Sample covariance random matrices arising in artificial neural networks

Clement Chouard

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Clément CHOUARD

Le 16 octobre 2023

Matrices aléatoires de covariance et réseaux de neurones artificiels

---

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Mathématiques et Applications**

Unité de recherche :
IMT : Institut de Mathématiques de Toulouse

Thèse dirigée par
Mireille CAPITAINE et Guillaume CEBRON

Jury

M. Fan ZHOU, Rapporteur
M. Lucas BéNIGNI, Examinateur
M. Michel LEDOUX, Examinateur
Mme Mireille CAPITAINE, Directrice de thèse
M. Guillaume CéBRON, Co-directeur de thèse
M. Jamal NAJIM, Président

Remerciements

Je suis très heureux d'avoir été encadré par Mireille Capitaine et Guillaume Cébron pour réaliser cette thèse. Je vous remercie pour votre confiance, votre disponibilité et l'enthousiasme avec lequel vous avez partagé vos idées mathématiques pendant toutes ces années. Je vous remercie pour votre bienveillance et vos qualités humaines inestimables qui m'ont aidé à traverser les moments difficiles.

Je remercie mes rapporteurs Jamal Najim et Zhou Fan qui ont lu et évalué mon travail. Je remercie Lucas Bégnini et Michel Ledoux qui ont accepté de faire partie de mon jury. Je suis reconnaissant envers le Ministère de l'Enseignement Supérieur et de la Recherche ainsi que l'École Universitaire de Recherche MINT (référence CANR-18-EURE-0023) qui ont financé cette thèse.

Je tiens à remercier mes collègues de l'Institut de Mathématiques de Toulouse. J'ai eu autant de plaisir à vous côtoyer au bureau qu'à l'extérieur de l'Université. Je remercie aussi tous les membres de l'attachante communauté des matrices aléatoires. Je pense en particulier aux personnes rencontrées pendant mes visites à Bordeaux, Lausanne, ou bien au sein du groupe de recherche MEGA.

Je remercie tous les amis venus me rendre visite pendant ma thèse, qui ont toujours su trouver les bons mots pour m'encourager. Je remercie toutes les merveilleuses personnes rencontrées à Toulouse qui m'ont permis de m'épanouir dans cette ville. Une mention spéciale va aux bénéloves de l'Escabel et des associations auxquelles j'ai participé. Ces manières alternatives de contribuer à la société sont essentielles pour moi et tout à fait complémentaires de mon travail de recherche. Je remercie enfin toutes les personnes avec qui j'ai eu la chance de vivre ces dernières années. Cette thèse n'aurait pas abouti sans votre présence et votre affection.

Table des matières

1 Introduction **1**

1.1 Théorie classique des matrices aléatoires de covariance 3

1.1.1 Transformée de Stieltjes 3

1.1.2 Lois globales 5

1.1.3 Résolvante et notion d'équivalent déterministe 7

1.1.4 Lois locales 9

1.2 Équivalent déterministe pour les modèles avec une structure de dépendance partielle 12

1.2.1 Concentration lipschitzienne 12

1.2.2 Résultat principal 14

1.2.3 Stratégie de preuve 16

1.2.4 Convergence en distance de Kolmogorov 17

1.3 Réseaux de neurones artificiels et liens avec les matrices aléatoires 20

1.3.1 Principes de fonctionnement des réseaux de neurones artificiels 20

1.3.2 Lois globales pour le modèle du noyau conjugué 21

1.4 Équivalent déterministe pour le modèle du noyau conjugué . . 25

1.4.1 Résultat principal 25

1.4.2 Idées de démonstration 28

1.4.3 Modèle linéaire gaussien équivalent 29

1.4.4 Lien avec l'erreur d'apprentissage du réseau 31

2 Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure **35**

2.1 Introduction 37

2.2 Setting 40

2.2.1 Notations and definitions 40

2.2.2 Main results 41

2.2.3 Application to Kolmogorov distances 42

2.2.4 Application to kernel methods 43

2.2.5 Organization of the paper 43

2.3 Concentration framework 45

2.4 Properties of resolvent matrices 48

2.4.1 Sample covariance resolvent 48

2.4.2 LOO resolvent and co-resolvent 49
2.5 First deterministic equivalent 53
2.5.1 General properties of the deterministic equivalents . . . 53
2.5.2 Introduction of parameters $\mathfrak{a}$ and $\mathfrak{b}$ 53
2.5.3 Proof of the first deterministic equivalent 54
2.6 Second deterministic equivalent 57
2.6.1 Reformulation as a fixed point problem 57
2.6.2 Stability of $\mathcal{F}$ in the real case 59
2.6.3 Stability of $\mathcal{F}$ in the complex case 60
2.6.4 Proof of the second deterministic equivalent 64
2.7 Proof of the main results 66
2.8 Bounds in Kolmogorov distance 68
2.8.1 Bound for fixed measures 68
2.8.2 Asymptotic bound 69
2.8.3 Application to the empirical spectral distribution of sample covariance matrices 70
2.9 Application to kernel methods 74
2.9.1 Kernel ridge regression 74
2.9.2 Random features method 74
2.9.3 Effective ridge parameter 75

3 Deterministic equivalent of the conjugate kernel matrix associated to artificial neural networks 79
3.1 Introduction 81
3.1.1 Overview of the article 83
3.1.2 General notations and definitions 83
3.2 Technical tools 85
3.2.1 Concentration framework 85
3.2.2 Polynomial bounds in z and notation $O_z(\epsilon_n)$ 86
3.3 Covariance matrices of coordinate-wise functions of Gaussian vectors 89
3.3.1 Hermite polynomials 89
3.3.2 Iterated Hadamard products 91
3.3.3 General expansion of the covariance matrix Σ 92
3.3.4 Approximation of Σ for weakly correlated Gaussian vectors 92
3.3.5 Linearization of Σ in specific settings 94
3.4 Deterministic equivalent of sample covariance matrices with a general dependence structure 96
3.4.1 General results 97
3.4.2 Regularity of the Stieltjes transform with respect to the free convolution 98

3.4.3 Approximation of the deterministic equivalent built from deterministic matrices 102
3.4.4 Concentration of the deterministic equivalent built from random matrices 103
3.5 Single-layer neural network with deterministic data 105
3.5.1 Setting 105
3.5.2 Technicalities and linearization of Σ 106
3.5.3 Propagation of the approximate orthogonality 107
3.5.4 Deterministic equivalent and consequences 108
3.5.5 Application to another model involving entry-wise operations 111
3.6 Single-layer neural network with random data 113
3.6.1 Setting 113
3.6.2 Deterministic equivalent and consequences 114
3.6.3 Application to data matrices with i.i.d. columns 117
3.7 Multi-layer neural network model 119
3.8 Appendix : Bounds on Kolmogorov distances between empirical measures 123

Bibliographie **125**

Chapitre 1

Introduction

Cette thèse porte sur certains modèles de matrices aléatoires en lien avec les réseaux de neurones artificiels. Elle est composée de deux textes principaux, les chapitres 2 et 3, directement issus d'articles destinés à être publiés, et qui sont de ce fait rédigés en anglais.

Le premier de ces deux textes étudie une classe générale de matrices de covariance empirique présentant une structure de dépendance partielle. Cette analyse complète plusieurs travaux antérieurs dans le domaine des matrices aléatoires, et présente donc un intérêt théorique en elle-même. Elle trouve également des applications pratiques dans certains problèmes d'apprentissage machine.

Le deuxième article qui constitue cette thèse porte sur le modèle du noyau conjugué. Ce modèle est un ensemble de matrices de covariance directement inspiré du fonctionnement de certains types de réseaux de neurones artificiels. Nous établissons certaines propriétés théoriques de ce modèle, qui généralisent et complètent les travaux antérieurs dans ce domaine.

L'objectif de cette introduction est de présenter dans un premier temps les concepts et les résultats fondateurs sur les matrices aléatoires de covariance (section 1.1). Nous pourrons alors contextualiser nos recherches portant sur les modèles avec une structure de dépendance partielle (section 1.2). Dans un second temps, nous expliquerons comment les matrices aléatoires interviennent naturellement dans l'étude des réseaux de neurones artificiels (section 1.3). Nous décrirons ensuite nos travaux sur le modèle dit du noyau conjugué (section 1.4), et enfin nous explorerons quelques implications pratiques de ces résultats.

Contents

1.1 Théorie classique des matrices aléatoires de covariance **3**
1.1.1 Transformée de Stieltjes 3
1.1.2 Lois globales 5
1.1.3 Résolvante et notion d'équivalent déterministe 7
1.1.4 Lois locales 9
1.2 Équivalent déterministe pour les modèles avec une structure de dépendance partielle **12**
1.2.1 Concentration lipschitzienne 12
1.2.2 Résultat principal 14
1.2.3 Stratégie de preuve 16
1.2.4 Convergence en distance de Kolmogorov 17
1.3 Réseaux de neurones artificiels et liens avec les matrices aléatoires **20**
1.3.1 Principes de fonctionnement des réseaux de neurones artificiels 20
1.3.2 Lois globales pour le modèle du noyau conjugué . . . 21
1.4 Équivalent déterministe pour le modèle du noyau conjugué . **25**
1.4.1 Résultat principal 25
1.4.2 Idées de démonstration 28
1.4.3 Modèle linéaire gaussien équivalent 29
1.4.4 Lien avec l'erreur d'apprentissage du réseau 31

1.1 Théorie classique des matrices aléatoires de covariance

L'étude des matrices aléatoires en grande dimension a commencé dans les années 1950 avec les travaux fondateurs de Wigner ([Wig57]). Pour modéliser des problèmes issus de la physique nucléaire, où les niveaux d'énergie d'un système correspondent aux valeurs propres d'un opérateur hamiltonien, Wigner considère des matrices carrées symétriques à entrées indépendantes et identiquement distribuées (i.i.d.). Il démontre la convergence de la mesure spectrale empirique de telles matrices vers une mesure de probabilité déterministe appelée loi du semi-cercle.

Quelques années plus tard, Marčenko et Pastur s'intéressent à une autre classe de matrices aléatoires, les matrices dites de covariance empirique, qui seront au cœur de cette thèse. Ce type de matrices est particulièrement utilisé en statistiques depuis les travaux de Wishart [Wis28], et la compréhension de leurs propriétés asymptotiques se révèle très utile pour établir des tests ou estimateurs statistiques en grande dimension. Dans l'article [MP67], les auteurs démontrent la convergence de la mesure spectrale empirique de certains modèles de matrices de covariance empirique vers une mesure de probabilité déterministe, depuis connue sous le nom de loi de Marčenko-Pastur.

Les résultats qui décrivent le comportement spectral asymptotique de ces matrices aléatoires peuvent être obtenus de différentes manières : par des méthodes combinatoires comme la méthode des moments, des méthodes déterminantales lorsque la loi jointe des valeurs propres est explicite, ou bien encore en utilisant des méthodes analytiques comme celles dans lesquelles s'inscrit cette thèse. Introduisons donc quelques objets essentiels qui nous permettront d'aborder les résultats fondateurs de ce domaine de recherche.

1.1.1 Transformée de Stieltjes

Nous notons $\mathbb{K}^{p\times n}$ l'ensemble des matrices avec p lignes, n colonnes, et à coefficients à valeurs dans $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Si $M \in \mathbb{K}^{p\times n}$, on note sa transposée $M^\top$, et dans le cas d'une matrice carrée sa trace $\mathrm{Tr}(M) = \sum_{i=1}^n M_{ii}$. Nous utilisons les normes vectorielles et matricielles suivantes : la norme euclidienne $\|\cdot\|$, la norme de Frobenius $\|\cdot\|_F$, la norme spectrale $|||\cdot|||$, et la norme du maximum des entrées $\|\cdot\|_{\max}$.

Une des préoccupations principales de la théorie des matrices aléatoires est de comprendre les propriétés spectrales de ces matrices lorsque leurs dimensions tendent simultanément vers $+\infty$.

Parmi les nombreuses approches possibles, nous utilisons dans cette thèse des méthodes dites analytiques, qui consistent à étudier certaines fonctions

complexes en lien avec le spectre de ces matrices.

La transformée de Stieltjes associée à une mesure réelle μ est la fonction d'une variable complexe :

$$g_\mu : z \mapsto g_\mu(z) = \int_{\mathbb{R}} \frac{1}{t-z} d\mu(t).$$

Cette intégrale a un sens lorsque la variable z, aussi appelée paramètre spectral de la transformation, se situe en dehors du support de μ. La transformée de Stieltjes g_μ est donc toujours bien définie sur le demi-plan complexe ouvert supérieur $\mathbb{C}^+ = \{z \in \mathbb{C} \text{ tels que } \Im(z) > 0\}$.

Considérons $K \in \mathbb{R}^{p\times p}$ une matrice diagonalisable, et regroupons ses valeurs propres dans un ensemble appelé spectre de K, et noté $\mathrm{Sp}K = \{\lambda \text{ valeurs propres de } K\}$. La mesure spectrale empirique de K est la mesure de probabilité $\mu_K = \frac{1}{p}\sum_{\lambda\in \mathrm{Sp}K} \delta_\lambda$, où δ_λ désigne la mesure de Dirac centrée en λ.

Pour une matrice diagonalisable $K \in \mathbb{R}^{p\times p}$, la fonction $z \in \mathbb{C}^+ \mapsto g_K(z)$ est par définition la transformée de Stieltjes associée à la mesure spectrale empirique de K :

$$g_K(z) = g_{\mu_K}(z) = \int_{\mathbb{R}} \frac{1}{t-z} d\mu_K(t) = \frac{1}{p}\sum_{\lambda\in \mathrm{Sp}K} \frac{1}{\lambda - z}.$$

On peut voir que g_K est une fonction méromorphe dont les pôles sont les valeurs propres de K, ce qui est particulièrement adapté pour étudier le spectre de K.

Un des principaux avantages de la transformée de Stieltjes est justement qu'elle caractérise la mesure dont elle est issue. On se réfère à [BS10] pour la démonstration des propriétés suivantes :

Proposition 1.1.1. *1. La transformée de Stieltjes d'une mesure de probabilité réelle μ est une fonction holomorphe sur $\mathbb{C}^+$, et vérifie $\Im(g_\mu(z)) > 0$ pour tout $z \in \mathbb{C}^+$, $|g_\mu(z)| \leq 1/\Im(z)$, et $\lim_{t\to\infty} itg_\mu(it) = -1$.*

Nous proposons d'appeler par la suite fonctions candidates *les fonctions holomorphes sur $\mathbb{C}^+$ qui vérifient ces trois propriétés.*

2. Si a et b ne sont pas des atomes de μ, alors :

$$\mu([a,b]) = \lim_{\epsilon\to 0^+} \frac{1}{\pi}\int_{[a,b]} \Im(g_\mu(t+i\epsilon))dt.$$

Cette formule est connue sous le nom de formule d'inversion de Frobenius-Perron. En conjonction avec le théorème de représentation de Herglotz pour les fonctions holomorphes, on obtient l'importante conséquence suivante :

3. *La transformée de Stieltjes est une bijection de l'ensemble des mesures de probabilité réelles sur l'ensemble des fonctions candidates, c'est à dire des fonctions g holomorphes sur $\mathbb{C}^+$, telles que $\Im(g(z)) > 0$ pour tout $z \in \mathbb{C}^+$, $|g(z)| \leq 1/\Im(z)$, et $\lim_{t\to\infty} itg(it) = -1$.*
   La transformée de Stieltjes d'une mesure caractérise donc cette mesure.
4. *Si (μ_n) et μ_∞ sont des mesures de probabilité réelles, la convergence faible de (μ_n) vers μ_∞ est équivalente à la convergence simple sur $\mathbb{C}^+$ de la suite des transformées de Stieltjes (g_{μ_n}) vers g_{μ_∞}.*

Signalons enfin que certains auteurs utilisent une convention pour les signes opposée à la nôtre, et travaillent donc avec la transformée de Cauchy égale à $-g_\mu(z) = \int_{\mathbb{R}} \frac{1}{z-t} d\mu(t)$. Nous verrons d'ailleurs à plusieurs instances dans ce manuscrit l'inverse de cette transformée de Cauchy, égale à $l_\mu(z) = -1/g_\mu(z)$.

1.1.2 Lois globales

Considérons $Y \in \mathbb{R}^{p\times n}$ une matrice rectangulaire, à entrées réelles. La matrice de covariance empirique associée à Y est par définition la matrice $K = \frac{1}{n} YY^\top$. K est une matrice carrée de dimension p, réelle, symétrique, et semi-définie positive. K est donc diagonalisable en base orthonormée, avec p valeurs propres réelles positives ou nulles, dont on cherche à comprendre le comportement asymptotique.

La loi de Marčenko-Pastur de paramètre $\gamma > 0$, notée $\mathrm{MP}(\gamma)$, est la mesure de probabilité réelle μ, avec une masse en 0 égale à $\mu(\{0\}) = 1 - 1/\gamma$ si $\gamma > 1$, et qui admet la densité suivante sur l'intervalle $\left[\left(1-\sqrt{\gamma}\right)^2, \left(1+\sqrt{\gamma}\right)^2\right]$:

$$f(t) = \frac{\sqrt{t - \left(1-\sqrt{\gamma}\right)^2}\sqrt{\left(1+\sqrt{\gamma}\right)^2 - t}}{2\pi\gamma t}.$$

De manière équivalente, $\mathrm{MP}(\gamma)$ est caractérisée par sa transformée de Stieltjes $g(z)$, seule fonction candidate (cf. Proposition 1.1.1) qui vérifie l'équation de point fixe :

$$g(z) = \frac{1}{1 - \gamma - \gamma z g(z) - z}.$$

Nous pouvons maintenant présenter quelques théorèmes centraux de la théorie des matrices aléatoires de covariance empirique. Ces résultats sont connus sous le nom de lois globales car ils décrivent le comportement asymptotique global des spectres aléatoires de ces matrices.

Commençons par le cas le plus simple où les entrées de Y sont i.i.d. On dit alors que la matrice de covariance empirique K est une matrice de Wishart.

Avertissement : les paragraphes qui suivent ont pour but d'établir un historique de notre domaine de recherche et de pouvoir replacer dans leur contexte les résultats obtenus pendant cette thèse. Nous ne proposons absolument pas les hypothèses les plus générales, ni les résultats les plus fins.

Nous prévenons aussi notre lecteur que, pour plus de lisibilité et conformément à l'usage en matrices aléatoires, nous omettrons souvent de préciser la dépendance en la dimension des objets manipulés. Ainsi dans les résultat suivants, Y, K, ... sont tous implicitement des suites indexées par n.

Théorème 1.1.1 ([MP67]). *Soit $Y \in \mathbb{R}^{p\times n}$ une matrice aléatoire telle que :*

1. *Le rapport des dimensions $\gamma_n = \frac{p}{n}$ converge vers $\gamma_\infty > 0$.*
2. *Les entrées de Y sont i.i.d. et de variance 1.*

Alors la mesure spectrale empirique de $K = \frac{1}{n}YY^\top$ converge faiblement presque sûrement (p.s.) vers la mesure de probabilité déterministe $\mathrm{MP}(\gamma_\infty)$.

Si μ est une mesure de probabilité à support dans $\mathbb{R}^+$, nous appelons produit de convolution libre de $\mathrm{MP}(\gamma)$ avec μ la mesure de probabilité caractérisée par sa transformée de Stieltjes $g(z)$, seule fonction candidate (cf. Proposition 1.1.1) qui vérifie l'équation de point fixe :

$$g(z) = \int_{\mathbb{R}} \frac{1}{(1-\gamma-\gamma z g(z))t - z} d\mu(t).$$

Nous notons $\mathrm{MP}(\gamma) \boxtimes \mu$ cette nouvelle mesure, aussi appelée loi de Marčenko-Pastur déformée par μ.

La notation $\boxtimes$ adoptée ici est issue de la théorie des probabilités libres. Sans rentrer dans le détail, $\nu \boxtimes \mu$ correspond à la loi (au sens des probabilités libres) d'un produit de deux variables aléatoires non commutatives, libres entre elles, de lois ν et μ respectivement. Nous utiliserons exclusivement dans cette thèse des produits de convolution libre avec des lois de Marčenko-Pastur, pour lesquels la définition à l'aide de la transformée de Stieltjes est suffisante.

Lorsque Y est obtenue par transformation linéaire d'une matrice à entrées i.i.d., on parle de structure de dépendance linéaire, ou bien encore de matrice de Wishart colorée pour K. Il est aussi possible dans ce cas de caractériser la limite spectrale empirique à l'aide de l'opérateur de convolution décrit ci-dessus.

Théorème 1.1.2 ([Sil95]). *Soit $\Sigma \in \mathbb{R}^{p\times p}$ une matrice déterministe symétrique semi-définie positive, $X \in \mathbb{R}^{p\times n}$ une matrice aléatoire, et $Y = \Sigma^{1/2}X$. On suppose que :*

1. *Le rapport des dimensions $\gamma_n = \frac{p}{n}$ converge vers $\gamma_\infty > 0$.*
2. *Les entrées de X sont i.i.d. et de variance 1.*

3. μ_Σ *converge faiblement vers une mesure de probabilité* μ_∞*.*

Alors la mesure spectrale empirique de $K = \frac{1}{n}YY^\top$ *converge faiblement p.s. vers la mesure de probabilité déterministe* $\mathrm{MP}(\gamma_\infty) \boxtimes \mu_\infty$*.*

Il est enfin possible de considérer le cas plus général où les colonnes de Y sont indépendantes, mais où les entrées à l'intérieur de ces colonnes peuvent être corrélées. C'est ce type de modèle présentant une structure de dépendance partielle qui va principalement nous intéresser pendant notre thèse. La loi globale correspondante est la suivante :

Théorème 1.1.3 ([BZ08])**.** *Soit* $Y \in \mathbb{R}^{p\times n}$ *une matrice aléatoire telle que :*

1. *Le rapport des dimensions* $\gamma_n = \frac{p}{n}$ *converge vers* $\gamma_\infty > 0$*.*
2. *Les colonnes de* Y *sont i.i.d. suivant la loi d'un vecteur aléatoire* y*.*
3. y *satisfait la propriété de concentration suivante : pour toute matrice déterministe* $A \in \mathbb{R}^{p\times p}$*,* $\mathrm{Var}\left(y^\top Ay\right) \leq |||A|||^2 o(n^2)$*.*
4. *Avec* $\Sigma = \mathbb{E}\left[yy^\top\right]$*,* $|||\Sigma|||$ *est bornée, et* μ_Σ *converge faiblement vers une mesure de probabilité* μ_∞*.*

Alors la mesure spectrale empirique de $K = \frac{1}{n}YY^\top$ *converge faiblement p.s. vers la mesure de probabilité déterministe* $\mathrm{MP}(\gamma_\infty) \boxtimes \mu_\infty$*.*

Il est possible de compléter ces résultats dans de nombreuses directions. Tout d'abord on peut chercher à rendre quantitative la convergence des mesures spectrales, soit en utilisant une distance sur les mesures comme la distance de Kolmogorov, soit en fournissant des estimées précises sur les transformées de Stieltjes. Ces deux questions sont d'ailleurs très proches et nous y reviendrons aux sections 1.2.4 et 2.8.

Un autre problème intéressant est de s'intéresser aux valeurs propres extrémales et à leurs fluctuations, ce qui donne lieu à une littérature très riche. Dans cette thèse, nous optons pour une approche encore différente, qui consiste à étudier un outil analytique plus précis que la transformée de Stieltjes, la résolvante de la matrice de covariance.

1.1.3 Résolvante et notion d'équivalent déterministe

Considérons $K \in \mathbb{R}^{p\times p}$ une matrice symétrique réelle, et donc en particulier diagonalisable en base orthonormée. La résolvante de K est la fonction d'une variable complexe et à valeurs matricielles :

$$\mathcal{G}_K : z \mapsto \mathcal{G}_K(z) = (K - zI_p)^{-1}.$$

La transformée de Stieltjes d'une matrice est égale à la trace normalisée de sa résolvante, autrement dit elles sont reliées par la formule $g_K(z) = \frac{1}{p}\mathrm{Tr}\mathcal{G}_K(z)$.

De plus si $\{\mathbf{e}_\lambda, \lambda \in \mathrm{Sp}K\}$ désigne une base orthonormée associée aux valeurs propres de K, il est facile de voir que :

$$\mathcal{G}_K(z) = \sum_{\lambda \in \mathrm{Sp}K} \frac{1}{\lambda - z} \mathbf{e}_\lambda \mathbf{e}_\lambda^\top.$$

Ainsi $\mathcal{G}_K(z)$ est une fonction méromorphe sur le plan complexe, dont les pôles sont les valeurs propres de K, et dont les résidus correspondent aux projecteurs orthonormaux associés à ces valeurs propres. Une compréhension fine du comportement asymptotique de la résolvante apporte donc non seulement des informations sur le spectre de la matrice, mais aussi sur ses directions propres.

La résolvante d'une matrice possède une structure algébrique très riche et de nombreuses autres propriétés, notamment de concentration, dont il est possible de tirer parti pour étudier les matrices aléatoires. Nous renvoyons à la section 2.4 pour une analyse approfondie de cet objet.

Nous avons vu que la convergence faible d'une suite de mesures aléatoires vers une mesure déterministe p.s. était équivalente à la donnée d'une limite déterministe p.s. pour leurs transformées de Stieltjes, et nous aurions pu formuler les Théorèmes 1.1.1, 1.1.2 et 1.1.3 en ce sens. À la différence de la transformée de Stieltjes cependant, la résolvante est une suite de fonctions à valeurs matricielles dont les dimensions tendent vers $+\infty$, ce qui rend la notion de limite plus difficile à formuler. Nous utilisons pour cela le concept d'équivalent déterministe de la résolvante, en suivant une terminologie issue de [HLN07].

Si $K \in \mathbb{R}^{p\times p}$ est une matrice aléatoire, un équivalent déterministe de sa résolvante $\mathcal{G}_K(z)$ est une fonction à valeurs matricielles $\mathbf{G} : \mathbb{C}^+ \to \mathbb{C}^{p\times p}$, telle que pour toute matrice $A \in \mathbb{R}^{p\times p}$, vérifiant certaines contraintes de normalisation suivant les résultats, la quantité $\mathrm{Tr}(\mathcal{G}_K(z)A - \mathbf{G}(z)A)$ converge vers 0 p.s.

