin-glass phase*, Physical Review Letters, 47 (1981), p. 359.

[269] ——, *Relaxational dynamics of the edwards-anderson model and the mean-field theory of spin-glasses*, Physical Review B, 25 (1982), p. 6860.

[270] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, *A jamming transition from under- to over-parametrization affects generalization in deep learning*, Journal of Physics A: Mathematical and Theoretical, 52 (2019), p. 474001.

[271] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, *A jamming transition from under-to over-parametrization affects generalization in deep learning*, Journal of Physics A: Mathematical and Theoretical, 52 (2019), p. 474001.

[272] I. Steinwart, D. R. Hush, C. Scovel, et al., *Optimal rates for regularized least squares regression.*, in COLT, 2009, pp. 79–93.

[273] M. Stojnic, *A framework to characterize performance of lasso algorithms*, arXiv preprint arXiv:1303.7291, (2013).

[274] P. Sur, Y. Chen, and E. J. Candès, *The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square*, Probability Theory and Related Fields, 175 (2019), pp. 487–558.

[275] T. Suzuki, *Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality*, arXiv preprint arXiv:1810.08033, (2018).

[276] G. Szamel, *Simple theory for the dynamics of mean-field-like models of glass-forming fluids*, Physical review letters, 119 (2017), p. 155502.

[277] T. Takahashi and Y. Kabashima, *Macroscopic analysis of vector approximate message passing in a model-mismatched setting*, IEEE Transactions on Information Theory, (2022).

[278] K. Takeuchi, *Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements*, in 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 501–505.

[279] M. Talagrand et al., *Spin glasses: a challenge for mathematicians: cavity and mean field models*, vol. 46, Springer Science & Business Media, 2003.

[280] D. J. Thouless, P. W. Anderson, and R. G. Palmer, *Solution of'solvable model of a spin glass'*, Philosophical Magazine, 35 (1977), pp. 593–601.

[281] C. Thrampoulidis, E. Abbasi, and B. Hassibi, *Precise error analysis of regularized m-estimators in high dimensions*, IEEE Transactions on Information Theory, 64 (2018), pp. 5592–5628.

[282] C. Thrampoulidis, S. Oymak, and B. Hassibi, *Regularized linear regression: A precise analysis of the estimation error*, in Conference on Learning Theory, PMLR, 2015, pp. 1683–1709.

[283] C. Thrampoulidis, S. Oymak, and M. Soltanolkotabi, *Theoretical insights into multiclass classification: A high-dimensional asymptotic view*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 8907–8920.

[284] R. J. Tibshirani, *The lasso problem and uniqueness*, Electronic Journal of statistics, 7 (2013), pp. 1456–1490.

[285] A. M. Tulino, S. Verdú, et al., *Random matrix theory and wireless communications*, Foundations and Trends® in Communications and Information Theory, 1 (2004), pp. 1–182.

[286] S. S. Varadhan, *Large deviations and applications*, SIAM, 1984.

[287] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, arXiv:1011.3027, (2010).

[288] ——, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.

[289] C. Villani, *Optimal transport: old and new*, vol. 338, Springer, 2009.

[290] M. J. Wainwright, M. I. Jordan, et al., *Graphical models, exponential families, and variational inference*, Foundations and Trends® in Machine Learning, 1 (2008), pp. 1–305.

[291] K. Wang and C. Thrampoulidis, *Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization*, preprint, (2021).

[292] L. Wasserman, *All of statistics: a concise course in statistical inference*, vol. 26, Springer, 2004.

[293] ——, *All of nonparametric statistics*, Springer Science & Business Media, 2006.

[294] T. L. Watkin, A. Rau, and M. Biehl, *The statistical mechanics of learning a rule*, Reviews of Modern Physics, 65 (1993), p. 499.

[295] P. Weiss, *L'hypothèse du champ moléculaire et la propriété ferromagnétique*, J. Phys. Theor. Appl., 6 (1907), pp. 661–690.

[296] C. K. I. Williams, *Computing with infinite networks*, in Proceedings of the 9th International Conference on Neural Information Processing Systems, NIPS'96, Cambridge, MA, USA, 1996, MIT Press, p. 295–301.

[297] D. Wu and J. Xu, *On the optimal weighted ℓ_2 regularization in overparameterized linear regression*, in Advances in Neural Information Processing Systems, vol. 33, 2020.

[298] H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms*, arXiv preprint arXiv:1708.07747, (2017).