Cette formulation d'un équivalent déterministe traduit également un phénomène de concentration typique des matrices aléatoires. En effet lorsque les dimensions des espaces augmentent, les objets aléatoires ne restent pas proches de leur espérance en norme, et ils ne peuvent donc pas être approchés par des objets déterministes. En revanche les formes linéaires issues de ces matrices aléatoires se concentrent bien autour de leur espérance, et il est possible d'utiliser ce phénomène pour parler d'équivalent déterministe de matrices en grande dimension. Nous discuterons plus en détail de ces phénomènes de concentration à la section 1.2.1.

Explorons quelques conséquences immédiates apportées par un équivalent déterministe de la résolvante : en choisissant dans la définition $A = \frac{1}{p} I_p$, on obtient que $g_K(z)$ et $\frac{1}{p}\mathrm{Tr}\mathbf{G}(z)$ ont la même limite si elle existe, et donc que μ_K

converge p.s. vers une mesure admettant cette limite comme transformée de Stieltjes. Ainsi un équivalent déterministe de la résolvante est une propriété plus forte qu'une loi globale sur la mesure spectrale empirique.

En fixant un vecteur unitaire $\mathbf{u} \in \mathbb{R}^p$ et $A = \mathbf{u}\mathbf{u}^\top$, on obtient également une limite p.s. pour la mesure $\mu_{K,\mathbf{u}} = \sum_{\lambda \in \mathrm{Sp}K} (\mathbf{u}^\top \mathbf{e}_\lambda)^2 \delta_\lambda$, connue sous le nom de mesure spectrale empirique dans la direction $\mathbf{u}$, ou v.e.s.d. en anglais (cf.la section 2.2 et [Noi21]). En prenant pour A des matrices de base $\mathbf{E}_{ij}$, on obtient $\mathcal{G}_K(z)_{ij} - \mathbf{G}(z)_{ij} \to 0$ p.s., et on peut même obtenir cette convergence uniformément sur les entrées : $\|\mathcal{G}_K(z) - \mathbf{G}(z)\|_{\max} \to 0$ p.s.

1.1.4 Lois locales

La convergence de $\mathrm{Tr}(\mathcal{G}_K(z)A - \mathbf{G}(z)A)$ vers 0 dans notre définition d'équivalent déterministe peut être raffinée de manière quantitative en la dimension et le paramètre spectral z. Lorsque c'est le cas, de tels résultats d'équivalents déterministes quantitatifs portent le nom de lois locales. En effet, comme nous l'avons vu la résolvante contient des informations non seulement sur les valeurs propres, mais aussi sur les directions propres de la matrice.

Ce terme est aussi justifié par le fait que des estimées suffisamment précises en z peuvent conduire à des résultats fins sur les valeurs propres, à une échelle plus petite que l'ensemble des valeurs propres comme dans les lois globales. Nous n'avons pas obtenu de résultats permettant d'explorer cette direction dans notre thèse, mais nous renvoyons le lecteur intéressé par ces techniques à consulter [KY17, section 10].

Les premières lois locales obtenues concernaient les matrices de Wigner ([ESY09]). Quelques années après, le résultat suivant fournissait un équivalent pour les matrices de covariance de type Wishart.

Théorème 1.1.4. *[BEK+14] Soit $Y \in \mathbb{R}^{p\times n}$ une matrice aléatoire telle que :*

1. *Le rapport des dimensions $\gamma_n = \frac{p}{n}$ est majoré et minoré par des constantes strictement positives.*
2. *Les entrées de Y sont i.i.d., centrées, de variance 1 et de classe d'intégrabilité L^4.*

Alors la fonction à valeurs matricielles $\mathbf{G}(z) = g_{\mathrm{MP}(\gamma_n)}(z) I_p$ est un équivalent déterministe de $\mathcal{G}_K(z)$ au sens suivant : avec

$$\Psi(z) = \sqrt{\frac{\Im\Big(g_{\mathrm{MP}(\gamma_n)}(z)\Big)}{n\Im(z)}} + \frac{1}{n\Im(z)},$$

pour toute constante fixée $\tau > 0$, uniformément en $z \in \{z \in \mathbb{C}^+ \text{ tels que } |z| \geq \tau,\ |\Re(z)| \leq \tau^{-1},\ \text{et } n^{\tau-1} \leq \Im(z) \leq \tau^{-1}\}$, et uniformément en les vecteurs unitaires déterministes $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$:

$$\Big|\mathrm{Tr}\Big((\mathcal{G}_K(z) - \mathbf{G}(z))\mathbf{u}\mathbf{v}^\top\Big)\Big| \leq O\Big(\sqrt{\log n}\Psi(z)\Big) \quad \text{p.s.}$$

Prenons quelques instants pour discuter ce résultat fondamental. Tout d'abord le théorème prend bien la forme d'un équivalent déterministe de la résolvante comme nous l'avons présenté ci-dessus. Cet équivalent autorise à faire varier le paramètre spectral avec n, et notamment de le rapprocher de l'axe réel, pour peu que $\Im(z)$ ne tende pas trop vite vers 0 (strictement moins vite que $1/n$).

L'équivalent déterministe est quantitatif, avec des vitesses de convergence typiques en $1/\sqrt{n}$ pour la dimension et $1/\Im(z)$ pour le paramètre spectral. En effet $\Im\Big(g_{\mathrm{MP}(\gamma_n)}(z)\Big) \leq 1/\Im(z)$, donc $\Psi(z)$ est majorée par $O\Big(\frac{1}{\sqrt{n}\Im(z)}\Big)$. Le facteur $\sqrt{\log n}$ qui apparaît dans l'inégalité est une simple technicité pour obtenir une propriété p.s. Le résultat n'est de fait pas énoncé sous cette forme dans [BEK+14], mais en utilisant le concept propre à cet article de domination stochastique. Les auteurs obtiennent ainsi leurs estimées sur des évènements de grande probabilité, ce qui implique des bornes asymptotiques p.s. en utilisant le lemme de Borel-Cantelli, au prix d'un facteur $\sqrt{\log n}$. Nous retrouverons beaucoup de ces caractéristiques dans nos propres énoncés.

Nous pouvons aussi remarquer l'apparition de l'hypothèse de moyenne nulle pour les entrées de Y. Cette hypothèse était absente des lois globales présentés au début de cette section. Centrer une matrice aléatoire revient en effet à ajouter une matrice déterministe de rang 1, ce qui ne change pas la distribution spectrale asymptotique de cette matrice. Ce n'est plus le cas lorsqu'on considère la résolvante de cette matrice. Pour obtenir un équivalent déterministe de la résolvante, le centrage, ou au minimum un contrôle suffisant de l'espérance des colonnes, devient alors une hypothèse cruciale.

Dans le cas d'entrées i.i.d., la résolvante se comporte donc asymptotiquement comme une homothétie dont le rapport est la transformée de Stieltjes d'une loi de Marčenko-Pastur. On peut parler d'isotropie car les vecteurs propres de K n'ont pas de direction privilégiée.

Ce n'est plus le cas en général pour d'autres modèles de matrices de covariance, et on parle alors de lois locales anisotropiques. Pour les modèles avec une structure de dépendance linéaire, autrement dénommées matrices de Wishart colorées, le résultat correspondant est le suivant :

Théorème 1.1.5 ([KY17]). *Soit $\Sigma \in \mathbb{R}^{p\times p}$ une matrice déterministe symétrique semi-définie positive, $X \in \mathbb{R}^{p\times n}$ une matrice aléatoire, et $Y = \Sigma^{1/2}X$. On suppose que :*

1. *Le rapport des dimensions $\gamma_n = \frac{p}{n}$ est majoré et minoré par des constantes strictement positives.*
2. *Les entrées de X sont i.i.d., centrées, de variance 1, et de classe d'intégrabilité L^q pour tout $q \in \mathbb{N}$.*
3. *Σ est inversible et $|||\Sigma|||$ est bornée.*

Alors avec $\nu = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$ *et* $\tilde{\nu} = (1-\gamma_n)\cdot\delta_0 + \gamma_n \cdot \nu$, $\tilde{\nu}$ *est bien une mesure de probabilité, et la fonction à valeurs matricielles* $\mathbf{G}(z) = \left(-z g_{\tilde{\nu}}(z)\Sigma - z I_p\right)^{-1}$ *est un équivalent déterministe de* $\mathcal{G}_K(z)$ *au sens suivant : avec*

$$\Psi(z) = \sqrt{\frac{\Im(g_\nu(z))}{n\Im(z)}} + \frac{1}{n\Im(z)},$$

pour toute constante fixée $\tau > 0$, *uniformément en* $z \in \{z \in \mathbb{C}^+ \text{ tels que } |z| \geq \tau,\ |\Re(z)| \leq \tau^{-1},\ \text{et } n^{\tau-1} \leq \Im(z) \leq \tau^{-1}\}$, *et uniformément en les vecteurs unitaires déterministes* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ *:*

$$\left|\mathrm{Tr}\left(\Sigma^{-1/2}(\mathcal{G}_K(z) - \mathbf{G}(z))\Sigma^{-1/2}\mathbf{u}\mathbf{v}^\top\right)\right| \leq O\left(\sqrt{\log n}\Psi(z)\right) \quad \text{p.s.}$$

On peut montrer que l'identité $g_\nu(z) = \frac{1}{p}\mathrm{Tr}\mathbf{G}(z)$ a lieu (Proposition 2.6.2), et donc que cette loi locale implique bien à nouveau la loi globale 1.1.2. À la lumière de ces résultats, on peut se demander s'il n'est pas possible d'obtenir une loi locale pour les modèles ayant simplement une structure de dépendance partielle. Nous avons répondu positivement à cette question pendant notre thèse, ce qui fait l'objet de la section suivante.

1.2 Équivalent déterministe pour les modèles avec une structure de dépendance partielle

Dans cette partie nous présentons les résultats et idées principales de l'article [Cho22], duquel est directement issu le deuxième chapitre de cette thèse. Il nous semble important de nous arrêter quelques instants auparavant sur les propriétés de concentration des vecteurs et matrices aléatoires en grande dimension, qui seront au cœur des démonstrations.

1.2.1 Concentration lipschitzienne

Le phénomène de concentration de la mesure peut être décrit comme une propriété des objets aléatoires en grande dimension, selon laquelle les fonctions régulières de ces objets apparaissent comme presque constantes autour de leur moyenne. Un exemple typique est celui de la mesure uniforme sur la sphère euclidienne, qui concentre toute sa masse autour de l'équateur à mesure que la dimension tend vers $+\infty$.

Nous avons adopté dans cette thèse le récent formalisme de concentration lipschitzienne de Louart et Couillet ([LC20]), qui est particulièrement adapté pour travailler avec les matrices aléatoires et leurs transformations. Rappelons les définitions de norme spectrale $|||M||| = \sup_{\|x\|=1} \|Mx\|$, norme de Frobenius $\|M\|_F = \mathrm{Tr}(MM^\top)^{1/2} = \left(\sum_{i,j} M_{ij}^2\right)^{1/2}$, et norme du maximum des entrées $\|M\|_{\max} = \sup_{i,j} |M_{ij}|$.

Définition 1.2.1 (cf. Définition 2.3.1)**.** Soit $(E_n, \|\cdot\|)$ une suite d'espaces vectoriels normés de dimension finie, et X_n une suite de vecteurs aléatoires à valeurs dans E_n.

On dit que la suite X_n a la propriété de concentration lipschitzienne avec une suite de diamètres observables $\sigma_n > 0$, s'il existe une constante $C > 0$, indépendante de la dimension, telle que pour toute suite de fonctions 1-lipschitziennes $f_n : (E_n, \|\cdot\|) \to \mathbb{C}$, et pour tous $n \in \mathbb{N}$ et $t \geq 0$:

$$\mathbb{P}(|f_n(X_n) - \mathbb{E}[f_n(X_n)]| \geq t) \leq Ce^{-\frac{1}{C}\left(\frac{t}{\sigma_n}\right)^2}.$$

On note cette propriété $X_n \propto_{\|\cdot\|} \mathcal{E}(\sigma_n)$, ou bien $X_n \propto \mathcal{E}(\sigma_n)$ s'il n'y a pas d'ambiguïté sur le choix de la norme.

La norme $\|\cdot\|$ apparaît de façon cruciale dans cette définition car elle détermine la famille des fonctions 1-Lipschitz pour lesquelles l'inégalité de concentration est vérifiée uniformément. Remarquons que plus une norme est grande, plus elle induit une notion de concentration qui est forte. En particulier la concentration en norme de Frobenius d'une suite de matrices aléatoires est une propriété plus forte que sa concentration en norme spectrale, car la

norme spectrale d'une matrice est toujours inférieure ou égale à sa norme de Frobenius.

Démentons tout de suite une idée fausse que l'on pourrait se faire au sujet de cette notion de concentration : lorsque les dimensions des espaces E_n tendent vers $+\infty$, $X_n \propto_{\|\cdot\|} \mathcal{E}(\sigma_n)$ ne signifie pas que X_n est proche de son espérance en norme, mais seulement que les observations lipschitziennes unidimensionnelles se situent à une distance typique $O(\sigma_n)$ de leur espérance. Plus précisément, $|f_n(X_n) - \mathbb{E}[f_n(X_n)]| \leq \sqrt{\log n} O(\sigma_n)$ p.s., et $\mathrm{Var}[f_n(X_n)] \leq O(\sigma_n^2)$ (Proposition 2.3.3).

On peut en revanche comparer X_n à son espérance en norme, au prix d'un facteur qui dépend de la nature des espaces vectoriels normés. Par exemple, si $X_n \in \mathbb{R}^{p\times n} \propto_{|||\cdot|||} \mathcal{E}(\sigma_n)$, alors $|||X - \mathbb{E}[X]||| \leq O(\sqrt{n+p}\,\sigma_n)$ p.s. (Proposition 2.3.4).

Comme exemple fondamental d'objets aléatoires concentrés, si un vecteur aléatoire $X_n \in \mathbb{R}^n$ a des coordonnées i.i.d. gaussiennes centrées réduites, alors $X_n \propto_{\|\cdot\|} \mathcal{E}(1)$. De manière similaire, une matrice Y à entrées i.i.d. gaussiennes centrées réduites vérifie $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$. On se réfère à [LC20] et [Led01] pour de nombreux autres exemples de vecteurs ou matrices vérifiant ce phénomène de concentration.

Toute transformation continue lipschitzienne préserve la propriété de concentration lipschitzienne, en modifiant éventuellement la suite des diamètres observables si la constante de Lipschitz dépend de la dimension (Proposition 2.3.2).

Pour les matrices de Wishart colorées des Théorèmes 1.1.2 et 1.1.5 par exemple, si les entrées de $X \in \mathbb{R}^{p\times n}$ sont i.i.d. gaussiennes centrées réduites, et que $\Sigma \in \mathbb{R}^{p\times p}$ est une matrice déterministe, symétrique semi-définie positive, alors $Y = \Sigma^{1/2}X \propto_{\|\cdot\|_F} \mathcal{E}(|||\Sigma|||)$ (ceci découle de l'inégalité $\|AB\|_F \leq |||A|||\|B\|_F$).

De même si $f : \mathbb{R} \to \mathbb{R}$ est une fonction continue lipschitzienne, et que $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$, alors la matrice $f(Y)$ obtenue en appliquant la fonction f entrée par entrée vérifie également $f(Y) \propto_{\|\cdot\|_F} \mathcal{E}(1)$. Ceci permettra notamment d'obtenir sans effort des propriétés de concentration pour les modèles de matrices issus de réseaux de neurones artificiels, qui apparaîtront plus tard dans ce manuscrit.

En ce qui concerne les résolvantes de matrices de covariance, il est possible de montrer que l'application :

$$Y \in \mathbb{R}^{p\times n} \mapsto \mathcal{G}_K(z) = (K - zI_p)^{-1} = \left(\frac{1}{n}YY^\top - zI_p\right)^{-1}$$

est continue lipschitzienne par rapport à la norme de Frobenius, avec comme paramètre $\frac{2^{3/2}|z|^{1/2}}{\sqrt{n}\Im(z)^2}$ (Proposition 2.4.3). Ainsi en partant d'une matrice

$Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$, on a automatiquement $\mathcal{G}_K(z) \propto_{\|\cdot\|_F} \mathcal{E}\left(\frac{|z|^{1/2}}{\sqrt{n}\Im(z)^2}\right)$ (Corollaire 2.4.4).

Les développements récents de cette notion de concentration lipschitzienne vont bien au-delà des quelques faits énoncés ici, ou de ceux dont nous avons eu besoin pendant notre thèse. Nous encourageons le lecteur intéressé à poursuivre par la lecture de l'article [LC20], spécifiquement centré sur ces questions.

1.2.2 Résultat principal

Nous pouvons maintenant énoncer la loi locale que nous avons obtenue pour les matrices de covariance empirique avec une structure de dépendance partielle.

Théorème 1.2.1 (cf. Proposition 2.2.4)**.** *Soit $Y \in \mathbb{R}^{p\times n}$ une matrice aléatoire telle que :*

1. *Le rapport des dimensions $\gamma_n = \frac{p}{n}$ est majoré et minoré par des constantes strictement positives.*
2. *Y vérifie la propriété de concentration $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$.*
3. *Les colonnes de Y sont i.i.d. suivant la loi d'un vecteur aléatoire y.*
4. *$\|\mathbb{E}[y]\|$ est bornée, et avec $\Sigma = \mathbb{E}\left[yy^\top\right]$, $|||\Sigma|||$ est bornée.*

Alors avec $\nu = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$ et $\tilde{\nu} = (1-\gamma_n)\cdot\delta_0 + \gamma_n \cdot \nu$, $\tilde{\nu}$ est bien une mesure de probabilité, et la fonction à valeurs matricielles

$$\mathbf{G}(z) = (-z g_{\tilde{\nu}}(z)\Sigma - zI_p)^{-1}$$

est un équivalent déterministe de $\mathcal{G}_K(z)$ au sens suivant : il existe une constante $C > 0$, telle que avec $\kappa(z) = \frac{|z|^{5/2}}{\Im(z)^9}$, pour toute constante fixée $\tau > 0$, uniformément en $z \in \{z \in \mathbb{C}^+ \text{ tels que } \frac{|z|^7}{\Im(z)^{16}} \leq Cn \text{ et } \Im(z) \leq \tau\}$, et uniformément en la matrice déterministe $A \in \mathbb{R}^{p\times p}$ de norme de Frobenius égale à 1 :

$$|\mathrm{Tr}((\mathcal{G}_K(z) - \mathbf{G}(z))A)| \leq O\left(\sqrt{\frac{\log n}{n}}\kappa(z)\right) \quad \text{p.s.}$$

On peut remarquer que la matrice d'équivalent déterministe $\mathbf{G}(z)$ est exactement la même que celle obtenue dans la loi locale 1.1.5. On peut aussi voir que les directions propres de $\mathbf{G}(z)$ sont les mêmes que celles de Σ. Ainsi lorsque Σ est une matrice diagonale, $\mathbf{G}(z)$ reste aussi diagonale, et si de plus Σ est un multiple de l'identité on retrouve une loi locale isotropique.

La structure de notre théorème est également très similaire : le paramètre spectral z peut varier avec n, à condition que $\Im(z)$ ne tende pas trop vite vers

0 (et également ici $|z|/\Im(z)$ puisque nous n'avons pas imposé de condition sur la partie réelle de z).

Le résultat prend bien la forme d'un résultat de concentration, en un sens plus général que les lois globales précédemment énoncées puisqu'il autorise toutes les formes linéaires issues d'une matrice normalisée en norme de Frobenius, et pas seulement les formes quadratiques avec des vecteurs unitaires (ce qui équivaut à considérer des matrices de rang 1).

Enfin l'équivalent est quantitatif, avec des vitesses de convergence typiques en $1/\sqrt{n}$ pour la dimension, et $\kappa(z)$ pour le paramètre spectral, c'est à dire polynomiales en $1/\Im(z)$ et $|z|/\Im(z)$. Nous touchons ici cependant à l'inconvénient principal de notre résultat comparé au Théorème 1.1.5, car notre estimée quantitative n'apparaît absolument pas optimale en z.

Ce résultat est dans la continuité directe des travaux [LC21], où les auteurs obtiennent essentiellement le même équivalent déterministe, sans toutefois l'aspect quantitatif en z, et la possibilité de faire se rapprocher z de l'axe réel lorsque n tend vers $+\infty$ qui sont nos principales contributions.

En revanche, dans l'article [LC21] les colonnes de Y peuvent éventuellement avoir des distributions différentes. Dans ce cas il faut remplacer dans l'équivalent déterministe la transformée de Stieltjes $g_{\tilde{\nu}}(z)$ par une matrice diagonale qui vérifie une équation de point fixe matricielle, généralisant en un sens la définition de la convolution libre multiplicative avec une loi de Marčenko-Pastur.

Il serait du plus grand intérêt d'obtenir l'analogue de notre résultat quantitatif dans ce cas plus général. Cependant la technicité des preuves augmente grandement lorsqu'on est amené à considérer des équations de point fixe matricielles, et nous n'avons pu nous rapprocher de ce but.

Énonçons enfin quelques corollaires immédiats de ce résultat d'équivalent déterministe, tels que ceux présentés au paragraphe 1.1.3. En utilisant l'identité $g_\nu(z) = \frac{1}{p}\mathrm{Tr}\mathbf{G}(z)$ (Proposition 2.6.2), et en choisissant la matrice $A = I_p/\sqrt{p}$, qui est de norme de Frobenius égale à 1, on voit que le Théorème 1.2.1 implique bien la loi globale 1.1.3 correspondant à ce modèle. Il fournit de plus une convergence quantitative pour les transformées de Stieltjes. En conservant les mêmes notations que celles du théorème, on a en effet :

$$|g_K(z) - g_\nu(z)| \leq O\left(\frac{\sqrt{\log n}}{n}\kappa(z)\right) \quad \text{p.s.}$$

Rappelons la définition des mesures spectrales empiriques directionnelles : $\mu_{K,\mathbf{u}} = \sum_{\lambda\in \mathrm{Sp}K}(\mathbf{u}^\top \mathbf{e}_\lambda)^2\delta_\lambda$, où les vecteurs $\mathbf{e}_\lambda$ forment une base orthonormée associée aux valeurs propres λ. Pour tout vecteur unitaire déterministe $\mathbf{u} \in \mathbb{R}^p$,

on a :

$$\left|g_{\mu_{K,\mathbf{u}}}(z) - \mathbf{u}^\top \mathbf{G}(z)\mathbf{u}\right| \leq O\left(\sqrt{\frac{\log n}{n}}\kappa(z)\right) \quad \text{p.s.}$$

On peut enfin obtenir la convergence uniforme suivante sur les entrées de la résolvante : $\|\mathcal{G}_K(z) - \mathbf{G}(z)\|_{\max} \leq O\left(\sqrt{\frac{\log n}{n}}\kappa(z)\right)$ p.s.

1.2.3 Stratégie de preuve

Pour obtenir le Théorème 1.2.1, on se ramène tout d'abord à travailler en espérance. En effet comme vu précédemment la simple hypothèse $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$ implique automatiquement pour la résolvante que $\mathcal{G}_K(z) \propto_{\|\cdot\|_F} \mathcal{E}\left(\frac{|z|^{1/2}}{\sqrt{n}\Im(z)^2}\right)$. Uniformément en la matrice déterministe $A \in \mathbb{R}^{p\times p}$ de norme de Frobenius égale à 1, on a donc :

$$|\mathrm{Tr}((\mathcal{G}_K(z) - \mathbb{E}[\mathcal{G}_K(z)])A)| \leq O\left(\sqrt{\log n}\frac{|z|^{1/2}}{\sqrt{n}\Im(z)^2}\right) \quad \text{p.s.}$$

Puisque $\frac{|z|^{1/2}}{\Im(z)^2} \leq O(\kappa(z))$, pour obtenir un équivalent déterministe il suffit donc d'approcher l'espérance de la résolvante en norme de Frobenius, et plus précisément de montrer que $\|\mathbf{E}[\mathcal{G}_K(z)] - \mathbf{G}(z)\|_F \leq O\left(\frac{\kappa(z)}{\sqrt{n}}\right)$ (Théorème 2.2.3).

Suivant une idée attribuée à Silverstein ([Sil86]), et adaptée dans ce cadre par Louart et Couillet dans [LC18], on cherche une première approximation de $\mathbb{E}[\mathcal{G}_K(z)]$ à partir de la résolvante de Σ, non plus considérée en le paramètre spectral z mais en autre point, dépendant à la fois de z et de n.

Pour exploiter l'indépendance entre les colonnes de Y, on utilise des variantes des matrices de covariance de notre modèle, où on a laissé de côté la première colonne y_1 de la matrice Y (leave-one-out ou LOO en anglais). On définit ainsi $Y_- = (0, y_2, \ldots, y_n)$, $K_- = \frac{1}{n}Y_-Y_-^\top = K - y_1y_1^\top/n$, et $\mathcal{G}_{K_-}(z) = (K_- - zI_p)^{-1}$. On se sert aussi de la co-résolvante, définie en intervertissant Y et $Y^\top$ dans la matrice de covariance empirique : $\check{K} = \frac{1}{n}Y^\top Y$, et $\mathcal{G}_{\check{K}}(z) = \left(\frac{1}{n}Y^\top Y - zI_n\right)^{-1}$.

Nous détaillons les caractéristiques de ces résolvantes alternatives à la section 2.4, notamment leurs propriétés de concentration, et nous pensons que leur utilisation mériterait d'être généralisée dans le cadre des méthodes analytiques en matrices aléatoires.

À l'aide d'identités purement algébriques entre ces objets (formules de Sherman-Morrison, Proposition 2.4.9), l'indépendance entre y_1 et les autres colonnes et donc tous les objets LOO, et enfin des propriétés de concentration des formes quadratiques de type Hanson-Wright (Proposition 2.3.5), on arrive à démontrer que $\mathbb{E}[\mathcal{G}_K(z)]$ est proche de $\left(\frac{z}{\mathfrak{b}(z)}\Sigma - zI_p\right)^{-1}$, où $\mathfrak{b}(z)$ peut être

défini par deux formulations équivalentes : $\mathfrak{b}(z) = z + \frac{z}{n}\mathrm{Tr}\Big(\Sigma\mathbb{E}\Big[\mathcal{G}_{K_-}(z)\Big]\Big)$, ou bien $\mathfrak{b}(z) = \mathbb{E}[-1/\mathcal{G}_{\check{K}}(z)_{11}]$ (Théorème 2.5.3).

La dernière partie de la preuve (section 2.6) consiste à tirer parti de la suite d'approximations suivantes :

$$\begin{aligned}\mathfrak{b}(z) &= z + \frac{z}{n}\mathrm{Tr}\Big(\Sigma\mathbb{E}\Big[\mathcal{G}_{K_-}(z)\Big]\Big)\\ &\approx z + \frac{z}{n}\mathrm{Tr}(\Sigma\mathbb{E}[\mathcal{G}_K(z)])\\ &\approx z + \frac{z}{n}\mathrm{Tr}\left(\Sigma\left(\frac{z}{\mathfrak{b}(z)}\Sigma - zI_p\right)^{-1}\right).\end{aligned}$$

$\mathfrak{b}(z)$ est donc une solution approchée d'une certaine équation de point fixe (Proposition 2.6.4), qui est exactement l'équation vérifiée par $l_{\tilde{\nu}}(z) = -1/g_{\tilde{\nu}}(z)$, l'inverse de la transformée de Cauchy de la mesure $\tilde{\nu}$ (Proposition 2.6.2). On s'attend donc à ce que $\mathfrak{b}(z)$ se rapproche asymptotiquement de $l_{\tilde{\nu}}(z)$.

Ceci est d'ailleurs suggéré par la seconde formulation de $\mathfrak{b}(z)$ utilisant la co-résolvante. Au moins heuristiquement, on a en effet $\mathfrak{b}^{-1} = \mathbb{E}[-1/\mathcal{G}_{\check{K}}(z)_{11}] \approx \mathbb{E}[-1/g_{\check{K}}(z)] \approx -1/g_{\check{\nu}}(z) = l_{\tilde{\nu}}(z)$. Remarquons d'ailleurs que l'équivalent $\mathbf{G}(z)$ s'écrit bien comme $\mathbf{G}(z) = \left(\frac{z}{l_{\tilde{\nu}}(z)}\Sigma - zI_p\right)^{-1}$.

Pour montrer rigoureusement que le pseudo point fixe $\mathfrak{b}(z)$ converge vers le vrai point fixe $l_{\tilde{\nu}}(z)$, on utilise des propriétés de stabilité. L'équation peut en effet s'écrire à l'aide d'une fonction contractante pour une semi-métrique bien choisie sur $\mathbb{C}^+$ (Proposition 2.6.11). En utilisant des théorèmes de convergence quantitatifs adaptés (Lemme 2.6.14), on obtient que $\mathfrak{b}(z) - l_{\tilde{\nu}}(z)$ tend vers 0 (Proposition 2.6.15), et donc finalement que $\mathbb{E}[\mathcal{G}_K(z)]$ est proche de $\mathbf{G}(z)$ avec les bornes quantitatives recherchées en norme de Frobenius.

Signalons d'ailleurs que c'est pour faire fonctionner cet argument de stabilité qu'apparaît la condition selon laquelle $\Im(z)$ ne tend pas trop vite vers 0. En effet, à mesure que $\Im(z)$ se rapproche de 0, la fonction est de moins en moins contractante, et dans le même temps $\mathfrak{b}(z)$ est de moins en moins un point fixe approché de cette fonction. La condition sus-mentionnée est donc nécessaire pour lever toute indétermination.

1.2.4 Convergence en distance de Kolmogorov

Étant donné deux mesures réelles μ et ν, nous notons $\mathcal{F}_\mu$ et $\mathcal{F}_\nu$ leur fonctions de répartition respectives, et nous rappelons la définition de la distance de Kolmogorov entre μ et ν :

$$D(\mu,\nu) = \sup_{t\in\mathbb{R}} |\mathcal{F}_\mu(t) - \mathcal{F}_\nu(t)|.$$

La convergence en distance de Kolmogorov implique la convergence faible des mesures, et il y a même équivalence lorsque la fonction de répartition de la mesure limite est continue höldérienne ([GH03]). Prouver une convergence en distance de Kolmogorov, qui plus est quantitative, est donc un résultat plus précis que la simple convergence des lois globales mentionnées au paragraphe 1.1.2.

De tels résultats ont été démontrés d'abord dans le cas de matrices de Wishart classiques [Bai93], puis dans le cas de matrices de covariance avec une structure de dépendance linéaire, aussi appelées matrices de Wishart colorées [BHZ12]. Dans le cas d'une structure de dépendance partielle en colonnes, nous avons établi cette convergence en distance de Kolmogorov. La vitesse obtenue n'est toutefois pas optimale puisque nous sommes partis de notre équivalent déterministe qui n'était pas optimal en premier lieu.

La convergence en distance de Kolmogorov est une conséquence assez classique de la théorie des matrices aléatoires, dès lors que l'on dispose d'un équivalent quantitatif sur les transformées de Stieltjes entre deux distributions. Une méthodologie assez générale a été développée pour répondre à cette question, d'abord par Bai ([Bai08, Théorème 2.1], Lemme 2.8.1) puis Banna et Mai ([BM20, section 5].

Nous avons complété cette méthode dans notre article pour correspondre à nos hypothèses légèrement différentes, et notamment le fait que notre estimée sur les transformées de Stieltjes n'a pas lieu asymptotiquement sur tout le demi-plan complexe supérieur $\mathbb{C}^+$ (Théorème 2.8.6). Nous obtenons alors le résultat suivant :

Proposition 1.2.2 (cf. Corollaire 2.2.8)**.** *Soit $Y \in \mathbb{R}^{p\times n}$ une matrice aléatoire telle que :*

1. *Le rapport des dimensions $\gamma_n = \frac{p}{n}$ converge vers $\gamma_\infty > 0$.*
2. *Y vérifie la propriété de concentration $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$.*
3. *Les colonnes de Y sont i.i.d. suivant la loi d'un vecteur aléatoire y.*
4. *$\|\mathbb{E}[y]\|$ est bornée, $\Sigma = \mathbb{E}\left[yy^\top\right]$ est inversible, et $|||\Sigma|||$ et $|||\Sigma^{-1}|||$ sont bornées.*
5. *μ_Σ converge faiblement vers une mesure de probabilité μ_∞.*

Alors μ_K converge faiblement p.s. vers $\nu_\infty = \mathrm{MP}(\gamma_\infty) \boxtimes \mu_\infty$. De plus :

$$D(\mu_K, \nu_\infty) \leq D(\mu_\Sigma, \mu_\infty) + O(|\gamma_n - \gamma_\infty|) + O\left(n^{-\frac{1}{70}}\right) \quad \text{p.s.}$$

Signalons au passage que, pour les besoins de notre démonstration, nous avons établi le résultat suivant concernant la distance de Kolmogorov entre deux distributions de Marčenko-Pastur avec des paramètres différents (Lemme 2.8.10) :

$$D(\mathrm{MP}(\gamma), \mathrm{MP}(\gamma')) \leq \frac{|\gamma - \gamma'|}{\max(\gamma, \gamma')}$$

Bien qu'il s'agisse d'une question naturelle et assez facile à résoudre, nous n'avions pu trouver de référence correspondant à ce problème dans la littérature.

Terminons enfin par quelques mots sur la potentielle convergence en distance de Kolmogorov des mesures spectrales empiriques directionnelles, pour lesquelles on dispose également d'une borne quantitative sur $g_{\mu_{K,\mathbf{u}}}(z) - \mathbf{u}^\top \mathbf{G}(z)\mathbf{u}$. À l'aide de la Proposition 1.1.1, il n'est pas difficile de montrer que la fonction de variable complexe $g_{\nu_\mathbf{u}}(z) = \mathbf{u}^\top \mathbf{G}(z)\mathbf{u}$ est bien une transformée de Stieltjes, et donc qu'elle correspond à une mesure de probabilité $\nu_\mathbf{u}$. Pour mettre en œuvre comme précédemment la méthodologie du Théorème 2.8.6, il est cependant nécessaire de vérifier certaines hypothèses de régularité höldérienne sur la fonction de répartition de la distribution spectrale équivalente $\nu_\mathbf{u}$.

Dans le cas de la mesure spectrale empirique classique, ces hypothèses sont facilement vérifiées puisque l'on travaille avec des convolutions multiplicatives libres avec des lois de Marčenko-Pastur (cf. Lemme 2.8.7). Pour les mesures spectrales empiriques directionnelles en revanche, il n'est pas possible de fournir de telles garanties générales. Il faut donc adapter les résultats de convergence en distance de Kolmogorov en fonction du modèle, et notamment de la résolvante de Σ dont est directement issu l'équivalent déterministe $\mathbf{G}(z)$.

1.3 Réseaux de neurones artificiels et liens avec les matrices aléatoires

1.3.1 Principes de fonctionnement des réseaux de neurones artificiels

Les réseaux de neurones artificiels forment une large classe d'algorithmes en intelligence artificielle. Inspirés par le fonctionnement d'un cerveau biologique, ces réseaux sont constitués d'une collection de neurones artificiels qui échangent et traitent des informations. Une tâche complexe peut ainsi être réalisée par une suite d'opérations simples au niveau des neurones du réseau.

Mathématiquement parlant, un neurone artificiel est une fonction $x \mapsto f(Wx)$, où x est un vecteur de données, W est une matrice de poids, et f une fonction réelle appelée fonction d'activation, qui est appliquée coordonnée par coordonnée aux vecteurs. La fonction f joue un rôle central dans le réseau de neurones car elle permet de reproduire des comportements non linéaires. Entre autres fonctions couramment utilisées, on peut citer la fonction unité linéaire rectifiée $\mathrm{ReLU}(x) = \max(x, 0)$, sa version lissée $\mathrm{Softplus}(x) = \log(1 + e^x)$, la fonction sigmoïde $f(x) = 1/(1+e^{-x})$, ou bien encore la fonction arc-tangente.

Nous allons nous concentrer sur les réseaux de neurones à propagation avant, où les connections entre les neurones ne forment pas de cycle. Un tel réseau à L couches transforme un vecteur d'entrée $x_0 \in \mathbb{R}^{d_0}$ en un vecteur de sortie $x_L \in \mathbb{R}^{d_L}$, éventuellement d'une autre dimension voire simple scalaire, de la façon suivante :

$$x_0 \to x_1 = f(W_1 x_0) \to x_2 \cdots \to \cdots \to x_L = f(W_L x_{L-1}),$$

et où la fonction f est appliquée aux coordonnées des vecteurs.

Le postulat fondateur des réseaux de neurones artificiels est que toute tâche complexe peut être réalisée en ajustant les paramètres de poids internes du réseau W_l. Il est en effet possible de montrer théoriquement que de tels réseaux peuvent approcher une large classe de fonctions ([Cyb89]). Tout l'intérêt et la difficulté des réseaux de neurones résident cependant dans le choix des paramètres internes du réseau, qui est communément appelé phase d'apprentissage.

Une manière courante de procéder est de disposer d'une base de données, formée de n vecteurs d'entrée associés à des vecteurs de sortie qui correspondent à la tâche demandée. En reconnaissance d'images par exemple, on pourrait imaginer comme vecteurs d'entrée des photographies de chats ou de vélos en coordonnées RGB, comme vecteurs de sortie des étiquettes "chat" ou "vélo", et la tâche complexe à réaliser consisterait à discriminer la nature animale ou mécanique des images présentées.

Lorsqu'on regroupe les vecteurs d'entrée et de sortie en matrices, on dispose d'une paire $(X_0, Y) \in \mathbb{R}^{d_0 \times n} \times \mathbb{R}^{d_L \times n}$, appelée ensemble d'apprentissage. Notons $\mathcal{F}$ la fonction matricielle du réseau de neurones, c'est à dire la fonction $\mathcal{F} : X_0 \mapsto X_L$, avec :

$$X_0 \to X_1 = f(W_1 X_0) \to X_2 \cdots \to \cdots \to X_L = f(W_L X_{L-1}),$$

et où la fonction f est appliquée aux entrées des matrices. On compare Y à la prédiction $\mathcal{F}(X_0)$ fournie par le réseau de neurones, à l'aide d'une fonction de coût $\mathcal{L}$. On cherche alors à optimiser les paramètres de poids du réseau pour minimiser ce coût sur la base de données. Une fois entraîné, on espère que le réseau sera capable de réaliser correctement la tâche sur de nouvelles données, ce qui est souvent le cas en pratique et explique le succès de tels algorithmes.

De nombreux choix sont possibles à ce stade d'apprentissage, plus ou moins pertinents suivant le domaine d'application. Pour fixer les idées, on peut d'abord choisir aléatoirement tous les paramètres W, et ensuite optimiser les poids à l'aide d'une fonction d'erreur quadratique moyenne avec un terme de pénalisation en crête $\gamma > 0$:

$$\mathcal{L}(W) = \frac{1}{n} \|\mathcal{F}(W, X_0) - Y\|_F^2 + \gamma \|W\|_F^2.$$

Cette optimisation est typiquement réalisée à l'aide d'un algorithme de descente de gradient stochastique, combinée avec une méthode de rétro-propagation du gradient. Le paysage d'apprentissage n'étant pas convexe, de telles méthodes n'ont toutefois pas de garanties de succès théoriques, et un savoir-faire spécialisé est nécessaire pour assurer le bon déroulement de la procédure.

1.3.2 Lois globales pour le modèle du noyau conjugué

Le modèle du noyau conjugué est inspiré par un réseau de neurones à propagation avant, lors de sa phase d'initialisation où les paramètres du réseau sont choisis de manière aléatoire, et dans le régime où toutes les dimensions des matrices tendent vers $+\infty$. Le modèle avec L couches est donc constitué :

— d'une fonction d'activation $f : \mathbb{R} \to \mathbb{R}$,

— d'une matrice de données initiale $X_0 \in \mathbb{R}^{d_0 \times n}$,

— de matrices de poids $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$.

On définit X_l pour $l \in [\![1, L]\!]$ par récurrence avec la normalisation suivante :

$$X_{l+1} = f\left(\frac{W_{l+1} X_l}{\sqrt{d_l}}\right),$$

et où la fonction f est appliquée entrée par entrée aux matrices. Les matrices des noyaux conjugués sont les matrices de covariance empirique associées aux matrices de données de chaque couche :

$$K_l = \frac{X_l^\top X_l}{d_l}.$$

Ce modèle de matrices aléatoires a été popularisé dans l'article [PW17], et est relié à des questions très concrètes concernant les performances des réseaux de neurones artificiels, dont nous reparlerons à la section 1.4.4. Dressons pour l'instant un panorama des résultats théoriques existants.

On désigne par $\mathcal{N}$ la loi d'une variable aléatoire gaussienne centrée réduite, ou cette variable elle-même suivant le contexte.

Théorème 1.3.1 ([BP21], Théorème 2.2)**.** *Dans le cas d'un réseau de neurones à une seule couche, on suppose que :*

1. *Les rapports des dimensions $\gamma_n^{(0)} = \frac{n}{d_0}$ et $\gamma_n^{(1)} = \frac{n}{d_1}$ convergent respectivement vers $\gamma_\infty^{(0)} > 0$ et $\gamma_\infty^{(1)} > 0$.*
2. *La fonction d'activation f est infiniment dérivable et vérifie l'hypothèse suivante (plus forte que l'analycité) : il existe des constantes $A, C, c > 0$ telles que pour tout $A \geq A_0$, $\sup_{t\in[-A,A]} |f^{(n)}(t)| \leq CA^{cn}$. De plus $\mathbb{E}[f(\mathcal{N})] = 0$.*
3. *Les poids W_1 et les données X_0 sont des matrices aléatoires, indépendantes, à entrées i.i.d. et avec des queues de distribution légères dans le sens suivant : $\mathbb{P}(|(W_1)_{11}| \geq t) \leq e^{-Ct^c}$ et $\mathbb{P}(|(X_0)_{11}| \geq t) \leq e^{-Ct^c}$ pour certaines constantes $C, c > 0$.*

Alors la mesure spectrale empirique de K_1 converge faiblement p.s. vers une mesure de probabilité déterministe à support compact μ.

Cette mesure μ sera d'abord caractérisée par sa transformée de Stieltjes, seule solution d'une équation de point fixe polynomiale d'ordre 4 faisant intervenir les paramètres du modèle, et notamment certains moments gaussiens de la fonction f (cf. [BP21, Théorème 2.3]). Peu après, [Pé19, Théorème 1.4] établira que cette distribution est également la distribution spectrale asymptotique d'un autre modèle de covariance empirique, de type information plus bruit. Nous ne nous attardons pas sur ces premières descriptions de la mesure limite, car elles seront englobées et généralisées par le prochain théorème présenté.

Ces résultats sont obtenus par des méthodes combinatoires. Dans le cas de fonctions d'activation polynomiales, trouver les moments de la distribution limite revient à compter certains graphes spécifiques. Ces propriétés asymptotiques sont ensuite étendues à des fonctions non polynomiales par un argument d'approximation.

De telles méthodes ont permis de démontrer d'autres résultats sur le modèle du noyau conjugué, notamment sur le comportement de la plus grande valeur propre dans [LC21], ou bien pour un modèle avec un bruit additif de rang 1 dans [PS21].

Le résultat suivant, obtenu par des méthodes analytiques, réalise une percée majeure dans la compréhension globale du modèle du noyau conjugué. Pour une matrice carrée M, on note $\mathrm{diag}(M)$ la matrice diagonale extraite de M.

Théorème 1.3.2 ([FW20], Théorème 3.4)**.** *Dans le cas d'un réseau de neurones à L couches, on suppose que :*

1. *Les rapports des dimensions $\gamma_n^{(l)} = \frac{n}{d_l}$ convergent respectivement vers $\gamma_\infty^{(l)} > 0$.*
2. *La fonction d'activation f est deux fois dérivable, et ses dérivées sont uniformément bornées. De plus $\mathbb{E}[f(\mathcal{N})] = 0$ et $\mathbb{E}[f(\mathcal{N})^2] = 1$.*
3. *Les poids $W_1, \dots, W_L$ sont aléatoires, indépendants entre eux, à entrées i.i.d. gaussiennes centrées réduites.*
4. *X_0 est aléatoire, indépendant des autres matrices, et approximativement orthonormale dans le sens suivant (cf. [FW20, Définition 3.1]) : $|||K_0|||$ et $\|\mathrm{diag}(K_0) - I_n\|_F$ sont bornées, et $\|K_0 - I_n\|_{\max} = o(n^{-1/4})$.*
5. *μ_{K_0} converge faiblement p.s. vers une mesure de probabilité déterministe $\nu_\infty^{(0)}$.*

Alors pour $l \in [\![1, L]\!]$, les mesures spectrales empiriques de K_l convergent faiblement p.s. vers les mesures de probabilité déterministes $\nu_\infty^{(l)}$, définies par récurrence comme suit, avec $\zeta = \mathbb{E}[f'(\mathcal{N})]$:

$$\nu_\infty^{(l+1)} = \mathrm{MP}(\gamma_\infty^{(l)}) \boxtimes \Big((1-\zeta^2) + \zeta^2 \nu_\infty^{(l)}\Big).$$

La condition dite d'orthonormalité approximative (4.) est vérifiée pour une large classe de matrices X_0 à colonnes indépendantes, ce qui correspond à prendre des vecteurs indépendants en entrée du réseau de neurones (cf. [FW20, Proposition 3.3] et Proposition 3.6.6).

En particulier dans le cas où les entrées de X_0 sont i.i.d., centrées et de variance 1, cette propriété est vérifiée avec $\|K_0 - I_n\|_{\max} = O\Big(\sqrt{\log n / n}\Big) \leq o(n^{-1/4})$ sur une ensemble de grande probabilité, avec comme distribution spectrale asymptotique $\mu_{K_0} \to \nu_\infty^{(0)} = \mathrm{MP}(\gamma_\infty^{(0)})$ p.s. La distribution μ du Théorème 1.3.2 est donc égale à $\mu = \mathrm{MP}(\gamma_\infty^{(1)}) \boxtimes \Big((1-\zeta^2) + \zeta^2 \mathrm{MP}(\gamma_\infty^{(0)})\Big)$. On peut vérifier que ce résultat est cohérent avec les descriptions de la mesure μ évoquées précédemment ([BP21, Théorème 2.3]) et [Pé19, Théorème 1.4]) en utilisant certaines relations entre convolutions libres ([BG10, Théorème 3]).

Le Théorème 1.3.2 peut toutefois s'accommoder de corrélations à l'intérieur des colonnes de X_0, et la distribution $\nu_\infty^{(0)}$ n'est donc pas nécessairement une loi de Marčenko-Pastur.

Mentionnons enfin l'existence de l'article [WZ21], qui établit un résultat similaire dans le cas d'un réseau de neurones à une seule couche, mais où le rapport des dimensions $\gamma_n^{(1)}$ tend vers 0. La limite est alors une convolution libre multiplicative de $\mathrm{MP}(\gamma_\infty^{(0)})$, non plus avec une loi de Marčenko-Pastur, mais avec une loi du semi-cercle.

1.4 Équivalent déterministe pour le modèle du noyau conjugué

Cette section a pour but de présenter les principales conclusions et idées de l'article [Cho23], duquel est directement issu le troisième chapitre de ce manuscrit.

1.4.1 Résultat principal

Pour étudier le modèle du noyau conjugué, nous allons travailler avec les équivalents déterministes issus de notre premier article [Cho22]. Ces estimées sont quantitatives en la dimension n et le paramètre spectral z, mais non optimales en z comme nous avons discuté précédemment. Nous avons donc choisi de ne garder que le caractère polynomial en z de cette convergence, ce qui motive la définition suivante.

Définition 1.4.1 (cf. Définition 3.2.4)**.** Soit $\zeta : \mathbb{N} \times \mathbb{C}^+ \to \mathbb{R}^+$ une fonction et $\epsilon_n > 0$ une suite. On dit que ζ est bornée par ϵ_n en n, et polynomialement en z, s'il existe une constante $\alpha \geq 0$ telle que, pour tout $A > 0$, uniformément en $n \in \mathbb{N}$ et $z \in \{z \in \mathbb{C} \text{ tels que } 0 < \Im(z) \leq A\}$:

$$\zeta(n, z) \leq O\left(\epsilon_n \frac{|z|^\alpha}{\Im(z)^{2\alpha}}\right).$$

On note cette propriété $\zeta(n, z) \leq O_z(\epsilon_n)$.

On peut montrer que cette définition est compatible avec les règles de calcul classiques en analyse : $O_z(\epsilon_n) + O_z(\epsilon'_n) = O_z(\epsilon_n + \epsilon'_n)$ et $O_z(\epsilon_n)O_z(\epsilon'_n) = O_z(\epsilon_n\epsilon'_n)$ (cf.Remarque 3.2.5). Un autre avantage apporté par cette approche est de pouvoir se débarrasser de certaines restrictions sur z qui étaient présentes dans les énoncés d'équivalents déterministes. Nous renvoyons au Lemme 3.2.6 pour les détails techniques concernant cette mécanique, et au Théorème 3.4.5 pour une version simplifiée de la loi locale 1.2.1 sur les modèles de covariance avec une structure de dépendance partielle.

Pour plus de lisibilité dans cette section, nous proposons ci-dessous une version allégée de notre résultat d'équivalent déterministe 3.7.2. Nous mettons de côté volontairement la possibilité d'utiliser différentes fonctions d'activation, d'ajouter à chaque couche des matrices de biais, ou bien encore de suivre précisément l'évolution des erreurs d'approximation.

Rappelons ici les constituants du modèle du noyau conjugué : en partant d'une matrice de données X_0, de poids $W_1, \ldots, W_L$, et d'une fonction réelle

f, on définit les matrices X_l pour $l \in [\![1, L]\!]$ par la formule de récurrence suivante :

$$X_{l+1} = f\left(\frac{W_{l+1}X_l}{\sqrt{d_l}}\right),$$

où la fonction f est appliquée entrée par entrée aux matrices. Les matrices de covariance empirique associées aux vecteurs de sortie de chaque couche sont égales à $K_l = X_l^\top X_l/d_l$, et dans ce contexte appelées noyaux conjugués.

Théorème 1.4.1 (cf. Théorème 3.7.2)**.** *Dans le cas d'un réseau de neurones sans biais à L couches, on suppose que :*

1. *Les rapports des dimensions $\gamma_n^{(l)} = \frac{n}{d_l}$ sont majorés et minorés par des constantes strictement positives.*
2. *La fonction d'activation f est continue lipschitzienne, avec $\mathbb{E}[f(\mathcal{N})] = 0$ et $\mathbb{E}[f(\mathcal{N})^2] = 1$.*
3. *Les poids $W_1, \dots, W_L$ sont aléatoires, indépendants entre eux, à entrées i.i.d. gaussiennes centrées réduites.*
4. *X_0 est aléatoire, indépendante des autres matrices, et approximativement orthonormale dans le sens suivant : $|||K_0|||$ et $\|\mathrm{diag}(K_0) - I_n\|_F$ sont bornées, et $\|K_0 - I_n\|_{\max} = O(\epsilon_n)$ pour une certaine suite $\epsilon_n > 0$. De plus X_0 vérifie la propriété de concentration $X_0 \propto_{\|\cdot\|_F} \mathcal{E}(1)$.*
5. *On dispose d'équivalents déterministes pour la transformée de Stieltjes et la résolvante de K_0. Plus précisément, il existe :*
   - *une suite de mesures de probabilité déterministes $\nu_n^{(0)}$, à support dans $\mathbb{R}^+$, telles que $\left|\mathbb{E}[g_{K_0}(z)] - g_{\nu_n^{(0)}}(z)\right| \le O_z(1/n)$.*
   - *une suite de fonctions $\mathbf{G}_0 : \mathbb{C}^+ \to \mathbb{C}^{n\times n}$, vérifiant $|||\mathbf{G}_0(z)||| \le 1/\Im(z)$, telles que $|||\mathbb{E}[\mathcal{G}_{K_0}(z)] - \mathbf{G}_0(z)||| \le O_z(1/\sqrt{n})$.*

Soient $\zeta_1 = \mathbb{E}[f(\mathcal{N})\mathcal{N}]$, $\zeta_2 = \mathbb{E}[f(\mathcal{N})(\mathcal{N}^2 - 1)]$ et $\zeta_3 = \mathbb{E}[f(\mathcal{N})(\mathcal{N}^3 - \mathcal{N})]$. On définit par récurrence les mesures et fonctions suivantes :

$$\begin{aligned}
\nu_n^{(l)} &= \mathrm{MP}\left(\gamma_n^{(l)}\right) \boxtimes \left((1-{\zeta_1}^2) + {\zeta_1}^2 \nu_n^{(l-1)}\right),\\
\check{\nu}_n^{(l)} &= (1-\gamma_n^{(l)}) \cdot \delta_0 + \gamma_n^{(l)} \cdot \nu_n^{(l)},\\
\mathbf{G}_l(z) &= \frac{-1}{z\zeta_1 g_{\check{\nu}_n^{(l)}}(z)} \mathbf{G}_{l-1}\left(\frac{-1/g_{\check{\nu}_n^{(l)}}(z) - (1-{\zeta_1}^2)}{\zeta_1}\right) && \text{si } \zeta_1 \neq 0,\\
\mathbf{G}_l(z) &= g_{\mathrm{MP}(\gamma_n^{(l)})}(z) I_n && \text{si } \zeta_1 = 0.
\end{aligned}$$

Alors pour tout $l \in [\![1, L]\!]$, les fonctions $g_{K_l}(z)$ et $\mathbf{G}_l(z)$ sont des équivalents déterministes pour la transformée de Stieltjes et la résolvante de K_l respectivement, au sens suivant : uniformément en la matrice déterministe $A \in \mathbb{R}^{n\times n}$

de norme spectrale égale à 1 *:*

$$\left|g_{K_l}(z) - g_{\nu_n^{(l)}}(z)\right| \leq \sqrt{\log n}\, O_z\Big(\sqrt{\log n}/n + \epsilon_n + \sqrt{n}(\zeta_2^2\epsilon_n^2 + \zeta_3^2\epsilon_n^3)\Big) \text{ p.s.}$$
$$|\mathrm{Tr}((\mathcal{G}_{K_l}(z) - \mathbf{G}_l(z))A)| \leq \sqrt{\log n}\, O_z\Big(1/\sqrt{n} + \epsilon_n + n(\zeta_2^2\epsilon_n^2 + \zeta_3^2\epsilon_n^3)\Big) \quad \text{p.s.}$$

La forme que prend cet énoncé et les conséquences que l'on peut en tirer sont similaires à la loi locale 1.2.1, et nous renvoyons donc notre lecteur à la discussion qui suivait ce théorème. Signalons toutefois la différence consistant à normaliser les matrices en norme spectrale plutôt qu'en norme de Frobenius. Ceci est dû à un argument de linéarisation en norme spectrale sur lequel nous reviendrons au paragraphe suivant.