[299] J. S. Yedidia, W. T. Freeman, Y. Weiss, et al., *Understanding belief propagation and its generalizations*, Exploring artificial intelligence in the new millennium, 8 (2003), pp. 0018–9448.

[300] L. Zdeborová and F. Krzakala, *Statistical physics of inference: Thresholds and algorithms*, Advances in Physics, 65 (2016), pp. 453–552.

[301] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, in ICLR, 2017.

RÉSUMÉ

Les succès pratiques récents de l'apprentissage automatique dans toutes les tâches qui impliquent de l'analyse de données ont provoqué le besoin d'une théorie allant au-delà des statistiques classiques. A cet égard, le domaine de la physique statistique des milieux désordonnés propose une littérature conséquente dans l'analyse asymptotique exacte de systèmes aléatoires en grandes dimensions. Bien qu'ils soient efficaces, de nombreux outils issus de la physique statistique ne sont pas rigoureux et les modèles auxquels ils sont appliqués manquent de liens avec des scénarios réalistes d'apprentissage statistique. Cela motive l'introduction de modèles avec des données structurées et des méthodes d'apprentissage plus proches de l'état de l'art, ainsi que l'extension des méthodes de preuves existantes à ces problèmes. Cette thèse s'intéresse donc aux propriétés mathématiques d'une famille de fonctions implicites de grandes matrices aléatoires rencontrées en apprentissage supervisé ainsi qu'en inférence, notamment dans le contexte de la minimisation de risque empirique convexe. Nous établissons tout d'abord une extension des résultats de concentration existants pour la dynamique d'algorithmes de passage de messages approximés, et illustrons cette théorie sur des problèmes d'inférences dans des modèles probabilistes génératifs convolutionels multicouches. Nous montrons également que des méthodes de preuves similaires permettent d'obtenir des résultats asymptotiques pour la dynamique de la descente de gradient stochastique avec des données aléatoires. Nous utilisons ensuite ces résultats pour étudier le comportement statistique d'une famille de modèles linéaires généralisés convexes sous l'hypothèses de données aléatoires qui incluent des transformations de prédicteurs et de données allant au-delà de l'hypothèse i.i.d. Gaussienne, l'aggrégation de prédicteurs, les problèmes multiclasses, et différentes régularisations. Les évaluations numériques des formules établies montrent que, pour de nombreux modèles et tâches d'apprentissage, les courbes de performance obtenues par les prédictions théoriques correspondant à des modèles synthétiques Gaussiens corrélés dont les matrices de covariance sont celles des données empiriques, capturent exactement les courbes des problèmes réels. Les méthodes de preuve sont basées sur les éléments de théorie des probabilités inspirés de la physique statistique des verres de spin, l'optimisation et l'analyse convexe.

MOTS CLÉS

Physique statistique, Apprentissage Statistique, Probabilités en haute dimension, Analyse convexe, Théorie de l'information, Optimisation

ABSTRACT

The recent empirical success of machine learning in all fields involving data analysis has prompted the need for a quantitative theory that goes beyond classical statistics. In this regard, the field of statistical physics of disordered systems proposes a rich litterature in the asymptotically exact study of high-dimensional random systems. Although they are efficient, many of the tools found in statistical physics are non-rigorous and the models they are applied to lack links with realistic machine learning scenarios. This motivates the introduction of models with structured data and learning methods that are closer to the state of the art, as well as the extension of existing proof methods to those problems. With this goal in mind, the present work deals with the mathematical properties of a family of implicit functions of large random matrices encountered in supervised learning and inference, notably in the context of convex empirical risk minimization. We first establish an extension of existing concentration results for the dynamics of approximate message passing algorithms, and illustrate this theory on inference in probabilistic models with multilayer random convolutional generative priors. We also show how related ideas enable to obtain the high-dimensional dynamics of stochastic gradient descent with random data. We then use those results to study the statistical behaviour of a family of convex generalised linear models under the random design hypothesis including feature maps and data models going beyond the i.i.d. Gaussian setting, ensembling of predictors, multiclass problems and different regularisations. We also show numerically that for a wide range of tasks and realistic feature maps, the learning curves obtained from the theoretical prediction corresponding to the synthetic Gaussian models with matching covariances exactly capture those of the original problems. The proof methods are based on the elements of probability theory inspired by the statistical physics of spin glasses, optimization and convex analysis.

KEYWORDS

Statistical Physics, Statistical learning, High-dimensional probability, Convex analysis, Information Theory, Optimization