Remarquons aussi que les directions propres de $\mathbf{G}_0(z)$ sont laissées inchangées par la formule de récurrence définissant les équivalents déterministes successifs $\mathbf{G}_l(z)$. En particulier si X_0 se comporte de manière isotropique, c'est à dire si $\mathbf{G}_0(z)$ est un multiple de l'identité, alors on conserve une loi locale isotropique le long des couches du réseau. Ceci apparaît à première vue contre-intuitif, puisque réaliser des opérations entrée par entrée sur les matrices dépend intrinsèquement du choix d'une base de l'espace vectoriel.

Les constantes ζ_i sont appelées coefficients d'Hermite de f (cf. paragraphe 3.3.1). Lorsque f est deux fois dérivable, en utilisant les formules d'intégration par partie gaussiennes on peut montrer que $\zeta_1 = \mathbb{E}[f(\mathcal{N})\mathcal{N}] = \mathbb{E}[f'(\mathcal{N})]$ et $\zeta_2 = \mathbb{E}[f(\mathcal{N})(\mathcal{N}^2 - 1)] = \mathbb{E}[f''(\mathcal{N})]$, ce qui est cohérent avec la notation employée dans le Théorème 1.3.2.

Observons maintenant les termes d'erreur dans cette loi locale. Sous l'hypothèse $\epsilon_n = o((n\log n)^{-1/4})$, la différence des transformées de Stieltjes converge vers 0 à la vitesse $O_z\Big(\log n/n + \sqrt{n\log n}\,\epsilon_n^2\Big)$. On retrouve donc une version quantitative de la loi globale 1.3.2, que nous réécrirons sous la forme d'un corollaire ci-après.

La condition $\epsilon_n = o((n\log n)^{-1/4})$ n'est toutefois pas suffisante pour assurer que la différence entre les résolvantes converge bien vers 0. Pour ceci il faut en effet $\epsilon_n = o(n^{-1/2}(\log n)^{-1/4})$, ou bien des conditions supplémentaires sur les coefficients d'Hermite de f. Si ζ_2 est nul par exemple, ce qui se produit lorsque f est une fonction impaire, alors $\epsilon_n = o(n^{-1/3}(\log n)^{-1/6})$ suffit. Nous renvoyons à la Remarque 3.7.3 pour une exploration détaillée de ces conditions qui mêlent l'orthonormalité approximative de la matrice des données d'entrée et les coefficients d'Hermite de la fonction d'activation.

Corollaire 1.4.2 (cf. Corollaire 3.7.4)**.** *Avec les hypothèses et notations du Théorème 1.4.1, si de plus :*

1. *Les rapports des dimensions $\gamma_n^{(l)}$ convergent vers $\gamma_\infty^{(l)} > 0$.*
2. *$\epsilon_n = o((n\log n)^{-1/4})$.*

3. $\nu_n^{(0)}$ *converge faiblement p.s. vers une mesure de probabilité déterministe* $\nu_\infty^{(0)}$.

Alors pour $l \in [\![1, L]\!]$, *les mesures spectrales empiriques de* K_l *convergent faiblement p.s. vers les mesures de probabilité déterministes* $\nu_\infty^{(l)}$, *définies par la relation de récurrence suivante :*

$$\nu_\infty^{(l)} = \mathrm{MP}\left(\gamma_\infty^{(l)}\right) \boxtimes (1 - {\zeta_1}^2) + {\zeta_1}^2 \nu_\infty^{(l-1)}.$$

De plus, pour une certaine constante $\theta > 0$, *on a les inégalités suivantes en distance de Kolmogorov :*

$$D(\mu_{K_l}, \nu_\infty^{(l)}) \leq O\left(D(\nu_n^{(0)}, \nu_\infty^{(0)}) + \max_{1 \leq k \leq l} |\gamma_n^{(k)} - \gamma_\infty^{(k)}| + 1/n^\theta\right) \quad \text{p.s.}$$

1.4.2 Idées de démonstration

Pour obtenir la loi locale 1.4.1, notre démarche est proche de celle de l'article [FW20] qui avait permis d'obtenir la loi globale 1.3.2. Nous ne travaillons cependant plus au niveau des transformées de Stieltjes, mais au niveau des résolvantes elles-mêmes. Ceci est rendu possible par les développements concernant les matrices aléatoires de covariance que nous avons présentés précédemment, ainsi que par de nouveaux arguments propres au modèle du noyau conjugué.

La preuve consiste essentiellement en quatre étapes : une récurrence sur les propriétés d'un modèle avec une seule couche $X \to f(WX)$, un argument de conditionnement par rapport aux données X, l'utilisation d'un équivalent déterministe pour les matrices avec une structure de dépendance partielle, et enfin un procédé de linéarisation pour certaines matrices de covariance faisant intervenir des fonctions entrée par entrée. Présentons brièvement ces différents points.

Lorsque la matrice de données X est déterministe, les lignes de $f(WX)$ sont i.i.d. suivant la loi d'un vecteur $f(X^\top w)$, où w est un vecteur aléatoire à entrées i.i.d. gaussiennes. Ce modèle rentre donc dans le cadre de notre premier article, et le Théorème 2.2.4 fournit un équivalent déterministe pour la résolvante de $K = f(WX)^\top f(WX)/d$. Cet équivalent déterministe fait intervenir la résolvante de la matrice de covariance $\Sigma = \mathbb{E}\left[f(X^\top w) f(X^\top w)^\top\right]$, à priori difficile à appréhender compte-tenu de la fonction f appliquée entrée par entrée.

On contourne cette difficulté par un argument d'approximation en norme spectrale de Σ. En utilisant les propriétés de l'espace de Hilbert des fonctions L^2 par rapport à la mesure gaussienne, et les propriétés des polynômes d'Hermite qui en forment une base orthonormée, on peut en effet montrer que Σ est

proche de la matrice $\Sigma_{\text{lin}} = (1-\zeta_1^2)I_n + \zeta_1^2 XX^\top$, où $\zeta_1 = \mathbb{E}[f(\mathcal{N})\mathcal{N}]$ est le premier coefficient d'Hermite de f (Proposition 3.3.10). Cette proximité dépend à la fois de l'orthonormalité approximative de X et des premiers coefficients d'Hermite de la fonction d'activation f. On en déduit que les résolvantes de Σ et Σ_{lin} sont proches en norme spectrale, et on peut donc trouver un équivalent déterministe pour la résolvante de K, qui dépende de manière explicite de X (Théorème 3.5.7).

Si maintenant X est aléatoire et indépendante des autres matrices, on peut travailler conditionnellement à X pour utiliser les propriétés du modèle avec une matrice de données déterministe. Cette étape requiert aussi de bien comprendre les propriétés de concentration des objets qui apparaissent dans l'équivalent déterministe, un travail qui est réalisé à la section 3.4. Nous obtenons au passage des estimées sur les transformées de Stieltjes de mesures obtenues par convolution libre multiplicative, ce qui peut représenter un intérêt en lui-même (Théorème 3.4.7).

La dernière étape de notre raisonnement consiste à propager par récurrence les équivalents déterministes le long du réseau de neurones artificiels, en vérifiant que les conclusions pour une couche donnée correspondent bien aux hypothèses requises pour la couche suivante.

1.4.3 Modèle linéaire gaussien équivalent

Considérons le cas particulier où les vecteurs en entrée du réseau de neurones sont i.i.d. Dans cette situation la propriété d'orthonormalité asymptotique de X_0 est vérifiée sur un évènement de grande probabilité avec $\epsilon_n = \sqrt{\log n/n}$, sous des hypothèses assez générales (cf. [FW20, Proposition 3.3] et Proposition 3.6.6).

Si de plus X_0 est concentrée en norme de Frobenius, le Théorème 1.2.1 fournit les équivalents déterministes $g_{\text{MP}(\gamma_n^{(0)})\boxtimes\mu_\Sigma}(z)$ et $\mathbf{G}_0(z)$ pour la transformée de Stieltjes et la résolvante de K_0 respectivement. Ces équivalents déterministes sont les mêmes que ceux correspondant à une matrice de Wishart colorée de covariance Σ, que l'on peut d'ailleurs supposer gaussienne.

Le Théorème 1.4.1 fournit alors les équivalents déterministes $g_{\nu_n^{(l)}}(z)$ et $\mathbf{G}_l(z)$ pour les noyaux conjugués K_l. À l'aide de la version complète de notre résultat pour un modèle incorporant des biais (cf. Remarque 3.7.6), on peut voir que ces objets sont les mêmes que ceux d'un autre réseau de neurones, avec une fonction d'activation linéaire mais où on a rajouté une matrice de biais à chaque couche. Ce nouveau modèle, qui n'implique plus que des matrices gaussiennes et des opérations linéaires sur les matrices, peut être vu comme un modèle linéarisé gaussien équivalent du modèle du noyau conjugué dans le sens précis suivant :

Proposition 1.4.3. *On suppose que :*

1. *Les rapports des dimensions $\gamma_n^{(l)} = \frac{n}{d_l}$ sont majorés et minorés par des constantes strictement positives.*
2. *La fonction d'activation f est continue lipschitzienne, avec $\mathbb{E}[f(\mathcal{N})] = 0$ et $\mathbb{E}[f(\mathcal{N})^2] = 1$.*
3. *Les matrices $W_1, \dots, W_L, Z_0, \dots, Z_L$ sont aléatoires, indépendantes entre elles, à entrées i.i.d. gaussiennes centrées réduites.*
4. *X_0 est aléatoire, indépendante des autres matrices, et $X_0 \propto_{\|\cdot\|_F} \mathcal{E}(1)$. Les colonnes de X_0 sont i.i.d. suivant la loi d'un vecteur aléatoire centré x. Avec $\Sigma = \mathbb{E}\left[xx^\top\right]$, $|||\Sigma|||$ est bornée et $\mathrm{Tr}\Sigma = d_0$.*

On définit $X_0' = \Sigma^{1/2} Z_0$, $\zeta_1 = \mathbb{E}[f(\mathcal{N})\mathcal{N}]$, $\zeta_2 = \mathbb{E}[f(\mathcal{N})(\mathcal{N}^2 - 1)]$, et par récurrence pour $l \in [\![0, L-1]\!]$:

$$X_{l+1} = f\left(\frac{W_{l+1} X_l}{\sqrt{d_l}}\right),$$
$$X'_{l+1} = \zeta_1 \frac{W_{l+1} X'_l}{\sqrt{d_l}} + \sqrt{1 - \zeta_1^2} Z_{l+1}.$$

Alors les modèles de matrices de covariance empirique $K_l = X_l^\top X_l / d_l$ et $K'_l = X_l'^\top X'_l / d_l$ sont équivalents dans le sens suivant :

$$\left| g_{K_l}(z) - g_{K'_l}(z) \right| \leq O_z\left(\frac{\log n}{n}\right) \quad \text{p.s.}$$

Si de plus $\zeta_2 = 0$, alors uniformément en la matrice déterministe $A \in \mathbb{R}^{n \times n}$ de norme spectrale égale à 1 :

$$\left| \mathrm{Tr}\Big(\big(\mathcal{G}_{K_l}(z) - \mathcal{G}_{K'_l}(z) \big) A \Big) \right| \leq O_z\left(\frac{(\log n)^2}{\sqrt{n}}\right) \quad \text{p.s.}$$

Notre modèle de réseau de neurones artificiels, avec une fonction d'activation non linéaire et sans biais, apparaît donc équivalent à un autre modèle, qui ne fait plus intervenir que des matrices gaussiennes et des transformations linéaires, au prix de bruits aléatoires additionnels Z_l. Ce modèle équivalent est polynômial en un certain nombre de matrices aléatoires indépendantes, ce qui pourrait permettre d'utiliser des outils classiques en matrices aléatoires et en probabilités libres (cf. [Cap17]).

Ce phénomène se retrouve dans d'autres contextes en intelligence artificielle, où il est généralement appelé principe d'équivalence gaussienne ([GLR+22]). De manière plus générale, on peut même envisager un parallèle avec le domaine des systèmes dynamiques, où les non-linéarités induisent typiquement un comportement chaotique.

Il faut toutefois rester prudent sur le sens accordé à cette équivalence, qui ne signifie pas que toutes les propriétés des deux modèles sont les mêmes, notamment en ce qui concerne la phase d'apprentissage des paramètres. Au sens strict, nous avons juste prouvé que les transformées de Stieltjes des deux modèles sont proches, et donc que les mesures spectrales empiriques des deux modèles convergent vers la même limite si elle existe.

Si de plus $\zeta_2 = 0$, ce qui se produit notamment lorsque f est une fonction impaire, on sait que les résolvantes des deux modèles sont proches dans l'acceptation donnée ci-dessus. Entre autres conséquences, ces résolvantes sont proches entrée par entrée, et toutes les mesures spectrales empiriques directionnelles des deux modèles convergent également vers les mêmes limites si elle existent.

Signalons enfin que d'autres auteurs retrouvent des similarités avec ce même modèle, mais avec un sens différent accordé à cette équivalence. Dans [BP22] il est démontré que les plus grandes valeurs propres des deux modèles se comportent de la même manière, pour un réseau à une seule couche et dans le cas où $\zeta_2 = 0$ (et sous des hypothèses supplémentaires sur les distributions des entrées et la régularité de la fonction d'activation). Ce résultat est obtenu par une comparaison fine des moments des distributions en utilisant des méthodes combinatoires. Dans le cas $\zeta_2 \neq 0$, un autre modèle équivalent est proposé, que nous n'avons pas pu retrouver par nos méthodes analytiques.

1.4.4 Lien avec l'erreur d'apprentissage du réseau

Considérons le réseau de neurones artificiels suivant, couramment utilisé pour des tâches de type régression. Le réseau consiste en L premières couches avec une fonction d'activation f, à priori non linéaire, et une dernière couche purement linéaire pour produire la matrice de sortie Z :

$$X_0 \to X_1 = f(W_1 X_0/\sqrt{d_0}) \to \cdots \to X_L = f(W_L X_{L-1}/\sqrt{d_{L-1}}) \to Z = W X_L.$$

On dispose aussi d'un ensemble d'apprentissage, qui correspond à un ensemble de paires de données d'entrée X_0 et de cibles de régression Y. La phase d'apprentissage du réseau correspond à trouver les poids convenables pour que le réseau envoie la matrice X_0 sur la cible Y, c'est à dire que $Z(X_0, W_1, \ldots, W_L, W) \approx Y$.

On initialise aléatoirement $W_1, \ldots, W_L$, et par commodité on suppose que la dernière couche du réseau ne change pas les dimensions des matrices, c'est à dire que $W \in \mathbb{R}^{d_L \times d_L}$. On optimise uniquement les poids de la dernière couche, à l'aide d'une fonction d'erreur quadratique moyenne avec un terme

de pénalisation en crête $\gamma > 0$:

$$\begin{aligned}\mathcal{L}(W) &= \frac{1}{d_L}\|Z(X_0, W_1, \dots, W_L, W) - Y\|_F^2 + \gamma\|W\|_F^2 \\ &= \frac{1}{d_L}\left\|W^\top X_L - Y\right\|_F^2 + \gamma\|W\|_F^2.\end{aligned}$$

On peut vérifier facilement que le paramètre de poids optimal est :

$$\begin{aligned}\hat{W} &= \frac{1}{d_L} X_L \left(\frac{1}{d_L} X_L^\top X_L + \gamma I_n\right)^{-1} Y^\top \\ &= \frac{1}{d_L} X_L \mathcal{G}_{K_L}(-\gamma) Y^\top.\end{aligned}$$

L'erreur d'apprentissage quadratique est donc égale à :

$$\begin{aligned}E_{\text{train}} &= \frac{1}{d_L}\left\|Y^\top - \hat{W} X_L^\top\right\|_F^2 \\ &= \frac{\gamma^2}{d_L}\mathrm{Tr}\left(Y^\top Y \mathcal{G}_{K_L}(-\gamma)^2\right) \\ &= -\frac{\gamma^2}{d_L}\mathrm{Tr}\left(Y^\top Y \frac{\partial}{\partial\gamma}\mathcal{G}_{K_L}(-\gamma)\right),\end{aligned}$$

où on a utilisé la relation algébrique générale sur le carré de la matrice résolvante : $\frac{\partial}{\partial z}\mathcal{G}_{K_L}(z) = \mathcal{G}_{K_L}(z)^2$.

L'erreur d'apprentissage du modèle est donc directement reliée aux propriétés de la résolvante de K_L, considérée avec un paramètre spectral réel négatif. Le Théorème 1.4.1 n'est pas directement formulé pour correspondre à cette situation, mais il pourrait être étendu sans problème en ce sens (cf. la version de la loi locale 1.2.1 avec z réel négatif à la section 2.2).

Sous réserve que les hypothèses en usage dans cette section soient vérifiées, et avec les mêmes notations, on obtient l'équivalent déterministe $\mathcal{G}_{K_L}(-\gamma)^2 \approx -\frac{\partial}{\partial\gamma}\mathbf{G}_L(-\gamma)$, dans le sens du Théorème 1.4.1. En particulier :

$$E_{\text{train}} \approx -\frac{\gamma^2}{d_L}\mathrm{Tr}\left(Y^\top Y \frac{\partial}{\partial\gamma}\mathbf{G}_L(-\gamma)\right).$$

Ces dérivées peuvent être explicitées en remontant jusqu'aux fonctions $\mathbf{G}_0(z)$ et $g_{\nu_n^{(0)}}(z)$ à l'aide des formules de récurrence du Théorème 1.4.1.

Comme ces formules ne changent pas les directions propres, on peut d'ailleurs remarquer que l'erreur d'apprentissage dépend directement de l'alignement entre les directions propres de la matrice cible Y, et celles de la première matrice d'équivalent déterministe $\mathbf{G}_0(z)$.

Dans le cas le plus simple dit de mémorisation, on entraîne le réseau sur des paires aléatoires d'entrée et de sortie. La performance d'un réseau à effectuer une telle tâche est une bonne mesure de sa capacité, c'est à dire à quel point il peut reproduire des liens complexes entre données.

Si les cibles de régression sont des matrices i.i.d. normalisées de sorte que $\mathbb{E}\left[Y^\top Y\right] = d_L/n\, I_n$, et indépendantes des autres matrices, alors on a simplement besoin de la transformée de Stieltjes équivalente pour approximer l'erreur d'apprentissage moyenne :

$$\begin{aligned}\mathbb{E}[E_{\text{train}}] &= \mathbb{E}\left[-\frac{\gamma^2}{d_L}\text{Tr}\left(Y^\top Y \frac{\partial}{\partial\gamma}\mathcal{G}_{K_L}(-\gamma)\right)\right] \\ &= -\gamma^2\frac{\partial}{\partial\gamma}\mathbb{E}[g_{K_L}(-\gamma)] \\ &\approx -\gamma^2\frac{\partial}{\partial\gamma}g_{\nu_n^{(L)}}(-\gamma).\end{aligned}$$

Chapitre 2

Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure

This self-contained chapter consists in the article [Cho22], submitted, and available on arXiv.

Abstract

We study sample covariance matrices arising from rectangular random matrices with i.i.d. columns. It was previously known that the resolvent of these matrices admits a deterministic equivalent when the spectral parameter stays bounded away from the real axis. We extend this work by proving quantitative bounds involving both the dimensions and the spectral parameter, in particular allowing it to get closer to the real positive semi-line.

As applications, we obtain a new bound for the convergence in Kolmogorov distance of the empirical spectral distributions of these general models. We also apply our framework to the problem of regularization of random features models in machine learning without Gaussian hypothesis.

Contents

2.1 Introduction **37**
2.2 Setting **40**
2.2.1 Notations and definitions 40
2.2.2 Main results 41
2.2.3 Application to Kolmogorov distances 42
2.2.4 Application to kernel methods 43
2.2.5 Organization of the paper 43
2.3 Concentration framework **45**
2.4 Properties of resolvent matrices **48**
2.4.1 Sample covariance resolvent 48
2.4.2 LOO resolvent and co-resolvent 49
2.5 First deterministic equivalent **53**
2.5.1 General properties of the deterministic equivalents . 53
2.5.2 Introduction of parameters $\mathfrak{a}$ and $\mathfrak{b}$ 53
2.5.3 Proof of the first deterministic equivalent 54
2.6 Second deterministic equivalent **57**
2.6.1 Reformulation as a fixed point problem 57
2.6.2 Stability of $\mathcal{F}$ in the real case 59
2.6.3 Stability of $\mathcal{F}$ in the complex case 60
2.6.4 Proof of the second deterministic equivalent 64
2.7 Proof of the main results **66**
2.8 Bounds in Kolmogorov distance **68**
2.8.1 Bound for fixed measures 68
2.8.2 Asymptotic bound 69
2.8.3 Application to the empirical spectral distribution of sample covariance matrices 70
2.9 Application to kernel methods **74**
2.9.1 Kernel ridge regression 74
2.9.2 Random features method 74
2.9.3 Effective ridge parameter 75

2.1 Introduction

Since the pioneering works in Statistics of [Wis28], an object of great importance to consider is the sample covariance matrix $K = \frac{1}{n}XX^\top$ where X is a rectangular $p \times n$ random matrix made of n independent columns. One is particularly interested in the asymptotic behaviour of the empirical spectral distribution (ESD) $\mu_K = \frac{1}{p}\sum_{\lambda \text{ eigenvalue of } K} \delta_\lambda$ when n and p go to infinity, and the ratio $\frac{p}{n}$ converges to a fixed constant $\gamma > 0$. When all the entries of X are independent and identically distributed (i.i.d.), it can be shown ([MP67], [Sil89]) that μ_K converges almost surely to a deterministic probability measure $\mathrm{MP}(\gamma)$, now called the Marčenko-Pastur distribution with shape parameter γ. From this article on, many works have provided generalizations beyond the i.i.d. case. This is the case for instance in [SB95], where the authors investigated the linearly dependent case. Namely when $X = \Sigma^{\frac{1}{2}}Y$, with Σ a real symmetric matrix and Y filled with i.i.d. entries, μ_K converges almost surely (a.s.) to a deterministic probability measure. This deterministic probability measure can be described as the free multiplicative convolution $\mathrm{MP}(\gamma) \boxtimes \mu_\infty$, where μ_∞ is the limiting ESD of Σ. Under the sole assumption that the columns of X are independent, [BZ08] proved that μ_K still converges a.s. to $\mathrm{MP}(\gamma) \boxtimes \mu_\infty$ where μ_∞ is the limiting ESD of $\Sigma = \mathbb{E}[K]$.

A key object in random matrix theory is the resolvent matrix $\mathcal{G}_K(z) = (K - zI_p)^{-1}$, where z is a complex number called spectral argument. This matrix is related to the Stieltjes transform $g_K(z) = \int_{\mathbb{R}} \frac{1}{t-z}\mu_K(dt)$ via the identity $g_K(z) = \frac{1}{p}\mathrm{Tr}(\mathcal{G}_K(z))$. This integral transform is a powerful tool that characterizes the law and the weak convergence of probability measures. For more advanced applications to the distribution of eigenvalues and eigenvectors, the Stieltjes tranform is not powerful enough and one needs control of the whole resolvent matrix $\mathcal{G}_K(z)$. This was investigated in [BEK+14] in the i.i.d. case, establishing that $\mathcal{G}_K(z)$ is close to $g_K(z)I_p$ with quantitative bounds involving the dimensions and the spectral parameter z. This analysis was later carried to the linearly dependent case in [KY17], showing that $\mathcal{G}_K(z)$ is close to a deterministic matrix $\mathbf{G}(z)$, which is not a multiple of the identity matrix in general. Following the terminology of [HLN07], we call the matrix $\mathbf{G}(z)$ a deterministic equivalent of $\mathcal{G}_K(z)$. In the most general case dealing with independent columns, [LC21] found a similar deterministic equivalent. Remarkably, they consider columns with different distributions, that were not examined in the previous literature.

This last article however did not allow the spectral argument z to vary with the dimensions, and in particular to get closer to the real axis with quantitative bounds. We complete it by quantifying the convergence towards the deterministic equivalent when the underlying random matrix has i.i.d. columns. Our result encompasses two different settings : when z are complex numbers with positive imaginary parts that do not vanish too quickly, and when z are

negative real numbers. In both cases, we prove the existence of polynomial bounds in n, $\Im(z)$ and $|z|$, uniform in the sense that the bounds only involve a few well-identified parameters, and not the whole law of the matrices. Our results also have the additional benefit of explaining the appearance of free convolution operators in the deterministic equivalent. Although not optimal compared to numerical simulations or previous similar statements in the linearly dependent case ([KY17]), our estimates still allow for important new applications.

In random matrix theory, a classic question arising after the existence of a weak limit for the ESD of a given model is to know whether this convergence can be quantified using a distance between probability distributions. [Bai08] first tackled the problem of convergence rate in the i.i.d. case, a result that was later improved in [GT10]. In the linearly dependent case, a similar result was proved in [BHZ12]. We complete this work by proposing a bound in Kolmogorov distance in our setting of matrices with i.i.d. columns.

Such a deterministic equivalent can also prove itself useful for machine learning applications. In the method of random features associated to kernel ridge regression, some natural predictors are systematically biased. Under Gaussian hypothesis, the authors of [JSS+20] proved that this issue could be solved by replacing the ridge parameter of the regression by a different parameter called effective ridge. We are able to extend this result beyond the Gaussian paradigm, showing in a sense the universality of the phenomenon regardless to the law of the random features.

This paper uses a wide variety of techniques. Concentration of measure theory in particular is a powerful tool to deal with large dimensional random objects. A core result in [Led01] highlights the importance of a class of concentrated vectors compatible with Lipschitz mappings. Logically flowing from this idea, we adopt the recent formalism of [LC20] as it is particularly well adapted to deal with random matrices and their transformations. Along with the classic concentration properties, we use generalizations of the so-called Hanson-Wright inequalities ([VW15], [Ada15]), which state the concentration of quadratic forms involving random vectors and matrices.

This paper also uses extensively the theory behind resolvent matrices. In addition to the usual sample covariance resolvent, we introduce two companion objects : the leave-one-out (LOO) resolvent and the co-resolvent. We provide a systematic analysis of the analytical and algebraic joint properties of these objects. In conjunction with concentration, this work allows to streamline some of the technical arguments, in particular we do not resort anymore to conditional concentration arguments. It also helps to better understand the properties of the deterministic equivalent and its links with free probability theory.

Finally to convert estimates on Stieltjes transforms into bounds in Kolmogorov distance, a general methodology was developed in [BM20]. We adapt these techniques to work with looser requirements better adapted to our needs.

2.2 Setting

2.2.1 Notations and definitions

The set of matrices with p lines, n columns, and entries belonging to a set $\mathbb{K}$ is denoted as $\mathbb{K}^{p\times n}$. We use the following norms for vectors and matrices : $\|\cdot\|$ the Euclidean norm, $\|\cdot\|_F$ the Frobenius norm, and $|||\cdot|||$ the spectral norm.

Given $M \in \mathbb{C}^{p\times p}$, we denote respectively by $M^\top$ its transpose, by $M^\dagger$ its transconjugate, and by $\mathrm{Tr}(M) = \sum_{1\leq i\leq p} M_{ii}$ its trace. If M is real and diagonalizable we denote by $\mathrm{Sp}M$ its spectrum, and by $\mu_M = \frac{1}{p}\sum_{\lambda\in \mathrm{Sp}M}\delta_\lambda$ its empirical spectral distribution (ESD). Let us define the sphere $\mathbb{S}^{p-1} = \{\mathbf{u} \in \mathbb{R}^p \text{ such that } \|\mathbf{u}\| = 1\}$. If $\mathbf{u} \in \mathbb{S}^{p-1}$, we also define the eigenvector empirical spectral distribution (VESD) of M in the direction $\mathbf{u}$ as $\mu_{M,\mathbf{u}} = \sum_{\lambda\in \mathrm{Sp}M}\left(\mathbf{u}^\top\mathbf{v}_\lambda\right)^2\delta_\lambda$, where $\mathbf{v}_\lambda \in \mathbb{S}^{p-1}$ are normalized eigenvectors associated to the eigenvalues λ of M.

Let us consider a spectral parameter z, either belonging to $\mathbb{R}^{*-} = (-\infty, 0)$ or $\mathbb{C}^+ = \{\omega \in \mathbb{C} \text{ such that } \Im(\omega) > 0\}$. If M is symmetric positive semi-definite, then $\mathrm{Sp}M \subset \mathbb{R}^+$, and we can define its resolvent $\mathcal{G}_M(z) = (M - zI_p)^{-1}$ and its Stieltjes transform $g_M(z) = \frac{1}{p}\mathrm{Tr}(\mathcal{G}_M(z))$.

Given a probability distribution ν supported on $\mathbb{R}^+$, its Stieltjes transform is $g_\nu(z) = \int_{\mathbb{R}} \frac{\nu(dt)}{t-z}$ where $z \in \mathbb{R}^{*-}$ or $z \in \mathbb{C}^+$. The Stieltjes transform of a matrix is the same as the Stieltjes transform of its ESD : $g_{\mu_M}(z) = g_M(z)$. There is a similar link for the VESD's : $g_{\mu_{M,\mathbf{u}}}(z) = \mathbf{u}^\top\mathcal{G}_M(z)\mathbf{u}$.

We denote by $\mathcal{F}_\nu$ the cumulative distribution function (CDF) of ν, and by $D(\nu,\mu) = \sup_{t\in\mathbb{R}}|\mathcal{F}_\nu(t) - \mathcal{F}_\mu(t)|$ the Kolmogorov distance between two measures ν and μ.

We denote by $\boxtimes$ the free multiplicative convolution of measures (see [BV93]). If $\mathrm{MP}(\gamma)$ denotes a Marčenko-Pastur distribution with shape parameter γ, the distribution $\mathrm{MP}(\gamma)\boxtimes\mu$ may be defined by its Stieltjes transform $g(z)$, which is the unique Stieltjes transform function that solves the self-consistent equation [MP67] :

$$g(z) = \int_{\mathbb{R}} \frac{1}{(1-\gamma-\gamma z g(z))t - z}\mu(dt).$$

Throughout this paper, we say that a property occurs almost surely eventually (a.s.e.) if there is a full measure set on which the property holds true for all but finitely many n. Equivalenty, we ask that there is a full measure set on which the property holds true for all integers n greater than a (possibly random) rank n_0. Such statements typically result from the application of Borel-Cantelli lemma.

For a better readability, we will sometimes omit indices n and parameters z in our notations.

2.2.2 Main results

Let $X \in \mathbb{R}^{p_n \times n}$ be a sequence of random matrices. The associated sample covariance matrix is $K = n^{-1}XX^\top$. We set $\Sigma = \mathbb{E}[K]$, and we also define the resolvent matrix $\mathcal{G}_K(z) = (K - zI_p)^{-1}$.

Assumptions 2.2.1.
1. The columns of X are i.i.d. sampled from the distribution of a random vector x, such that $\|\mathbb{E}[x]\|$ and $|||\Sigma||| = \left|\left|\left|\mathbb{E}[xx^\top]\right|\right|\right|$ are bounded.
2. The sequence X is $\propto \mathcal{E}_2(1)$ concentrated with respect to the Frobenius norm, in the sense that all 1-Lipschitz functions of X satisfy a concentration inequality in $e^{-(\cdot)^2}$ (see Definition 2.3.1 below).
3. $\gamma_n = \frac{p_n}{n}$ is bounded from above and from below : there is a constant $C > 0$ such that $C^{-1} \leq \gamma_n \leq C$.

We start with a simple proposition establishing the concentration of the resolvent of the sample covariance matrix K :

Proposition 2.2.2.
1. *If $z \in \mathbb{R}^{*-}$ are real spectral arguments, then $\mathcal{G}_K(z) \propto \mathcal{E}_2\left(\frac{\tau}{n^{\frac{1}{2}}}\right)$, where $\tau = \frac{1}{|z|^{\frac{3}{2}}}$.*
2. *If $z \in \mathbb{C}^+$ are complex spectral arguments, then $\mathcal{G}_K(z) \propto \mathcal{E}_2\left(\frac{\tau}{n^{\frac{1}{2}}}\right)$, where $\tau = \frac{|z|^{\frac{1}{2}}}{\Im(z)^2}$*

In order to state the main result of this article, we need to introduce the probability measures $\nu_n = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$ and $\check{\nu}_n = (1 - \gamma_n)\delta_0 + \gamma_n \nu_n$. We also build the sequence of matrices :

$$\mathbf{G}(z) = (-z g_{\check{\nu}}(z)\Sigma - zI_p)^{-1},$$

where $z \in \mathbb{R}^{*-}$ or $z \in \mathbb{C}^+$.

Theorem 2.2.3.
1. *If $z \in \mathbb{R}^{*-}$ are real spectral arguments, possibly varying with n, such that $|z|$ is bounded and $|z|^{-7} \leq O(n)$, then $\|\mathbb{E}[\mathcal{G}_K(z)] - \mathbf{G}(z)\|_F \leq O\left(\frac{\kappa}{n^{\frac{1}{2}}}\right)$, where $\kappa = \frac{1}{|z|^{\frac{11}{2}}}$.*
2. *There exists a constant $C > 0$ such that, if $z \in \mathbb{C}^+$ are complex spectral arguments with $\Im(z)$ bounded and $\Im(z)^{-16}|z|^7 \leq Cn$, then $\|\mathbb{E}[\mathcal{G}_K(z)] - \mathbf{G}(z)\|_F \leq O\left(\frac{\kappa}{n^{\frac{1}{2}}}\right)$, where $\kappa = \frac{|z|^{\frac{5}{2}}}{\Im(z)^9}$.*

These results are uniform in the sense that the implicit constants in the $O(\cdot)$ notations only depend on the constants in the hypothesis chosen for X, γ_n, Σ, $\mathbb{E}[x]$ and z.

This result can be combined with the concentration of the resolvent to obtain a so-called deterministic equivalent of the resolvent. We remind our reader that a.s.e. stands for almost surely eventually.

Proposition 2.2.4. *With the same assumptions and notations as in Theorem 2.2.3, $\mathbf{G}(z)$ is a deterministic equivalent for $\mathcal{G}_K(z)$. More precisely, uniformly in any deterministic matrix $A \in \mathbb{R}^{p\times p}$, we have :*

$$|\mathrm{Tr}(\mathcal{G}_K(z)A - \mathbf{G}(z)A)| \leq \|A\|_F \, O\left(\frac{\kappa(\log n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \quad \text{a.s.e.}$$

$$\text{and} \quad \mathrm{Var}\left[\mathrm{Tr}(\mathcal{G}_K(z)A)\right] \leq \|A\|_F^2 \, O\left(\frac{\kappa^2}{n}\right).$$

Specifying the matrix A and using the relation $p^{-1}\mathrm{Tr}\mathbf{G}(z) = g_\nu(z)$, which shall be proved in the course of this article, yields the following consequences :

Corollary 2.2.5. *With the same assumptions and notations as in Theorem 2.2.3, uniformly in any deterministic vector $\mathbf{u} \in \mathbb{S}^{p-1}$, we have :*

$$|g_K(z) - g_\nu(z)| \leq O\left(\frac{\kappa(\log n)^{\frac{1}{2}}}{n}\right) \quad \text{a.s.e.}$$

$$\left|g_{\mu_{K,\mathbf{u}}}(z) - \mathbf{u}^\top \mathbf{G}(z)\mathbf{u}\right| \leq O\left(\frac{\kappa(\log n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \quad \text{a.s.e.}$$

$$\max_{1\leq i,j\leq p} |\mathcal{G}_K(z) - \mathbf{G}(z)|_{ij} \leq O\left(\frac{\kappa(\log n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \quad \text{a.s.e.}$$

Remark 2.2.6. Similar statements can be made about the variances, for instance $\mathrm{Var}[g_K(z)] \leq O\left(\frac{\kappa^2}{n^2}\right)$. The Stieltjes transform of the ESD is better concentrated than the Stieltjes transforms of the VESD's. That was to be expected since the ESD corresponds to an average over a basis of orthonormal vectors.

Finally all these results admit a transposed version for models of random matrices $X' \in \mathbb{R}^{n\times p}$ having i.i.d. sampled rows from a random vector with covariance Σ. In this setting the covariance matrix has to be chosen as $K' = \frac{1}{n}X'^\top X'$ and the equivalent objects are left unchanged.

2.2.3 Application to Kolmogorov distances

Our main results give quantitative bounds for spectral arguments that get close the real axis. This allows to use general methods that bound the Kolmogorov distance of two distributions given a precise control on their Stieltjes transforms. The consequences are the following :

Theorem 2.2.7. *Under the Assumptions 2.2.1 and the additional hypothesis that Σ is invertible and $|||\Sigma^{-1}|||$ is bounded, we have :*

$$D(\mu_K, \nu) \leq O\Big(n^{-\frac{1}{70}}\Big) \quad \text{a.s.e.}$$

Corollary 2.2.8. *If additionally $\mu_\Sigma \to \mu_\infty$ weakly and $\gamma_n \to \gamma_\infty$, then μ_K converges weakly to $\nu_\infty = \mathrm{MP}(\gamma_\infty) \boxtimes \mu_\infty$ a.s. Moreover :*

$$D(\mu_K, \nu_\infty) \leq D(\mu_\Sigma, \mu_\infty) + O(|\gamma_n - \gamma_\infty|) + O\Big(n^{-\frac{1}{70}}\Big) \quad \text{a.s.e.}$$

2.2.4 Application to kernel methods

This second application to the regularization of random feature models in machine learning closely follows the works of [JSS+20]. Random feature (RF) models are used as an efficient approximation of kernel ridge regression (KRR) methods. In the regime on the dimensions where this approximation has a computational interest, a systematic asymptotic bias appears unfortunately.

Under Gaussian assumptions, the authors of [JSS+20] showed an implicit regularization phenomenon : the bias can be removed by replacing λ, the ridge parameter used in the regression, by another parameter $\tilde{\lambda}$. They called this new parameter effective ridge, and characterized it by means of a self-consistent equation.

Using our deterministic equivalent, we extend the aforementioned work when the data are not necessarily Gaussian distributed, and we also explain how the effective ridge parameter $\tilde{\lambda}$ is related to a free multiplicative convolution of measures. We refer to the Section 2.9 for more details on these results.

2.2.5 Organization of the paper

The bulk of this paper is devoted to the proof of the deterministic equivalent Theorem 2.2.3, which is the content of Sections 2.5-2.7.

In Section 2.3 we present the concentration framework and the key concentration results that we shall use throughout the proofs. In Section 2.4 we study thoroughly the properties of the resolvent of sample covariance matrices and two companion objects : the leave-one-out (LOO) resolvent and the co-resolvent.

In Section 2.5 we perform the first step of the proof by establishing an intermediate deterministic equivalent based on algebraic identities. In Section 2.6 we pass from this first equivalent to a second deterministic equivalent corresponding to our main result. The key ingredient of this step consists in rewriting the problem as a fixed point equation, implying a function that is a contraction mapping for a well-chosen metric. The arguments for real and complex spectral parameters differ slightly and are treated separately. In Section 2.7 we wrap up the proofs of Theorem 2.2.3 and its corollaries.

In Section 2.8, as an illustration of the power of our quantitative results, we present an application to the convergence in Kolmogorov distance for general models of sample covariance matrices with i.i.d. columns.

In Section 2.9, we explain how our work is related to the problem of kernel ridge regression (KRR) in machine learning, and we retrieve similar results to those of [JSS+20] under looser hypothesis.

2.3 Concentration framework

In this preliminary section we introduce some basic definitions and results to handle concentrated random vectors and matrices. We adopt the recent formalism of Lipschitz concentration of Louart and Couillet as it is particularly well adapted to deal with random matrices and their transformations. We advice the reader to consult the articles [LC18], [LC20], and [LC21] for a more comprehensive coverage of these notions.

Definition 2.3.1. We say that a sequence of random vectors X_n, belonging to normed vector spaces $(E_n, \|\cdot\|)$, is concentrated with parameter $q > 0$ and observable diameter $\sigma_n > 0$, if there is a constant $C > 0$ such that for any sequence of 1-Lipschitz maps $f_n : E_n \to \mathbb{C}$, for any $n \in \mathbb{N}$ and $t \geq 0$:

$$\mathbb{P}(|f_n(X_n) - \mathbb{E}[f_n(X_n)]| \geq t) \leq Ce^{-\frac{1}{C}\left(\frac{t}{\sigma_n}\right)^q}.$$

We denote by $X_n \propto \mathcal{E}_q(\sigma_n)$ this notion of concentration.

As explained in the beginning of this paper, we will omit most of the n indices for a better readability. We also use the following generalization of the definition : $X_n \propto \mathcal{E}_q(\sigma_n)+\mathcal{E}_{q'}(\sigma'_n)$ means that the above property holds with the right hand side term of the inequality replaced by $Ce^{-\frac{1}{C}\left(\frac{t}{\sigma_n}\right)^q} + Ce^{-\frac{1}{C}\left(\frac{t}{\sigma'_n}\right)^{q'}}$.

In this paper, we mostly consider $\mathbb{R}^n$ endowed with the Euclidean norm, or equivalently $\mathbb{R}^{p\times n}$ endowed with the Frobenius norm. In this setting, any standard Gaussian vector is concentrated independently of its dimension : if the entries of X_n are i.i.d. sampled from a standard Gaussian random variable, then $X_n \propto \mathcal{E}_2(1)$ ([Led01]). The class of Lipschitz concentrated random vectors goes way beyond Gaussian vectors, see [Led01], [Tal95], and [Tao12] for more examples.

We deduce immediately from the definition the following property :

Proposition 2.3.2. *If $X_n \propto \mathcal{E}_q(\sigma_n)$ and φ_n is a sequence of k_{φ_n}-Lipschitz functions on the support of X_n, then $\varphi_n(X_n) \propto \mathcal{E}_q(k_{\varphi_n}\sigma_n)$, with new concentration constants that do not depend on φ_n.*

Let us now mention a simple link between observable diameter and variance for random variables.

Proposition 2.3.3. *If a random variable X_n is $\mathcal{E}_q(\sigma_n)$ concentrated, then $|X_n - \mathbb{E}[X_n]| \leq O\left(\sigma_n(\log n)^{\frac{1}{q}}\right)$ almost surely eventually (a.s.e.) and $\mathrm{Var}[X_n] \leq O(\sigma_n^2)$. These bounds only depend on the concentration constant of X_n, not on the whole law of X_n.*

Proof. For the first assertion, let us choose $t_n = \sigma_n(2C\log n)^{\frac{1}{q}}$ in the concentration inequality. We obtain :

$$\mathbb{P}(|X_n - \mathbb{E}[X_n]| \geq t_n) \leq Ce^{-\frac{1}{C}\left(\frac{t_n}{\sigma_n}\right)^q} = \frac{C}{n^2}.$$

This bound is summable, so Borel-Cantelli lemma implies that $|X_n - \mathbb{E}[X_n]| \leq t_n \leq O\Big(\sigma_n(\log n)^{\frac{1}{q}}\Big)$ a.s.e. This $O(\cdot)$ bound only depends on C, the concentration constant of X_n.

For the second assertion, we integrate the concentration inequality :

$$\begin{aligned}\mathrm{Var}[X_n] &= \int_0^\infty \mathbb{P}\left(|X_n - \mathbb{E}[X_n]|^2 \geq t\right) dt \\ &\leq C\int_0^\infty \exp\left(-\frac{1}{C}\frac{t^{\frac{q}{2}}}{\sigma_n^q}\right) dt \\ &= \frac{2C^{1+\frac{2}{q}}}{q}\Gamma\left(\frac{2}{q}\right)\sigma_n^2.\end{aligned}$$

□

Thanks to a classic argument of ϵ-nets ([Tao12, Section 2.3.1]), we also have a control on the spectral norm of concentrated random matrices :

Proposition 2.3.4. *If a sequence of random matrices $X \in \mathbb{R}^{p\times n}$ is $\propto \mathcal{E}_2(\sigma_n)$ concentrated with respect to the spectral norm, then $|||X - \mathbb{E}[X]||| \leq O((n+p)^{\frac{1}{2}}\sigma_n)$ a.s.e.*

Proof. Upon centering and rescaling, we can assume without loss of generality that X is centered and $\propto \mathcal{E}_2(1)$ concentrated.

Let us consider a covering of the unit Euclidean ball of $\mathbb{R}^p$ with 9^p balls of radius $1/3$, centered in $u_1, \ldots, u_{9^p}$ vectors belonging to the unit Euclidean ball. Similarly let us consider a covering of the unit Euclidean ball of $\mathbb{R}^n$ with 9^n balls of radius $1/3$, centered in $v_1, \ldots, v_{9^n}$ vectors belonging to the unit Euclidean ball. The applications $M \mapsto u_i^\top M v_j$ all are 1-Lipschitz with respect with the spectral norm. From Propositions 2.3.2 and 2.3.3, there is $C > 0$ such that for any i, j and $t \geq 0$:

$$\mathbb{P}\Big(\left|u_i^\top X v_j\right| \geq t\Big) \leq Ce^{-\frac{t^2}{C}}.$$

With $t_n = C^{\frac{1}{2}}((n+p)\log 9 + 2\log n)^{\frac{1}{2}}$, a union bound over the pairs (i,j) shows that :

$$\mathbb{P}\left(\sup_{i,j}|u_i^\top X v_j| \geq t_n\right) \leq C9^{n+p}e^{-\frac{t_n^2}{C}} = \frac{C}{n^2}.$$

This bound is summable, so Borel-Cantelli lemma implies that $\sup_{i,j}|u_i^\top X v_j| \leq t_n$ a.s.e.

For any $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^n$ unit-normed vectors, we can find i and j such that $\|u - u_i\|$ and $\|v - v_j\| \leq \frac{1}{3}$, hence :

$$\begin{aligned}|u^\top X v| &\leq \left|u_i^\top X v_j\right| + \left|(u-u_i)^\top X v_j\right| + \left|u^\top X (v - v_j)\right| \\ &\leq |u_i^\top X v_j| + \|u - u_i\| |||X||| \|v_j\| + \|u^\top\| |||X||| \|v - v_j\| \\ &\leq \sup_{i,j} |u_i^\top X v_j| + \frac{2}{3}|||X|||.\end{aligned}$$

We deduce that $|||X||| = \sup_{\|u\|=\|v\|=1} |u^\top X v| \leq \sup_{i,j} |u_i^\top X v_j| + \frac{2}{3}|||X|||$, thus $|||X||| \leq 3 \sup_{i,j} |u_i^\top X v_j| \leq 3t_n \leq O((n+p)^{\frac{1}{2}})$ a.s.e.

□

The next proposition is sometimes referred to as Hanson-Wright type inequality after the seminal paper [HW71]. It establishes the concentration of some random quadratic forms, and will prove itself to be crucial in the analysis of resolvents. A proof of this result can be found in [Ada15, Theorem 2.4].

Proposition 2.3.5 (Hanson-Wright)**.** *Let $X_n \in \mathbb{R}^n$ be a random vector, such that $X_n \propto \mathcal{E}_2(1)$ with respect to the Euclidean norm, and $\|\mathbb{E}[X_n]\|$ is bounded. Then uniformly in any deterministic matrix $A \in \mathbb{C}^{n \times n}$, we have $X_n^\top A X_n \propto \mathcal{E}_2(\|A\|_F) + \mathcal{E}_1(|||A|||)$, in the sense that the new concentration constants can be taken uniformly in A.*

We wrap up this section with a simple lemma that bounds the variance of a product of random variables when one of them is a.s. bounded.

Proposition 2.3.6 (Variance of a product)**.** *Let X and Y be L^2 complex random variables with $|X|$ a.s. bounded by $\|X\|_\infty > 0$. Then :*

$$\operatorname{Var}[XY] \leq 2\|X\|_\infty^2 \operatorname{Var}[Y] + 2|\mathbb{E}[Y]|^2 \operatorname{Var}[X]$$

Proof. We decompose the product XY into $XY = X(Y - \mathbb{E}[Y]) + X\mathbb{E}[Y]$, and we use the inequality $\operatorname{Var}[A+B] \leq 2\operatorname{Var}[A] + 2\operatorname{Var}[B]$:

$$\begin{aligned}\operatorname{Var}[XY] &= \operatorname{Var}\left[X(Y - \mathbb{E}[Y]) + X\mathbb{E}[Y]\right] \\ &\leq 2\operatorname{Var}\left[X(Y - \mathbb{E}[Y])\right] + 2\operatorname{Var}\left[X\mathbb{E}[Y]\right] \\ &\leq 2\mathbb{E}\left[|X|^2 |Y - \mathbb{E}[Y]|^2\right] + 2|\mathbb{E}[Y]|^2 \operatorname{Var}[X] \\ &\leq 2\|X\|_\infty^2 \operatorname{Var}[Y] + 2|\mathbb{E}[Y]|^2 \operatorname{Var}[X].\end{aligned}$$

□

2.4 Properties of resolvent matrices

This section deals with the resolvent of sample covariance matrices, and introduces two important companion objects : the leave-one-out (LOO) resolvent and the co-resolvent.

Given a matrix $M \in \mathbb{R}^{p\times p}$, the resolvent of M is the matrix function $\mathcal{G}_M(z) = (M - zI_p)^{-1}$, well defined when the spectral parameter $z \in \mathbb{C}$ is not an eigenvalue of M. In the upcoming sections, we will often discard the parameter z from our notations for better readability, especially in the course of technical proofs.

By resolvent identity we refer to the relation between invertible matrices $A-B = A(B^{-1}-A^{-1})B$, and its immediate corollary for resolvents $\mathcal{G}_M-\mathcal{G}_N = \mathcal{G}_M(N-M)\mathcal{G}_N$.

2.4.1 Sample covariance resolvent

If $X \in \mathbb{R}^{p\times n}$ is a rectangular matrix, we set its sample covariance $K = \frac{1}{n}XX^\top$. Given a spectral parameter $z \in \mathbb{R}^{*-}$ or $\mathbb{C}^+$, we define the resolvent matrix $\mathcal{G} = \mathcal{G}_K(z)$. In other terms :

$$\mathcal{G} = (K - zI_p)^{-1} = \left(\frac{1}{n}XX^\top - zI_p\right)^{-1}.$$

To unify the notations we let $\eta = |z|^{-1}$ when $z \in \mathbb{R}^{*-}$ and $\eta = \Im(z)^{-1}$ when $z \in \mathbb{C}^+$. Note that we always have $\eta|z| \geq 1$. Let us establish some elementary properties of the sample covariance resolvent.

Proposition 2.4.1. *$\mathcal{G}$ is well defined, $|||\mathcal{G}||| \leq \eta$ and $|||\mathcal{G}X||| \leq n^{\frac{1}{2}}\left(2^{\frac{1}{2}}\eta|z|^{\frac{1}{2}}\right)$.*

Proof. The eigenvalues of K are real non negative, thus all eigenvalues of $K - zI_p$ are bounded from below by $d(z, \mathbb{R}^+)$, and $K - zI_p$ is invertible with $\left|\left|\left|(K - zI_p)^{-1}\right|\right|\right| = |||\mathcal{G}||| \leq \eta$. From the identity $\mathcal{G}K = I_p + z\mathcal{G}$, we deduce that $|||\mathcal{G}X|||^2 = \left|\left|\left|\left(\mathcal{G}XX^\top\right)\mathcal{G}^\dagger\right|\right|\right| \leq |||\mathcal{G}nK||| \cdot |||\mathcal{G}||| \leq n(1+\eta|z|)\eta \leq 2n\eta^2|z|$. □

In the following proposition $\Im(M)$ denotes the entrywise imaginary part of a complex matrix M.

Proposition 2.4.2. *$\Im(\mathcal{G}) = \Im(z)\mathcal{G}\mathcal{G}^\dagger$ and $\Im(z\mathcal{G}) = \Im(z)K\mathcal{G}\mathcal{G}^\dagger$. In particular $\Im(\mathcal{G})$ and $\Im(z\mathcal{G})$ are positive semi-definite.*

Proof. Using the resolvent identity, $2i\Im(\mathcal{G}) = \mathcal{G} - \overline{\mathcal{G}} = \mathcal{G}(-\overline{z}I_p + zI_p)\overline{\mathcal{G}} = 2i\Im(z)\mathcal{G}\mathcal{G}^\dagger$. Given that $z\mathcal{G} = K\mathcal{G} - I_p$ we also get $\Im(z\mathcal{G}) = K\Im(\mathcal{G}) = \Im(z)K\mathcal{G}\mathcal{G}^\dagger$. The matrices $\mathcal{G}\mathcal{G}^\dagger$ and K commute and are both positive semi-definite, which achieves the proof. □

The sample covariance map $X \mapsto \frac{1}{n}XX^\top$ is not globally Lipschitz with respect to the Frobenius norm when $|||X|||$ is not bounded. However, we will prove that taking the resolvent of the sample covariance matrix is a Lipschitz function, which will greatly simplify all further concentration analysis.

Proposition 2.4.3. *The map* $X \mapsto \mathcal{G}(X) = \left(\frac{1}{n}XX^\top - zI_p\right)^{-1}$ *is globally Lipschitz with respect to the Frobenius norm with parameter* $n^{-\frac{1}{2}}\left(2^{\frac{3}{2}}\eta^2|z|^{\frac{1}{2}}\right)$.

Proof. Let us consider X and $H \in \mathbb{R}^{p\times n}$. We have :

$$\begin{aligned}\mathcal{G}(X) - \mathcal{G}(X+H) &= \mathcal{G}(X)\frac{1}{n}\Big((X+H)(X+H)^\top - XX^\top\Big)\mathcal{G}(X+H)\\ &= \frac{1}{n}\mathcal{G}(X)\Big(XH^\top + H(X+H)^\top\Big)\mathcal{G}(X+H).\end{aligned}$$

We bound this expression in the following way :

$$\begin{aligned}&\|\mathcal{G}(X) - \mathcal{G}(X+H)\|_F\\ &\leq \frac{1}{n}\Big(|||\mathcal{G}(X)X|||\left\|H^\top\right\|_F|||\mathcal{G}(X+H)||| + |||\mathcal{G}(X)|||\|H\|_F\left|\left|\left|(X+H)^\top\mathcal{G}(X+H)\right|\right|\right|\Big)\\ &\leq \frac{1}{n}\Big(n^{\frac{1}{2}}\left(2^{\frac{1}{2}}\eta|z|^{\frac{1}{2}}\right)\|H\|_F\eta + \eta\|H\|_F n^{\frac{1}{2}}\left(2^{\frac{1}{2}}\eta|z|^{\frac{1}{2}}\right)\Big) = n^{-\frac{1}{2}}\left(2^{\frac{3}{2}}\eta^2|z|^{\frac{1}{2}}\right)\|H\|_F.\end{aligned}$$

□

As an immediate consequence, we obtain by Proposition 2.3.2 the following key result for the concentration of sample covariance resolvents :

Corollary 2.4.4. *If* $X \propto \mathcal{E}_2(1)$, *then* $\mathcal{G} \propto \mathcal{E}_2\left(n^{-\frac{1}{2}}\eta^2|z|^{\frac{1}{2}}\right)$.

2.4.2 LOO resolvent and co-resolvent

Let $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$ be the columns of X. For a better readability we denote the first column x_1 simply by x. We define the leave-one-out (LOO) matrix $X_- = (\mathbf{0}_p, x_2, \ldots, x_n)$, the LOO sample covariance $K_- = \frac{1}{n}X_-{X_-}^\top$ and the LOO resolvent $\mathcal{G}_- = \mathcal{G}_{K_-}(z)$. We have the following identities :

$$\begin{aligned}K_- &= \frac{1}{n}\sum_{j\geq 2}x_jx_j^\top = K - \frac{1}{n}xx^\top,\\ \mathcal{G}_- &= (K_- - zI_p)^{-1} = \left(\frac{1}{n}X_-{X_-}^\top - zI_p\right)^{-1}.\end{aligned}$$

We also define the co-sample covariance matrix by swapping X and $X^\top$: $\check{K} = \frac{1}{n}X^\top X$, and the co-resolvent : $\check{\mathcal{G}} = \mathcal{G}_{\check{K}}(z)$. In other terms :

$$\check{\mathcal{G}} = \left(\check{K} - zI_n\right)^{-1} = \left(\frac{1}{n}X^\top X - zI_n\right)^{-1}.$$

The LOO resolvent and the co-resolvent naturally inherit all bounds and concentration properties of the sample covariance resolvent : $|||\mathcal{G}_-|||$ and $\left|\left|\left|\check{\mathcal{G}}\right|\right|\right| \leq \eta$, and if $X \propto \mathcal{E}_2(1)$, then $\mathcal{G}_-$ and $\check{\mathcal{G}} \propto \mathcal{E}_2\Big(n^{-\frac{1}{2}}\eta^2|z|^{\frac{1}{2}}\Big)$. When X has an additional structure of independence between the columns, the first column x is independent from the LOO resolvent $\mathcal{G}_-$, leading to further interesting properties.

Proposition 2.4.5. *Assume that $X \in \mathbb{R}^{p\times n}$ has i.i.d. columns, that X is $\propto \mathcal{E}_2(1)$ concentrated with respect to the Frobenius norm, and that $\|\mathbb{E}[x]\|$ and $\left|\left|\left|\mathbb{E}\left[xx^\top\right]\right|\right|\right|$ are bounded. Let $\sigma^2 = \eta^2(1+n^{-1}\eta^2|z|)$. Then uniformly in any deterministic matrix $B \in \mathbb{C}^{p\times p}$ with $\|B\|_F \leq 1$:*

$$\mathrm{Var}\left[x^\top \mathcal{G}_- x\right] \leq O\Big(p\sigma^2\Big), \tag{2.4.1}$$

$$\mathrm{Var}\left[x^\top B\mathcal{G}_- x\right] \leq O\Big(\sigma^2\Big), \tag{2.4.2}$$

$$\mathrm{Var}\left[x^\top \mathcal{G}_- B\mathcal{G}_- x\right] \leq O\Big(\eta^2\sigma^2\Big). \tag{2.4.3}$$

Remark 2.4.6. These variance bounds correspond exactly to the ones we would obtain by considering that $\mathcal{G}_-$ is deterministic and applying the concentration of quadratic forms given by Proposition 2.3.5. The full concentration property of these quantities remains however unclear.

Proof. Let $\Sigma = \mathbb{E}\left[xx^\top\right]$. Remark that $\Sigma = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n x_i x_i^\top\right] = \frac{1}{n}\mathbb{E}\left[XX^\top\right] = \mathbb{E}[K]$. Without loss of generality we can assume that $\|\mathbb{E}[x]\| \leq 1$ and $|||\Sigma||| \leq 1$. Let us first prove in full details the second inequality.

For any deterministic matrix $M \in \mathbb{R}^{p\times p}$ with $\|M\|_F \leq \eta$, from Proposition 2.3.5 we have $x^\top Mx \propto \mathcal{E}_2(\eta) + \mathcal{E}_1(\eta)$, thus $\mathrm{Var}\left[x^\top Mx\right] \leq O(\eta^2)$. If B is deterministic with $\|B\|_F \leq 1$, then $\|B\mathcal{G}_-\|_F \leq \eta$ and using independence :

$$\begin{aligned}
\mathbb{E}\left[\left|x^\top B\mathcal{G}_- x - \mathrm{Tr}(B\mathcal{G}_-\Sigma)\right|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left|x^\top B\mathcal{G}_- x - \mathrm{Tr}(B\mathcal{G}_-\Sigma)\right|^2 \,\middle|\, \mathcal{G}_-\right]\right] \\
&= \mathbb{E}_{\mathcal{G}_-}\left[\mathbb{E}_x\left[\left|x^\top Mx - \mathrm{Tr}(M\Sigma)\right|^2\right](M = B\mathcal{G}_-)\right] \\
&= \mathbb{E}_{\mathcal{G}_-}\left[\mathrm{Var}_x\left[x^\top Mx\right](M = B\mathcal{G}_-)\right] \leq O\Big(\eta^2\Big).
\end{aligned}$$

On the other hand $M \mapsto \mathrm{Tr}(BM\Sigma)$ is 1-Lipschitz with respect to the Frobenius norm, therefore by Proposition 2.3.2 $\mathrm{Tr}(B\mathcal{G}_-\Sigma) \propto \mathcal{E}_2\Big(n^{-\frac{1}{2}}\eta^2|z|^{\frac{1}{2}}\Big)$, and by Proposition 2.3.3 $\mathrm{Var}[\mathrm{Tr}(B\mathcal{G}_-\Sigma)] \leq O(n^{-1}\eta^4|z|)$. We combine both

estimates to obtain the inequality (2.4.2) :

$$\begin{aligned}\mathrm{Var}\left[x^\top B\mathcal{G}_- x\right] &= \mathbb{E}\Big[\Big|x^\top B\mathcal{G}_- x - \mathrm{Tr}(B\mathbb{E}[\mathcal{G}_-]\Sigma)\Big|^2\Big] \\ &\leq 2\mathbb{E}\Big[\Big|x^\top B\mathcal{G}_- x - \mathrm{Tr}(B\mathcal{G}_-\Sigma)\Big|^2\Big] \\ &\quad + 2\mathbb{E}\Big[|\mathrm{Tr}(B\mathcal{G}_-\Sigma) - \mathrm{Tr}(B\mathbb{E}[\mathcal{G}_-]\Sigma)|^2\Big] \\ &\leq O\Big(\eta^2\Big) + O\Big(n^{-1}\eta^4|z|\Big) = O(\sigma^2).\end{aligned}$$

At no point did the $O(\cdot)$ notations depend on B in the estimates, as explained in the concentration section.

The first inequality (2.4.1) can be seen as a consequence of the second inequality, applied to $B = p^{-\frac{1}{2}} I_p$ which satisfies $\|B\|_F \leq 1$. The proof of the third inequality (2.4.3) is quite similar to the first one. For this reason we will only point out the key differences. Using the deterministic bound $\|\mathcal{G}_- B\mathcal{G}_-\|_F \leq \eta^2$ we get :

$$\mathbb{E}\Big[\Big|x^\top \mathcal{G}_- B\mathcal{G}_- x - \mathrm{Tr}(\mathcal{G}_- B\mathcal{G}_-\Sigma)\Big|^2\Big] \leq O\Big(\eta^4\Big).$$

On the other hand the map $M \mapsto \mathrm{Tr}(MBM\Sigma) = \mathrm{Tr}(BM\Sigma M)$ is η- Lipschitz on the set $\{M \in \mathbb{C}^{p\times p} \text{ with } |||M||| \leq \eta\}$. Indeed :

$$\begin{aligned}\|M\Sigma M - M'\Sigma M'\|_F &\leq |||M|||\,|||\Sigma|||\,\|M - M'\|_F + \|M - M'\|_F\,|||\Sigma|||\,|||M'||| \\ &\leq \eta\|M - M'\|_F.\end{aligned}$$

This implies that $\mathrm{Tr}(\mathcal{G}_- B\mathcal{G}_-\Sigma) \propto \mathcal{E}_2\Big(n^{-\frac{1}{2}}\eta^3|z|^{\frac{1}{2}}\Big)$ and $\mathrm{Var}[\mathrm{Tr}(\mathcal{G}_- B\mathcal{G}_-\Sigma)] \leq O(n^{-1}\eta^6|z|)$. Combining the previous two estimates proves the inequality (2.4.3). □

We wrap up this section by mentioning some links between the resolvent, the LOO resolvent and the co-resolvent. We need the following linear algebra results :

Lemma 2.4.7 (Sherman-Morrison)**.** *Given an invertible matrix $M \in \mathbb{C}^{p\times p}$ and two vectors $u, v \in \mathbb{C}^p$, $M + uv^\top$ is invertible if and only if $1 + v^\top M^{-1}u \neq 0$, in which case the following identities hold true :*

$$\begin{aligned}\Big(M + uv^\top\Big)^{-1} &= M^{-1} - \frac{M^{-1}uv^\top M^{-1}}{1 + v^\top M^{-1}u}, \\ \Big(M + uv^\top\Big)^{-1}u &= \frac{M^{-1}u}{1 + v^\top M^{-1}u}.\end{aligned}$$

Lemma 2.4.8. $\frac{1}{n}X^\top \mathcal{G} X = I_n + z\check{\mathcal{G}}$.

Proof. Since $X^\top \left(\frac{1}{n} XX^\top - zI_p\right) = \left(\frac{1}{n} X^\top X - zI_n\right) X^\top$, we have $\frac{1}{n} X^\top \mathcal{G} X = \frac{1}{n} \left(\frac{1}{n} X^\top X - zI_n\right)^{-1} X^\top X = I_n + z\check{\mathcal{G}}$. □

As an immediate application of these lemmas, we get the following identities, which we will refer to as LOO identities. These will be extremely useful to exploit any column independence structure in the random matrix X, and subsequently to prove bounds and concentration properties that otherwise would have been far from obvious.

Proposition 2.4.9 (LOO identities)**.** $1 + \frac{1}{n} x^\top \mathcal{G}_- x \neq 0$, *and :*

$$\begin{aligned}
\frac{1}{n} x^\top \mathcal{G} x &= 1 - \frac{1}{1 + \frac{1}{n} x^\top \mathcal{G}_- x} = 1 + z\check{\mathcal{G}}_{11}, \\
\mathcal{G} &= \mathcal{G}_- - \frac{1}{n} \frac{\mathcal{G}_- x x^\top \mathcal{G}_-}{1 + \frac{1}{n} x^\top \mathcal{G}_- x} = \mathcal{G}_- + \frac{z}{n} \check{\mathcal{G}}_{11} \mathcal{G}_- x x^\top \mathcal{G}_-, \\
\mathcal{G} x &= \frac{\mathcal{G}_- x}{1 + \frac{1}{n} x^\top \mathcal{G}_- x} = -z\check{\mathcal{G}}_{11} \mathcal{G}_- x.
\end{aligned}$$

2.5 First deterministic equivalent

2.5.1 General properties of the deterministic equivalents

We place ourselves under the Assumptions 2.2.1 for the matrix X, and we set the sample covariance matrix $K = n^{-1}XX^\top$. For a better readability we will denote the first column of X as $x_1 = x$. The other columns are i.i.d. sampled with the same law as x, hence $\Sigma = \mathbb{E}[K] = \mathbb{E}[xx^\top]$. Without loss of generality, we can assume that $\|\mathbb{E}[x]\| \leq 1$ and $|||\Sigma||| \leq 1$.

We introduce respectively the classical resolvent $\mathcal{G} = \left(\frac{XX^\top}{n} - zI_p\right)^{-1}$, the leave-one-out (LOO) resolvent $\mathcal{G}_- = \left(\frac{XX^\top - xx^\top}{n} - zI_p\right)^{-1}$ and the co-resolvent $\check{\mathcal{G}} = \left(\frac{X^\top X}{n} - zI_n\right)^{-1}$ associated to X. An in-depth analysis of these resolvent matrices may be found in Section 2.4.

The general form of our deterministic equivalents is given by :

$$\mathbf{G}^{\mathfrak{l}} = \left(\frac{z}{\mathfrak{l}}\Sigma - zI_p\right)^{-1},$$

for some sequence of deterministic parameters $\mathfrak{l} \in \mathbb{C}$ that may vary with z.

We will first give a sufficient condition on $\mathfrak{l}$ so that $\mathbf{G}^{\mathfrak{l}}$ is well defined and bounded. To this end we define $\Omega = (-\infty, z]$ if $z \in \mathbb{R}^{*-}$, and $\Omega = \{\mathfrak{l} \in \mathbb{C}$ such that $\Im(\mathfrak{l}) \geq \Im(z)$ and $\Im(z^{-1}\mathfrak{l}) \geq 0\}$ if $z \in \mathbb{C}^+$. We also recall the definition of $\eta = |z|^{-1}$ if $z \in \mathbb{R}^{*-}$ and $\eta = \Im(z)^{-1}$ if $z \in \mathbb{C}^+$.

Lemma 2.5.1. *If $\mathfrak{l} \in \Omega$, then $\mathbf{G}^{\mathfrak{l}}$ is well defined, $\left|\left|\left|\mathbf{G}^{\mathfrak{l}}\right|\right|\right| \leq \eta$ and $\left|\left|\left|\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\right|\right|\right| \leq \eta$.*

Proof. The eigenvalues of $\frac{z}{\mathfrak{l}}\Sigma - zI_p$ are $\frac{z}{\mathfrak{l}}\lambda - z$ for $\lambda \in \mathrm{Sp}\Sigma \subset \mathbb{R}^+$. If $z \in \mathbb{R}^{*-}$, then $\frac{z}{\mathfrak{l}} \geq 0$ and $\frac{z}{\mathfrak{l}}\lambda - z \geq |z|$. If $z \in \mathbb{C}^+$, then $\Im\left(\frac{z}{\mathfrak{l}}\right) \leq 0$ and $\Im\left(\frac{z}{\mathfrak{l}}\lambda - z\right) \leq -\Im(z)$. In both cases, all the eigenvalues of $\frac{z}{\mathfrak{l}}\Sigma - zI_p$ are greater or equal than η^{-1} in modulus, hence $\mathbf{G}^{\mathfrak{l}} = \left(\frac{z}{\mathfrak{l}}\Sigma - zI_p\right)^{-1}$ is well defined and $\left|\left|\left|\mathbf{G}^{\mathfrak{l}}\right|\right|\right| \leq \eta$.

The argument for $\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}$ is similar : the eigenvalues of $\Sigma - \mathfrak{l}I_p$ are $\lambda - \mathfrak{l}$ for $\lambda \in \mathrm{Sp}\Sigma$. If $z \in \mathbb{R}^{*-}$ then $\mathfrak{l} \leq z$ and $\lambda - \mathfrak{l} \geq |z|$. If $z \in \mathbb{C}^+$ then $\Im(\mathfrak{l}) \geq \Im(z)$ and $\Im(\lambda - \mathfrak{l}) \leq -\Im(z)$. In both cases, all the eigenvalues of $\Sigma - \mathfrak{l}I_p$ are greater or equal than η^{-1} in modulus, hence $\left|\left|\left|\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\right|\right|\right| = \left|\left|\left|(\Sigma - \mathfrak{l}I_p)^{-1}\right|\right|\right| \leq \eta$. □

2.5.2 Introduction of parameters $\mathfrak{a}$ and $\mathfrak{b}$

Let us consider $\mathfrak{a} = z + \frac{z}{n}x^\top\mathcal{G}_-x$ and $\mathfrak{b} = \mathbb{E}[\mathfrak{a}] = z + \frac{z}{n}\mathrm{Tr}(\Sigma\mathbb{E}[\mathcal{G}_-])$. In the rest of this section we will prove that $\mathbf{G}^{\mathfrak{b}}$ is close to $\mathbb{E}[\mathcal{G}]$. The precise statement constitutes Theorem 2.5.3.

We can rewrite the LOO identities of Proposition 2.4.9 in the following fashion :

$$\begin{aligned}\frac{1}{n}x^\top \mathcal{G} x = 1 - \frac{z}{\mathfrak{a}} &= 1 + z\check{\mathcal{G}}_{11},\\ \mathcal{G} &= \mathcal{G}_- - \frac{z}{\mathfrak{a}}\frac{1}{n}\mathcal{G}_- x x^\top \mathcal{G}_-,\\ \mathcal{G}x &= \frac{z}{\mathfrak{a}}\mathcal{G}_- x.\end{aligned}$$

In particular $\mathfrak{a} = -\check{\mathcal{G}}_{11}^{-1}$. We also deduce the following properties :

Proposition 2.5.2. *$\mathfrak{a}$ and $\mathfrak{b} \in \Omega$. $\mathbf{G}^{\mathfrak{b}}$ is well defined, $|\mathfrak{a}^{-1}| \leq \eta$ and $\left\|\left|\frac{z}{\mathfrak{b}}\mathbf{G}^{\mathfrak{b}}\right|\right\| \leq \eta$.*

Proof. If $z \in \mathbb{R}^{*-}$, $\mathcal{G}_-$ is a real positive semi-definite matrix and $\frac{\mathfrak{a}}{z} = 1 + \frac{1}{n}x^\top \mathcal{G}_- x \geq 1$, hence $\mathfrak{a} \leq z$. If $z \in \mathbb{C}^+$, $\Im(\mathcal{G}_-)$ and $\Im(z\mathcal{G}_-)$ are positive semi-definite matrices using Proposition 2.4.2, thus $\Im(\mathfrak{a}) = \Im(z) + \frac{1}{n}x^\top \Im(z\mathcal{G}_-)x \geq \Im(z)$, and $\Im\left(\frac{\mathfrak{a}}{z}\right) = \frac{1}{n}x^\top \Im(\mathcal{G}_-)x \geq 0$. In any case, $\mathfrak{a} \in \Omega$, and Ω is stable through expectation so $\mathfrak{b} \in \Omega$ as well. The last assertions are immediate consequences of Lemma 2.5.1 and the definition of Ω. □

2.5.3 Proof of the first deterministic equivalent

Let us recall the definition of $\eta = |z|^{-1}$ if $z \in \mathbb{R}^{*-}$ and $\eta = \Im(z)^{-1}$ if $z \in \mathbb{C}^+$. We can now prove the following result :

Theorem 2.5.3 (First equivalent)**.** *Let $z \in \mathbb{R}^{*-}$ or $\mathbb{C}^+$ be spectral arguments, such that $\eta^{10}|z|^3 \leq O(n)$, and $|z| \leq O(1)$ in the real case, or $\Im(z) \leq O(1)$ in the complex case. Then :*

$$\left\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{b}}\right\|_F \leq O\left(\frac{\eta^5|z|^{\frac{3}{2}}}{n^{\frac{1}{2}}}\right)$$

Remark 2.5.4. In the case where z stays bounded away from $\mathbb{R}^+$, the theorem boils down to a $O(n^{-\frac{1}{2}})$ estimate, which was already obtained in [LC21, Theorem 4].

The assumption $\eta^{10}|z|^3 = O(n)$ is not strictly necessary to obtain this result, but it greatly simplifies the explicit bound. Indeed we can always obtain a $O\left(n^{-\frac{1}{2}}\kappa\right)$ bound, where $\sigma^2 = \eta^2(1 + n^{-1}\eta^2|z|)$ and $\kappa = \sigma^2\eta^2|z| + \sigma\eta^4|z|^{\frac{3}{2}}$. However $\kappa \geq \eta^5|z|^{\frac{3}{2}}$, so the above bound will not be interesting unless we have at least $\eta^{10}|z|^3 = O(n)$. The other assumption can be rewritten as $1 \leq O(\eta)$, which simplifies the results and is not restrictive for practical applications. Under these two hypothesis, $\kappa = O(\eta^5|z|^{\frac{3}{2}})$ which gives the same estimate.

The proof of Theorem 2.5.3 can be organized in three steps :

First step :

We introduce $\mathcal{G}_-$ using the LOO identities :

$$\begin{aligned}\mathcal{G}\left(\frac{z}{\mathfrak{b}}\Sigma - xx^\top\right) &= \mathcal{G}\frac{z}{\mathfrak{b}}\Sigma - \frac{z}{\mathfrak{a}}\mathcal{G}_- xx^\top \\ &= \frac{z}{\mathfrak{b}}\left((\mathcal{G}-\mathcal{G}_-)\Sigma + \mathcal{G}_-\left(\Sigma - xx^\top\right) + \left(1-\frac{\mathfrak{b}}{\mathfrak{a}}\right)\mathcal{G}_- xx^\top\right).\end{aligned}$$

$XX^\top$ is the sum of n iid copies of $xx^\top$, and $\mathbb{E}\left[\mathcal{G}_-\left(\Sigma - xx^\top\right)\right] = 0$ by independence, leading to :

$$\begin{aligned}\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{b}} &= \mathbb{E}\left[\mathcal{G}\left(\frac{z}{\mathfrak{b}}\Sigma - \frac{1}{n}XX^\top\right)\mathbf{G}^{\mathfrak{b}}\right] \\ &= \mathbb{E}\left[\mathcal{G}\left(\frac{z}{\mathfrak{b}}\Sigma - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\mathbf{G}^{\mathfrak{b}}\right] \\ &= \mathbb{E}\left[\mathcal{G}\left(\frac{z}{\mathfrak{b}}\Sigma - xx^\top\right)\right]\mathbf{G}^{\mathfrak{b}} \\ &= \frac{z}{\mathfrak{b}}\left(\mathbb{E}\left[\mathcal{G}-\mathcal{G}_-\right]\Sigma + \mathbb{E}\left[\frac{\mathfrak{a}-\mathfrak{b}}{\mathfrak{a}}\mathcal{G}_- xx^\top\right]\right)\mathbf{G}^{\mathfrak{b}}.\end{aligned}$$

Second step :

We estimate the above quantities using concentration and the properties of LOO quadratic forms.

First note that the term σ appearing in Proposition 2.4.5 boils down to $\sigma = O(\eta)$ under our assumptions. Indeed $\eta^2|z| \leq \eta^{\frac{5}{2}}|z|^{\frac{3}{2}} \leq O(\eta^5|z|^{\frac{3}{2}}) \leq O(n^{\frac{1}{2}})$, so $\sigma^2 = \eta^2(1+n^{-1}\eta^2|z|) = O(\eta^2)$.

We have estimates on the variance of both $\mathfrak{a}$ and $\mathfrak{a}^{-1}$: indeed $\mathrm{Var}[\mathfrak{a}] = |z|^2 n^{-2}\,\mathrm{Var}\left[x^\top\mathcal{G}_- x\right] \leq O(n^{-1}\eta^2|z|^2)$ from Proposition 2.4.5, and $\mathfrak{a}^{-1} = -\check{\mathcal{G}}_{11} \propto \mathcal{E}_2\left(n^{-\frac{1}{2}}\eta^2|z|^{\frac{1}{2}}\right)$ so $\mathrm{Var}\left[\mathfrak{a}^{-1}\right] \leq O(n^{-1}\eta^4|z|)$.

Now let us consider a matrix $B \in \mathbb{C}^{p\times p}$ with $\|B\|_F \leq 1$. We have $\mathrm{Tr}\left(B\mathbb{E}[\mathcal{G}-\mathcal{G}_-]\right) = \mathbb{E}\left[-\frac{z}{\mathfrak{a}}\frac{1}{n}\mathrm{Tr}\left(B\mathcal{G}_- xx^\top\mathcal{G}_-\right)\right] = -\frac{z}{n}\mathbb{E}\left[\frac{\zeta}{\mathfrak{a}}\right]$ with $\zeta = x^\top\mathcal{G}_- B\mathcal{G}_- x$.

We have $|\mathbb{E}[\zeta]| = |\mathrm{Tr}(\mathbb{E}[\mathcal{G}_- B\mathcal{G}_-]\Sigma)| \leq \eta^2 p^{\frac{1}{2}} \leq O(n^{\frac{1}{2}}\eta^2)$, and from Proposition 2.4.5 $\mathrm{Var}\left[\zeta\right] \leq O(\eta^4)$ uniformly in B. Using the deterministic bound $|\mathfrak{a}^{-1}| \leq \eta$, we have :

$$\begin{aligned}\left|\mathbb{E}\left[\mathfrak{a}^{-1}\zeta\right]\right| &\leq \left|\mathbb{E}\left[\mathfrak{a}^{-1}(\zeta - \mathbb{E}[\zeta])\right]\right| + \left|\mathbb{E}[\mathfrak{a}^{-1}]\mathbb{E}[\zeta]\right| \\ &\leq \eta\,\mathrm{Var}[\zeta]^{\frac{1}{2}} + \eta|\mathbb{E}[\zeta]| \leq O\left(n^{\frac{1}{2}}\eta^3\right).\end{aligned}$$

We deduce that : $\|\mathbb{E}[\mathcal{G}-\mathcal{G}_-]\|_F = \sup_{\|B\|_F=1}|\mathrm{Tr}\left(B\mathbb{E}[\mathcal{G}-\mathcal{G}_-]\right)| \leq O\left(n^{-\frac{1}{2}}|z|\eta^3\right)$.

For the second term, $\mathrm{Tr}\left(\mathbb{E}\left[B\frac{\mathfrak{a}-\mathfrak{b}}{\mathfrak{a}}\mathcal{G}_- xx^\top\right]\right) = \mathbb{E}\left[(\mathfrak{a}-\mathbb{E}[\mathfrak{a}])\frac{\xi}{\mathfrak{a}}\right]$ with $\xi = x^\top B\mathcal{G}_- x$. From Proposition 2.4.5, $\mathrm{Var}[\xi] \leq O(\eta^2)$ uniformly in B, and

$|\mathbb{E}[\xi]| = \mathbb{E}[\mathrm{Tr}(B\mathcal{G}_-\Sigma)] \leq p^{\frac{1}{2}}\eta \leq O(n^{\frac{1}{2}}\eta)$. We bound the variance of $\frac{\xi}{\mathfrak{a}}$ using Propositions 2.5.2 and 2.3.6 :

$$\begin{aligned}\mathrm{Var}\left[\frac{\xi}{\mathfrak{a}}\right] &\leq 2\left\|\mathfrak{a}^{-1}\right\|_\infty^2 \mathrm{Var}[\xi] + 2\,\mathrm{Var}\left[\mathfrak{a}^{-1}\right]|\mathbb{E}[\xi]|^2 \\ &\leq O\Big(\eta^4 + \eta^6|z|\Big) \leq O(\eta^6|z|),\end{aligned}$$

where we used that $1 \leq O(\eta) \leq O(\eta^2|z|)$. Cauchy-Schwartz inequality again implies that :

$$\left|\mathbb{E}\left[(\mathfrak{a} - \mathbb{E}[\mathfrak{a}])\frac{\xi}{\mathfrak{a}}\right]\right|^2 \leq \mathrm{Var}\left[\mathfrak{a}\right]\mathrm{Var}\left[\frac{\xi}{\mathfrak{a}}\right] \leq O\Big(n^{-1}\eta^8|z|^3\Big).$$

We deduce that :

$$\left\|\mathbb{E}\left[\frac{\mathfrak{a}-\mathfrak{b}}{\mathfrak{a}}\mathcal{G}_- xx^\top\right]\right\|_F = \sup_{\|B\|_F=1}\left|\mathrm{Tr}\left(\mathbb{E}\left[B\frac{\mathfrak{a}-\mathfrak{b}}{\mathfrak{a}}\mathcal{G}_- xx^\top\right]\right)\right| \leq O\Big(n^{-\frac{1}{2}}\eta^4|z|^{\frac{3}{2}}\Big).$$

Third step :

We wrap up the proof by combining the previous estimates :

$$\begin{aligned}\left\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{b}}\right\|_F &\leq \left|\left|\left|\frac{z}{\mathfrak{b}}\mathbf{G}^{\mathfrak{b}}\right|\right|\right| \cdot \left(\|\mathbb{E}\left[\mathcal{G} - \mathcal{G}_-\right]\|_F |||\Sigma||| + \left\|\mathbb{E}\left[\frac{\mathfrak{a}-\mathfrak{b}}{\mathfrak{a}}\mathcal{G}_- xx^\top\right]\right\|_F\right) \\ &\leq \eta \cdot \Big(O\Big(n^{-\frac{1}{2}}\eta^3|z|\Big) + O\Big(n^{-\frac{1}{2}}\eta^4|z|^{\frac{3}{2}}\Big)\Big) \\ &\leq n^{-\frac{1}{2}}O\Big(\eta^5|z|^{\frac{3}{2}}\Big).\end{aligned}$$

2.6 Second deterministic equivalent

We continue our analysis under the same assumptions as in Theorem 2.5.3.

2.6.1 Reformulation as a fixed point problem

Let us recall the definitions of $\nu = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$, $\check{\nu} = (1-\gamma_n)\delta_0 + \gamma_n \nu$ and $\mathbf{G}^{\mathfrak{l}} = \left(\frac{z}{\mathfrak{l}}\Sigma - zI_p\right)^{-1}$. We let $\mathfrak{c} = -g_{\check{\nu}}{}^{-1}$. In the rest of this section we will prove that $\mathbf{G}^{\mathfrak{c}}$ is close to $\mathbb{E}[\mathcal{G}]$. The precise statement constitutes Theorem 2.6.16. We first establish some properties of ν and $\check{\nu}$.

Lemma 2.6.1. *$\check{\nu}$ is a probability distribution, and $g_{\check{\nu}} = \frac{\gamma_n - 1}{z} + \gamma_n g_\nu$.*

Proof. $\check{\nu}$ has total mass 1 and is a positive measure, excepted maybe in $\{0\}$ if $\gamma_n > 1$. In this case, from [Bel03, Theorem 4.1] we have $\nu(\{0\}) \geq 1 - \gamma_n{}^{-1}$, thus $\check{\nu}(\{0\}) \geq 0$ and $\check{\nu}$ is a probability measure.

The identity between the Stieltjes transforms is easily verified :

$$\begin{aligned} g_{\check{\nu}}(z) &= \int_{\mathbb{R}} \frac{1}{t-z} \check{\nu}(dt) \\ &= (1-\gamma_n) \int_{\mathbb{R}} \frac{1}{t-z} \delta_0(dt) + \gamma_n \int_{\mathbb{R}} \frac{1}{t-z} \nu(dt) \\ &= \frac{1-\gamma_n}{-z} + \gamma_n g_\nu(z). \end{aligned}$$

□

Let us recall the definition of $\Omega = (-\infty, z]$ if $z \in \mathbb{R}^{*-}$, and $\Omega = \{\mathfrak{l} \in \mathbb{C}$ such that $\Im(\mathfrak{l}) \geq \Im(z)$ and $\Im(z^{-1}\mathfrak{l}) \geq 0\}$ if $z \in \mathbb{C}^+$. We introduce the following functional on Ω :

$$\mathcal{F}(\mathfrak{l}) = z + \frac{z}{n} \mathrm{Tr}\left(\mathbf{G}^{\mathfrak{l}}\Sigma\right).$$

Proposition 2.6.2. *$\mathfrak{c}$ is a fixed point of $\mathcal{F}$ and $\mathfrak{c} \in \Omega$. Moreover $\frac{1}{p}\mathrm{Tr}(\mathbf{G}^{\mathfrak{c}}) = g_\nu$.*

Proof. The fact that $\mathfrak{c} \in \Omega$ is a consequence of the classical properties of the Stieltjes transforms. Using the properties of the free multiplicative convolution we have $g_\nu(z) = \int_{\mathbb{R}} \frac{1}{(1-\gamma_n - \gamma_n z g_\nu(z))t - z} \mu_\Sigma(dt)$. Moreover $\frac{z}{\mathfrak{c}} = 1 - \gamma_n - \gamma_n z g_\nu(z)$ using Lemma 2.6.1. We deduce that $\frac{1}{p}\mathrm{Tr}(\mathbf{G}^{\mathfrak{c}}) = \int_{\mathbb{R}} \frac{1}{\frac{z}{\mathfrak{c}}t - z} \mu_\Sigma(dt) = g_\nu(z)$. On the

other hand :

$$\begin{aligned}\frac{z}{n}\mathrm{Tr}(\mathbf{G}^{\mathfrak{c}}\Sigma) &= \frac{\gamma_n}{p}\mathrm{Tr}\left(\left(\frac{\Sigma}{\mathfrak{c}} - I_p\right)^{-1}\Sigma\right)\\ &= \gamma_n \int_{\mathbb{R}} \frac{t}{\frac{t}{\mathfrak{c}} - 1}\mu_\Sigma(dt)\\ &= \gamma_n \int_{\mathbb{R}} \mathfrak{c}\left(1 + \frac{1}{\frac{t}{\mathfrak{c}} - 1}\right)\mu_\Sigma(dt)\\ &= \gamma_n \mathfrak{c}(1 + z g_\nu(z)) = \mathfrak{c} - z.\end{aligned}$$

This proves that $\mathfrak{c}$ is indeed a fixed point of $\mathcal{F}$. □

Remark 2.6.3. We will prove in Corollaries 2.6.8 and 2.6.12 that $\mathfrak{c}$ is the unique fixed point of $\mathcal{F}$ in Ω. The two characterizations we have for $\mathfrak{c}$ correspond to the two characterizations we mentioned for $\mathfrak{a}$ and $\mathfrak{b}$. Indeed by considering that $\mathbb{E}[\mathcal{G}_-] \approx \mathbb{E}[\mathcal{G}] \approx \mathbf{G}^{\mathfrak{b}}$ in the definition of $\mathfrak{b}$, we can see that $\mathfrak{b} \approx z + \frac{z}{n}\mathrm{Tr}\Big(\mathbf{G}^{\mathfrak{b}}\Sigma\Big)$. In other terms $\mathfrak{b}$ is an approximate fixed point of $\mathcal{F}$, and we can hope that it will get close to $\mathfrak{c}$ which is the true fixed point of $\mathcal{F}$.

The definition of $\mathfrak{c}$ using Stieltjes transforms also admits a heuristic interpretation using the co-resolvent $\check{\mathcal{G}}$. We will later understand that $\nu = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$ is a good approximation of μ_K, the spectral distribution of K, in the sense that their Stieltjes transforms are close. On the other hand, the spectra of $K = n^{-1}XX^\top$ and $\check{K} = n^{-1}X^\top X$ differ by $|n - d|$ zeroes, equivalently $\mu_{\check{K}} = (1 - \gamma_n)\delta_0 + \gamma_n \mu_K$. The measure $\check{\nu} = (1 - \gamma_n)\delta_0 + \gamma_n \nu$ is therefore a good approximation of $\mu_{\check{K}}$. With the help of concentration arguments, it makes sense that $\mathfrak{b}^{-1} = \mathbb{E}[\mathfrak{a}]^{-1} = \mathbb{E}\Big[-\check{\mathcal{G}}_{11}^{-1}\Big]^{-1} \approx \mathbb{E}[-g_{\check{K}}]^{-1} \approx -g_{\check{\nu}}^{-1}$, which enlightens the choice of $\mathfrak{c} = -g_{\check{\nu}}^{-1}$.

Our strategy for the rest of this section is as follows : we quantify the error made when saying that $\mathfrak{b}$ is an approximate fixed point of $\mathcal{F}$ in the next proposition. We then study the stability of the fixed point equation to control the gap between $\mathfrak{b}$ and $\mathfrak{c}$. This analysis will rely on the fact that $\mathcal{F}$ is a contraction mapping, with slightly different arguments in the real and in the complex case. We finally deduce that $\mathbf{G}^{\mathfrak{c}}$ is close to $\mathbf{G}^{\mathfrak{b}}$, itself close to $\mathbb{E}[\mathcal{G}]$, which proves Theorem 2.6.16.

Proposition 2.6.4. $|\mathfrak{b} - \mathcal{F}(\mathfrak{b})| \leq O\Big(n^{-1}\eta^5|z|^{\frac{5}{2}}\Big)$.

Proof. Let us remember the estimates obtained in the proof of Theorem 2.5.3 : $\|\mathbb{E}[\mathcal{G} - \mathcal{G}_-]\|_F \leq O\Big(n^{-\frac{1}{2}}|z|\eta^3\Big)$ and $\Big\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{b}}\Big\|_F \leq O\Big(n^{-\frac{1}{2}}\eta^5|z|^{\frac{3}{2}}\Big)$. We deduce

that :

$$\begin{aligned}|\mathfrak{b}-\mathcal{F}(\mathfrak{b})| &= \left|\frac{z}{n}\mathrm{Tr}\Big(\Big(\mathbb{E}[\mathcal{G}_-]-\mathbf{G}^{\mathfrak{b}}\Big)\Sigma\Big)\right| \\ &\leq \frac{|z|}{n}p^{\frac{1}{2}}\Big(\|\mathbb{E}[\mathcal{G}_- - \mathcal{G}]\|_F + \Big\|\mathbb{E}[\mathcal{G}]-\mathbf{G}^{\mathfrak{b}}\Big\|_F\Big) \\ &\leq O\Big(n^{-1}\eta^5|z|^{\frac{5}{2}}\Big).\end{aligned}$$

□

2.6.2 Stability of $\mathcal{F}$ in the real case

Let us working with z belonging to $\mathbb{R}^{*-}$. In this context we recall that $\Omega = (-\infty, z]$. Our first objective in this section is to prove that $\mathcal{F}$ is a contraction mapping around $\mathfrak{c}$ for the usual distance on $\mathbb{R}$.

Lemma 2.6.5. *For any $\mathfrak{l} \in \Omega$, $z - \gamma_n \leq \mathcal{F}(\mathfrak{l}) \leq z \leq 0$. In particular $1 \leq \frac{\mathfrak{c}}{z} \leq 1 + \frac{\gamma_n}{|z|}$.*

Proof. The matrices $\mathbf{G}^{\mathfrak{l}}$ and Σ commute and are both positive semi-definite, hence the product $\mathbf{G}^{\mathfrak{l}}\Sigma$ is positive semi-definite. We deduce that $0 \leq \mathrm{Tr}\Big(\mathbf{G}^{\mathfrak{l}}\Sigma\Big) \leq p|z|^{-1}$, and $-\gamma_n \leq \mathcal{F}(\mathfrak{l}) - z = \frac{z}{n}\mathrm{Tr}\Big(\mathbf{G}^{\mathfrak{l}}\Sigma\Big) \leq 0$. Applying these bounds to $\mathfrak{c} = \mathcal{F}(\mathfrak{c})$ proves the second assertion. □

Lemma 2.6.6. *For any $\mathfrak{l} \in \Omega$, $\left\|\left|\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\Sigma\right|\right\| \leq \frac{1}{1+|z|}$.*

Proof. The eigenvalues of $\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\Sigma = (\Sigma - \mathfrak{l}I_p)^{-1}\Sigma$ are given by $\frac{\lambda}{\lambda - l}$ where λ are the eigenvalues of Σ. All λ belong to $[0,1]$ and $l \leq z < 0$, so we have the following inequalities :

$$0 \leq \frac{\lambda}{\lambda - \mathfrak{l}} = \frac{\lambda}{\lambda + |\mathfrak{l}|} \leq \frac{\lambda}{\lambda + |z|} \leq \frac{1}{1+|z|}.$$

We deduce that all the eigenvalues of $\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\Sigma$ belong to $\Big[0, \frac{1}{1+|z|}\Big]$, which proves the bound in spectral norm. □

Proposition 2.6.7. *$\mathcal{F}$ is a contraction mapping around $\mathfrak{c}$. More precisely, $\mathcal{F}$ is $k_{\mathcal{F}}$-Lipschitz around $\mathfrak{c}$ with $k_{\mathcal{F}} = \frac{\gamma_n}{(1+|z|)(\gamma_n+|z|)} < 1$.*

Proof. For any $\mathfrak{l}$, using a resolvent identity :

$$\begin{aligned}\mathcal{F}(\mathfrak{c}) - \mathcal{F}(\mathfrak{l}) &= \frac{z}{n}\mathrm{Tr}\Big(\Big(\mathbf{G}^{\mathfrak{c}} - \mathbf{G}^{\mathfrak{l}}\Big)\Sigma\Big) \\ &= \frac{z}{n}\mathrm{Tr}\Big(\mathbf{G}^{\mathfrak{c}}\Big(\frac{z}{\mathfrak{l}} - \frac{z}{\mathfrak{c}}\Big)\Sigma\mathbf{G}^{\mathfrak{l}}\Sigma\Big) \\ &= \frac{\mathfrak{c}-\mathfrak{l}}{n}\mathrm{Tr}\Big(\frac{z}{\mathfrak{c}}\mathbf{G}^{\mathfrak{c}}\Sigma\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\Sigma\Big).\end{aligned}$$

A straightforward application of the last lemma would only yield a Lipschitz factor of $\gamma_n(1+|z|)^{-2}$ which may not be small enough. To go beyond, we exploit the fact that $\mathfrak{c}$ is a fixed point of $\mathcal{F}$, in particular $1-\frac{z}{\mathfrak{c}} = \frac{1}{n}\mathrm{Tr}\Big(\frac{z}{\mathfrak{c}}\mathbf{G}^{\mathfrak{c}}\Sigma\Big)$. All the matrices commute in the above expression, and we can use the bounds of the preceding lemmas to obtain :

$$\begin{aligned}
|\mathcal{F}(\mathfrak{c}) - \mathcal{F}(\mathfrak{l})| &\leq |\mathfrak{c}-\mathfrak{l}| \cdot \left|\frac{1}{n}\mathrm{Tr}\Big(\frac{z}{\mathfrak{c}}\mathbf{G}^{\mathfrak{c}}\Sigma\Big)\right| \cdot \left\|\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\Sigma\right\| \\
&\leq |\mathfrak{c}-\mathfrak{l}| \cdot \left(1-\frac{z}{\mathfrak{c}}\right) \cdot \frac{1}{1+|z|} \\
&\leq |\mathfrak{c}-\mathfrak{l}| \cdot \left(1-\frac{1}{1+\frac{\gamma_n}{|z|}}\right) \cdot \frac{1}{1+|z|} \\
&= |\mathfrak{c}-\mathfrak{l}| \cdot \frac{\gamma_n}{(1+|z|)(\gamma_n+|z|)}.
\end{aligned}$$

□

Corollary 2.6.8. *In the real case, $\mathfrak{c}$ is the unique fixed point of $\mathcal{F}$.*

Proof. If $\mathfrak{l}$ is a fixed point of $\mathcal{F}$, then $|\mathfrak{l}-\mathfrak{c}| = |\mathcal{F}(\mathfrak{l}) - \mathcal{F}(\mathfrak{c})| \leq k_{\mathcal{F}}|\mathfrak{l}-\mathfrak{c}|$. Given that $k_{\mathcal{F}} < 1$, necessarily $|\mathfrak{l}-\mathfrak{c}| = 0$ and $\mathfrak{l}$ must be equal to $\mathfrak{c}$. □

Proposition 2.6.9. $|\mathfrak{b}-\mathfrak{c}| \leq O\Big(n^{-1}|z|^{-\frac{7}{2}}\Big)$.

Proof. With $k_{\mathcal{F}} = \frac{\gamma_n}{(1+|z|)(\gamma_n+|z|)} \leq \frac{1}{1+|z|} < 1$, by Proposition 2.6.7 :

$$|\mathfrak{b}-\mathfrak{c}| \leq |\mathfrak{b}-\mathcal{F}(\mathfrak{b})| + |\mathcal{F}(\mathfrak{b}) - \mathcal{F}(\mathfrak{c})| \leq |\mathfrak{b}-\mathcal{F}(\mathfrak{b})| + k_{\mathcal{F}}|\mathfrak{b}-\mathfrak{c}|.$$

Given that $|\mathfrak{b}-\mathcal{F}(\mathfrak{b})| \leq O\Big(n^{-1}|z|^{-\frac{5}{2}}\Big)$ by Proposition 2.6.4 and $\frac{1}{1-k_{\mathcal{F}}} \leq \frac{1+|z|}{|z|} \leq O(|z|^{-1})$, we deduce that $|\mathfrak{b}-\mathfrak{c}| \leq \frac{|\mathfrak{b}-\mathcal{F}(\mathfrak{b})|}{1-k_{\mathcal{F}}} \leq O\Big(n^{-1}|z|^{-\frac{7}{2}}\Big)$. □

2.6.3 Stability of $\mathcal{F}$ in the complex case

We are now working with z belonging to $\mathbb{C}^+$. In this context we recall that $\Omega = \{\mathfrak{l} \in \mathbb{C} \text{ such that } \Im(\mathfrak{l}) \geq \Im(z) \text{ and } \Im(z^{-1}\mathfrak{l}) \geq 0\}$. As in the last section, we would like $\mathcal{F}$ to be a contraction mapping. To this end we need to ditch the usual metric on $\mathbb{C}^+$, and work instead with the following semi-metric :

$$d(\omega_1,\omega_2) = \frac{|\omega_1-\omega_2|}{\Im(\omega_1)^{\frac{1}{2}}\Im(\omega_2)^{\frac{1}{2}}}.$$

Lemma 2.6.10. *For any $\mathfrak{l} \in \Omega$, $\Im(z) \leq \Im(\mathcal{F}(\mathfrak{l})) \leq \Im(z) + \gamma_n|z|\eta$, and :*

$$\Im(\mathcal{F}(l)) - \Im(z) = \frac{|z|^2}{n}\frac{\Im(\mathfrak{l})}{|\mathfrak{l}|^2}\left\|\Sigma\mathbf{G}^{\mathfrak{l}}\right\|_F^2.$$

Proof. Applying the resolvent identity to $z\mathbf{G}^{\mathfrak{l}} = \left(\frac{\Sigma}{\mathfrak{l}} - I_p\right)^{-1}$ and $\overline{z\mathbf{G}^{\mathfrak{l}}} = \left(\frac{\Sigma}{\bar{\mathfrak{l}}} - I_p\right)^{-1}$ leads to :

$$\Im(z\mathbf{G}^{\mathfrak{l}}) = \frac{1}{2i} z\mathbf{G}^{\mathfrak{l}} \left(\frac{\Sigma}{\bar{\mathfrak{l}}} - \frac{\Sigma}{\mathfrak{l}} \frac{\Sigma}{\bar{\mathfrak{l}}} \right) \overline{z\mathbf{G}^{\mathfrak{l}}} = |z|^2 \frac{\Im(\mathfrak{l})}{|\mathfrak{l}|^2} \mathbf{G}^{\mathfrak{l}} \Sigma \overline{\mathbf{G}^{\mathfrak{l}}}.$$

From there we deduce that :

$$\begin{aligned}
\Im(\mathcal{F}(l)) &= \Im(z) + \frac{1}{n} \mathrm{Tr}\Big(\Sigma \Im\Big(z\mathbf{G}^{\mathfrak{l}} \Big) \Big) \\
&= \Im(z) + \frac{|z|^2}{n} \frac{\Im(\mathfrak{l})}{|\mathfrak{l}|^2} \mathrm{Tr}\Big(\Sigma \mathbf{G}^{\mathfrak{l}} \Sigma \overline{\mathbf{G}^{\mathfrak{l}}} \Big) \\
&= \Im(z) + \frac{|z|^2}{n} \frac{\Im(\mathfrak{l})}{|\mathfrak{l}|^2} \left\| \Sigma \mathbf{G}^{\mathfrak{l}} \right\|_F^2.
\end{aligned}$$

This identity proves the lower bound on $\Im(\mathcal{F}(\mathfrak{l}))$. The upper bound is a consequence of : $\left| \frac{1}{n} \mathrm{Tr}\Big(\Sigma \Im\Big(z\mathbf{G}^{\mathfrak{l}} \Big) \Big) \right| \leq \gamma_n \cdot |||\Sigma||| \cdot \left\| \Im\Big(z\mathbf{G}^{\mathfrak{l}} \Big) \right\| \leq \gamma_n |z| \eta.$ □

Proposition 2.6.11. *$\mathcal{F}$ is a contraction mapping with respect to d. More precisely, $\mathcal{F}$ is $k_{\mathcal{F}}$-Lipschitz with $k_{\mathcal{F}} = \frac{\gamma_n |z| \eta^2}{1 + \gamma_n |z| \eta^2} < 1$.*

Proof. For any $\mathfrak{l}, \mathfrak{l}' \in \Omega$, using a resolvent identity :

$$\begin{aligned}
\mathcal{F}(\mathfrak{l}) - \mathcal{F}(\mathfrak{l}') &= \frac{z}{n} \mathrm{Tr}\Big(\Big(\mathbf{G}^{\mathfrak{l}} - \mathbf{G}^{\mathfrak{l}'} \Big) \Sigma \Big) \\
&= \frac{z}{n} \mathrm{Tr}\left(\mathbf{G}^{\mathfrak{l}} \left(\frac{z}{\mathfrak{l}'} - \frac{z}{\mathfrak{l}} \right) \Sigma \mathbf{G}^{\mathfrak{l}'} \Sigma \right) \\
&= \frac{z^2}{n} \frac{\mathfrak{l} - \mathfrak{l}'}{\mathfrak{l}\mathfrak{l}'} \mathrm{Tr}\Big(\mathbf{G}^{\mathfrak{l}} \Sigma \mathbf{G}^{\mathfrak{l}'} \Sigma \Big).
\end{aligned}$$

We can recognize $d(\mathfrak{l}, \mathfrak{l}') = \frac{|\mathfrak{l} - \mathfrak{l}'|}{\Im(\mathfrak{l})^{\frac{1}{2}} \Im(\mathfrak{l}')^{\frac{1}{2}}}$ and the expression for $\Im(\mathcal{F}(\mathfrak{l})) - \Im(z)$ from the preceding lemma :

$$\begin{aligned}
|\mathcal{F}(\mathfrak{l}) - \mathcal{F}(\mathfrak{l}')| &\leq \frac{|z|^2}{n} \frac{|\mathfrak{l} - \mathfrak{l}'|}{|\mathfrak{l}||\mathfrak{l}'|} \left\| \mathbf{G}^{\mathfrak{l}} \Sigma \right\|_F \left\| \mathbf{G}^{\mathfrak{l}'} \Sigma \right\|_F \\
&= \frac{|\mathfrak{l} - \mathfrak{l}'|}{\Im(\mathfrak{l})^{\frac{1}{2}} \Im(\mathfrak{l}')^{\frac{1}{2}}} \left(\frac{|z|^2}{n} \frac{\Im(\mathfrak{l})}{|\mathfrak{l}|^2} \left\| \Sigma \mathbf{G}^{\mathfrak{l}} \right\|_F^2 \right)^{\frac{1}{2}} \left(\frac{|z|^2}{n} \frac{\Im(\mathfrak{l}')}{|\mathfrak{l}'|^2} \left\| \Sigma \mathbf{G}^{\mathfrak{l}'} \right\|_F^2 \right)^{\frac{1}{2}} \\
&= d(\mathfrak{l}, \mathfrak{l}') (\Im(\mathcal{F}(\mathfrak{l})) - \Im(z))^{\frac{1}{2}} (\Im(\mathcal{F}(\mathfrak{l}')) - \Im(z))^{\frac{1}{2}}.
\end{aligned}$$

We thus have :

$$\begin{aligned}
d(\mathcal{F}(\mathfrak{l}), \mathcal{F}(\mathfrak{l}')) &= \frac{|\mathcal{F}(\mathfrak{l}) - \mathcal{F}(\mathfrak{l}')|}{\Im(\mathcal{F}(\mathfrak{l}))^{\frac{1}{2}} \Im(\mathcal{F}(\mathfrak{l}'))^{\frac{1}{2}}} \\
&\leq d(\mathfrak{l}, \mathfrak{l}') \left(1 - \frac{\Im(z)}{\Im(\mathcal{F}(\mathfrak{l}))} \right)^{\frac{1}{2}} \left(1 - \frac{\Im(z)}{\Im(\mathcal{F}(\mathfrak{l}'))} \right)^{\frac{1}{2}}.
\end{aligned}$$

Given that $\Im(z) \leq \Im(\mathcal{F}(\mathfrak{l})) \leq \Im(z) + \gamma_n|z|\eta$, we have $0 \leq 1 - \frac{\Im(z)}{\Im(\mathcal{F}(\mathfrak{l}))} \leq 1 - (1+\gamma_n|z|\eta^2)^{-1} = \frac{\gamma_n|z|\eta^2}{1+\gamma_n|z|\eta^2}$, hence $d(\mathcal{F}(\mathfrak{l}), \mathcal{F}(\mathfrak{l}')) \leq \frac{\gamma_n|z|\eta^2}{1+\gamma_n|z|\eta^2} d(\mathfrak{l}, \mathfrak{l}')$. □

Corollary 2.6.12. *In the complex case, $\mathfrak{c}$ is the unique fixed point of $\mathcal{F}$.*

Proof. In order to obtain a contradiction, let $\mathfrak{l} \in \Omega$ be a fixed point of $\mathcal{F}$ such that $\mathfrak{l} \neq \mathfrak{c}$. Using Lemma 2.6.10, $\Im(\mathfrak{l}) = \Im(\mathcal{F}(\mathfrak{l})) = \Im(z) + \frac{|z|^2}{n}\frac{\Im(\mathfrak{l})}{|\mathfrak{l}|^2}\left\|\Sigma\mathbf{G}^{\mathfrak{l}}\right\|_F^2$. Since $\Im(z) > 0$, we have $1 > \frac{|z|^2}{n}\frac{1}{|\mathfrak{l}|^2}\left\|\Sigma\mathbf{G}^{\mathfrak{l}}\right\|_F^2$. Similarly $1 > \frac{|z|^2}{n}\frac{1}{|\mathfrak{c}|^2}\|\Sigma\mathbf{G}^{\mathfrak{c}}\|_F^2$.

As seen in the proof of Proposition 2.6.11, we have $\mathfrak{l} - \mathfrak{c} = \mathcal{F}(\mathfrak{l}) - \mathcal{F}(\mathfrak{c}) = \frac{z^2}{n}\frac{\mathfrak{l}-\mathfrak{c}}{\mathfrak{l}\mathfrak{c}}\mathrm{Tr}\left(\mathbf{G}^{\mathfrak{l}}\Sigma\mathbf{G}^{\mathfrak{c}}\Sigma\right)$. We obtain a contradiction after taking a Cauchy-Schwarz inequality :

$$1 = \frac{|z|^2}{n}\frac{1}{|\mathfrak{l}||\mathfrak{c}|}\mathrm{Tr}\left(\mathbf{G}^{\mathfrak{l}}\Sigma\mathbf{G}^{\mathfrak{c}}\Sigma\right) \leq \frac{|z|^2}{n}\frac{1}{|\mathfrak{l}|}\left\|\mathbf{G}^{\mathfrak{l}}\Sigma\right\|_F\frac{1}{|\mathfrak{c}|}\|\mathbf{G}^{\mathfrak{c}}\Sigma\|_F < 1.$$

□

Our second objective in this section is to use the contraction property of $\mathcal{F}$ to quantify how $\mathfrak{b}$ and $\mathfrak{c}$ are close. Since d is not a true metric (it lacks the triangular inequality), we need a few additional lemmas :

Lemma 2.6.13. *Given ω_1 and $\omega_2 \in \mathbb{C}^+$, $\left|\frac{1}{\Im(\omega_1)^{\frac{1}{2}}}\right| \leq \left|\frac{1}{\Im(\omega_2)^{\frac{1}{2}}}\right|(1 + d(\omega_1, \omega_2))$.*

Proof. It is simply a matter of computing :

$$\begin{aligned}
\frac{1}{\Im(\omega_1)^{\frac{1}{2}}} &= \frac{1}{\Im(\omega_2)^{\frac{1}{2}}}\left(1 + \frac{\Im(\omega_2)^{\frac{1}{2}} - \Im(\omega_1)^{\frac{1}{2}}}{\Im(\omega_1)^{\frac{1}{2}}}\right) \\
&= \frac{1}{\Im(\omega_2)^{\frac{1}{2}}}\left(1 + \frac{\Im(\omega_2) - \Im(\omega_1)}{\Im(\omega_1)^{\frac{1}{2}}\left(\Im(\omega_2)^{\frac{1}{2}} + \Im(\omega_1)^{\frac{1}{2}}\right)}\right), \\
\left|\frac{1}{\Im(\omega_1)^{\frac{1}{2}}}\right| &\leq \left|\frac{1}{\Im(\omega_2)^{\frac{1}{2}}}\right|\left(1 + \frac{|\Im(\omega_2) - \Im(\omega_1)|}{\Im(\omega_1)^{\frac{1}{2}}\left(\Im(\omega_2)^{\frac{1}{2}} + \Im(\omega_1)^{\frac{1}{2}}\right)}\right) \\
&\leq \left|\frac{1}{\Im(\omega_2)^{\frac{1}{2}}}\right|(1 + d(\omega_1, \omega_2)).
\end{aligned}$$

□

Lemma 2.6.14. *Let $f : \mathbb{C}^+ \to \mathbb{C}^+$ be k_f-Lipshitz with respect to d, let c be a fixed point of f and let $b \in \mathbb{C}^+$ be another point. We let $\Delta = d(b, f(b))$. If $k_f(1+\Delta) < 1$, then the following inequality holds true :*

$$|c - b| \leq \frac{|f(b) - b|}{1 - k_f(1+\Delta)}.$$

Proof. Using the above lemma with $\omega_1 = b$ and $\omega_2 = f(b)$, we have :

$$\frac{|c - f(b)|}{\Im(c)^{\frac{1}{2}}\Im(b)^{\frac{1}{2}}} \leq \frac{|f(c) - f(b)|}{\Im(c)^{\frac{1}{2}}\Im(f(b))^{\frac{1}{2}}}(1 + d(b, f(b))) \leq k_f\, d(c, b)(1 + \Delta).$$

Therefore :

$$\begin{aligned} d(c, b) &= \frac{|c - b|}{\Im(c)^{\frac{1}{2}}\Im(b)^{\frac{1}{2}}} \\ &\leq \frac{|c - f(b)|}{\Im(c)^{\frac{1}{2}}\Im(b)^{\frac{1}{2}}} + \frac{|f(b) - b|}{\Im(c)^{\frac{1}{2}}\Im(b)^{\frac{1}{2}}} \\ &\leq k_f d(c, b)(1 + \Delta) + \frac{|f(b) - b|}{\Im(c)^{\frac{1}{2}}\Im(b)^{\frac{1}{2}}}. \end{aligned}$$

Provided $k_f(1 + \Delta) < 1$, we obtain :

$$\begin{aligned} d(c, b) &\leq \frac{1}{1 - k_f(1 + \Delta)} \frac{|f(b) - b|}{\Im(c)^{\frac{1}{2}}\Im(b)^{\frac{1}{2}}} \\ |c - b| &\leq \frac{|f(b) - b|}{1 - k_f(1 + \Delta)}. \end{aligned}$$

□

We will now prove the equivalent of Proposition 2.6.9 in the complex case.

Proposition 2.6.15. *There is a constant $C > 0$ such that, if $\eta^8|z|^{\frac{7}{2}} \leq Cn$, then $|\mathfrak{b} - \mathfrak{c}| \leq O\big(n^{-1}\eta^7|z|^{\frac{7}{2}}\big)$.*

Proof. We want to apply Lemma 2.6.14 to the function $\mathcal{F}$, the true fixed point $\mathfrak{c}$ and the approximate fixed point $\mathfrak{b}$. $\mathcal{F}$ is $k_{\mathcal{F}}$-Lipschitz with $k_{\mathcal{F}} = \frac{\gamma_n\eta^2|z|}{1+\gamma_n\eta^2|z|}$. On the other hand, from Proposition 2.6.4 we have :

$$\Delta = d(\mathfrak{b}, \mathcal{F}(\mathfrak{b})) = \frac{|\mathfrak{b} - \mathcal{F}(\mathfrak{b})|}{\Im(\mathfrak{b})^{\frac{1}{2}}\Im(\mathcal{F}(\mathfrak{b})^{\frac{1}{2}}} \leq O\big(n^{-1}\eta^6|z|^{\frac{5}{2}}\big),$$

where we used that $\Im(\mathfrak{b})$ and $\Im(\mathcal{F}(\mathfrak{b})) \geq \Im(z) = \eta^{-1}$. By assumption γ_n is a bounded sequence, hence we can find a constant $C > 0$ such that $\gamma_n\Delta \leq Cn^{-1}\eta^6|z|^{\frac{5}{2}}$. Now if $\eta^8|z|^{\frac{7}{2}} \leq n/2C$, then $\gamma_n\eta^2|z|\Delta \leq Cn^{-1}\eta^8|z|^{\frac{7}{2}} \leq 1/2$, and for n large enough :

$$k_{\mathcal{F}}(1 + \Delta) = \frac{\gamma_n\eta^2|z| + \gamma_n\eta^2|z|\Delta}{1 + \gamma_n\eta^2|z|} \leq \frac{\gamma_n\eta^2|z| + 1/2}{1 + \gamma_n\eta^2|z|} = 1 - \frac{1}{2(1 + \gamma_n\eta^2|z|)} < 1.$$

From Lemma 2.6.14 we have : $\frac{1}{1-k_{\mathcal{F}}(1+\Delta)} \leq 2(1 + \gamma_n\eta^2|z|) \leq O(\eta^2|z|)$. We deduce that :

$$|\mathfrak{c} - \mathfrak{b}| \leq \frac{|\mathcal{F}(\mathfrak{b}) - \mathfrak{b}|}{1 - k_{\mathcal{F}}(1 + \Delta)} \leq O\big(n^{-1}\eta^7|z|^{\frac{7}{2}}\big).$$

□

2.6.4 Proof of the second deterministic equivalent

Let us recall the definitions of $\nu = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$, $\check{\nu} = (1-\gamma_n)\delta_0 + \gamma_n \nu$, $\mathfrak{c} = -g_{\check{\nu}}{}^{-1}$, and $\mathbf{G}^{\mathfrak{c}} = \left(\frac{z}{\mathfrak{c}}\Sigma - zI_p\right)^{-1}$.

Theorem 2.6.16 (Second equivalent)**.** *1. If $z \in \mathbb{R}^{*-}$ are real spectral arguments, possibly varying with n, such that $|z|$ is bounded and $|z|^{-7} \leq O(n)$, then :*

$$\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{c}}\|_F \leq O\left(\frac{1}{n^{\frac{1}{2}}|z|^{\frac{11}{2}}}\right).$$

2. There exists a constant $C > 0$ such that, if $z \in \mathbb{C}^+$ are complex spectral arguments with $\Im(z)$ bounded and $\eta^{16}|z|^7 \leq Cn$, then :

$$\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{c}}\|_F \leq O\left(\frac{\eta^9|z|^{\frac{5}{2}}}{n^{\frac{1}{2}}}\right).$$

Remark 2.6.17. In the real case, the assumptions on z are not strictly necessary to obtain an explicit bound. Indeed we can always find a $O\left(n^{-\frac{1}{2}}|z|^{-2}\kappa\right)$ estimate with the same parameters as in the remark following Theorem 2.5.3 : $\sigma^2 = |z|^{-2}(1 + n^{-1}|z|^{-1})$ and $\kappa = \sigma^2|z|^{-1} + \sigma|z|^{-\frac{5}{2}}$.

The situation in the complex case is however different this time, because we really need $\eta^8|z|^{\frac{7}{2}} \leq Cn$ to prove the stability of the fixed point problem. This is a stronger hypothesis than the hypothesis $\eta^{10}|z|^3 \leq O(n)$ we used for mere convenience in the first deterministic equivalent. In any case, the above bound will not be interesting unless we have at least $\eta^{18}|z|^5 \leq O(n)$.

Also note that the exponent in the real case is slightly better than the one we would obtain by replacing $\Im(z)$ in the complex setting by $|z|$ in the real setting (11/2 instead of 13/2). Surprisingly enough, this lack of symmetry does not come from the different methods we use to analyze the stability of the fixed point equation, but rather from a difference in the Lipschitz parameter of the application $\mathfrak{l} \in \Omega \mapsto \mathbf{G}^{\mathfrak{l}}$ as we will see in the upcoming proof.

Proof. We have quantified how close $\mathfrak{b}$ and $\mathfrak{c}$ are in Propositions 2.6.9 and 2.6.15, and we can derive a Lipschitz property for $\mathfrak{l} \mapsto \mathbf{G}^{\mathfrak{l}}$ using a resolvent identity :

$$\left\|\mathbf{G}^{\mathfrak{l}} - \mathbf{G}^{\mathfrak{l}'}\right\|_F = \left\|z|\mathfrak{l} - \mathfrak{l}'|\frac{\mathbf{G}^{\mathfrak{l}'}}{\mathfrak{l}'}\Sigma\frac{\mathbf{G}^{\mathfrak{l}}}{\mathfrak{l}}\right\|_F \leq p^{\frac{1}{2}}\frac{|\mathfrak{l} - \mathfrak{l}'|}{|z|}\left|\left|\left|\frac{z}{\mathfrak{l}'}\mathbf{G}^{\mathfrak{l}'}\Sigma\right|\right|\right| \cdot \left|\left|\left|\frac{z}{\mathfrak{l}}\mathbf{G}^{\mathfrak{l}}\right|\right|\right|.$$

This is where we can get better exponents in the real case using Lemma 2.6.6. $\left|\left|\left|\frac{z}{\mathfrak{b}}\mathbf{G}^{\mathfrak{b}}\Sigma\right|\right|\right| \leq \frac{1}{1+|z|} \leq 1$, thus :

$$\left\|\mathbf{G}^{\mathfrak{c}} - \mathbf{G}^{\mathfrak{b}}\right\|_F \leq p^{\frac{1}{2}}|z|^{-1}O\left(n^{-1}|z|^{-\frac{7}{2}}\right)|z|^{-1} \leq O\left(n^{-\frac{1}{2}}|z|^{-\frac{11}{2}}\right).$$

In the complex case, we only have $\left\|\left\|\frac{z}{\mathfrak{b}}\mathbf{G}^{\mathfrak{b}}\Sigma\right\|\right\| \leq \eta$ by Proposition 2.5.2, thus :

$$\left\|\mathbf{G}^{\mathfrak{c}} - \mathbf{G}^{\mathfrak{b}}\right\|_F \leq p^{\frac{1}{2}}|z|^{-1}O\Big(n^{-1}\eta^7|z|^{\frac{7}{2}}\Big)\eta^2 \leq O\Big(n^{-\frac{1}{2}}\eta^9 z^{\frac{5}{2}}\Big)$$

In both cases, the first deterministic equivalent in Theorem 2.5.3 gives a term of lesser magnitude for $\left\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{b}}\right\|_F$, which completes the proof since :

$$\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{c}}\|_F \leq \left\|\mathbb{E}[\mathcal{G}] - \mathbf{G}^{\mathfrak{b}}\right\|_F + \left\|\mathbf{G}^{\mathfrak{c}} - \mathbf{G}^{\mathfrak{b}}\right\|_F.$$

□

2.7 Proof of the main results

Most of the technical work is already achieved in Sections 2.4-2.6. For instance, the Proposition 2.2.2 stating the concentration of $\mathcal{G}_K(z)$ is a direct consequence of Corollary 2.4.4. Indeed with $\eta = |z|^{-1}$ if $z \in \mathbb{R}^{*-}$, we have $n^{-\frac{1}{2}}\eta^2|z|^{\frac{1}{2}} = n^{-\frac{1}{2}}|z|^{-\frac{3}{2}}$. With $\eta = \Im(z)^{-1}$ if $z \in \mathbb{C}^+$, we have $n^{-\frac{1}{2}}\eta^2|z|^{\frac{1}{2}} = n^{-\frac{1}{2}}\Im(z)^{-2}|z|^{\frac{1}{2}}$.

The conclusions of our main Theorem 2.2.3 are exactly the same as in Theorem 2.6.16 after carefully comparing the definitions of the deterministic equivalent. We need however to add a quick argument for the Proposition 2.2.4 and the Corollary 2.2.5. We remind our reader that a.s.e. stands for almost surely eventually.

Proof of Proposition 2.2.4. From Theorem 2.2.3, uniformly in any $A \in \mathbb{R}^{p\times p}$ we have :

$$\begin{aligned}|\mathrm{Tr}(\mathbb{E}[\mathcal{G}_K(z)]A - \mathbf{G}(z)A)| &\leq \|\mathbb{E}[\mathcal{G}_K(z)] - \mathbf{G}(z)\|_F \cdot \|A\|_F \\ &\leq \|A\|_F\, O\Big(\frac{\kappa}{n^{\frac{1}{2}}}\Big),\end{aligned}$$

where $\kappa = \frac{1}{|z|^{\frac{11}{2}}}$ in the real case, or $\kappa = \frac{|z|^{\frac{5}{2}}}{\Im(z)^9}$ in the complex case. On the other hand, from Propositions 2.2.2, 2.3.2 and 2.3.3, uniformly in any $A \in \mathbb{R}^{p\times p}$ we have :

$$|\mathrm{Tr}(\mathcal{G}_K(z)A - \mathbb{E}[\mathcal{G}_K(z)]A)| \leq \|A\|_F\, O\left(\frac{\tau(\log n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \quad \text{a.s.e.}$$

where $\tau = \frac{1}{|z|^{\frac{3}{2}}}$ in the real case, or $\tau = \frac{|z|^{\frac{1}{2}}}{\Im(z)^2}$ in the complex case.

In both scenarios, $\tau \leq O(\kappa)$. Indeed in the real case, $\tau\kappa^{-1} = |z|^4 \leq O(1)$. In the complex case, $\tau\kappa^{-1} = \Im(z)^7|z|^{-2} \leq \Im(z)^5 \leq O(1)$. We deduce that uniformly in any $A \in \mathbb{R}^{p\times p}$:

$$\begin{aligned}|\mathrm{Tr}(\mathcal{G}_K(z)A - \mathbf{G}(z)A)| &\leq |\mathrm{Tr}(\mathcal{G}_K(z)A - \mathbb{E}[\mathcal{G}_K(z)]A)| \\ &\quad + |\mathrm{Tr}(\mathbb{E}[\mathcal{G}_K(z)]A - \mathbf{G}(z)A)| \\ &\leq \|A\|_F\, O\left(\frac{\kappa(\log n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \quad \text{a.s.e.}\end{aligned}$$

□

Proof or Corollary 2.2.5. Considering $A = p^{-1}I_p$ which satisfies $\|A\|_F = p^{-\frac{1}{2}} = O(n^{-\frac{1}{2}})$, and using the identity $p^{-1}\mathrm{Tr}\mathbf{G}(z) = g_\nu(z)$ from Proposition 2.6.2

yields :

$$\left|\frac{1}{p}\mathrm{Tr}\mathcal{G}_K(z) - \frac{1}{p}\mathrm{Tr}\mathbf{G}(z)\right| = |g_K(z) - g_\nu(z)| \leq O\left(\frac{\kappa(\log n)^{\frac{1}{2}}}{n}\right) \quad \text{a.s.e.}$$

Uniformly in $\mathbf{u} \in \mathbb{S}^{p-1}$, with $A = \mathbf{u}\mathbf{u}^\top$, $\|A\|_F = \|\mathbf{u}\| = 1$:

$$\left|\mathrm{Tr}\Big(\mathcal{G}_K(z)\mathbf{u}\mathbf{u}^\top - \mathbf{G}(z)\mathbf{u}\mathbf{u}^\top\Big)\right| = \left|g_{\mu_{K,\mathbf{u}}}(z) - \mathbf{u}^\top\mathbf{G}(z)\mathbf{u}\right| \leq O\left(\frac{\kappa(\log n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \quad \text{a.s.e.}$$

Finally the third assertion is a consequence of the polarization identity for symmetric matrices, which relates the entries of any matrix M to terms like $\mathbf{u}^\top M\mathbf{u}$:

$$2M_{ij} = (\mathbf{e}_i + \mathbf{e}_j)^\top M(\mathbf{e}_i + \mathbf{e}_j) - \mathbf{e}_i^\top M\mathbf{e}_i - \mathbf{e}_j^\top M\mathbf{e}_j.$$

The entry-wise maximum is controlled by p^2 such terms, hence $\max_{1\leq i,j\leq p} |\mathcal{G}_K(z) - \mathbf{G}(z)|_{ij} \leq O\Big(\frac{\kappa(\log n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\Big)$ a.s.e. as a consequence of the second assertion. □

2.8 Bounds in Kolmogorov distance

The goal of this section is to prove Theorem 2.8.6, which provides an asymptotic bound in Kolmogorov distance given a quantitative bound on the Stieltjes transforms of probability distributions. It closely follows the ideas of [BM20, Theorem 5.2], with some key differences. First of all we do not ask for estimates of the Stieltjes transforms on the whole complex upper plane, but rather for specific sequences that will appear in the proof. This broader setting will be well adapted to the conclusions of Theorem 2.2.3. Compared to the work of [BM20], we also do not require weak convergence of the distributions, and we work without uniform bounds on the support of our distributions. More precisely, we open the possibility that their second moments grow slowly to ∞.

Let us remind the notations $\mathcal{F}_\nu$ for the cumulative distribution function (CDF) of ν, and $D(\nu,\mu) = \sup_{t\in\mathbb{R}} |\mathcal{F}_\nu(t) - \mathcal{F}_\mu(t)|$ for the Kolmogorov distance between ν and μ. The following well-known facts are well-surveyed in [GH03] : the convergence in Kolmogorov distance implies the weak convergence for probability measures, and there is even an equivalence if the limiting measure admits a Hölder continuous CDF.

2.8.1 Bound for fixed measures

To compare distributions in Kolmogorov distance we will first need a few technical lemmas for fixed probability measures μ and ν on $\mathbb{R}^+$. We start with the so-called Bai's Inequality [Bai08, Theorem 2.1] :

Lemma 2.8.1. *Let $\mathfrak{a} = \tan\left(\frac{3\pi}{8}\right)$. If $\int_{\mathbb{R}} |\mathcal{F}_\nu(t) - \mathcal{F}_\mu(t)|dt < \infty$, then for any $y > 0$:*

$$D(\mu,\nu) \leq \frac{2}{\pi}\left(\int_{\mathbb{R}} |g_\nu - g_\mu|(t+iy)dt + \frac{1}{y}\sup_{x\in\mathbb{R}}\int_{[\pm 2y\mathfrak{a}]} |\mathcal{F}_\mu(x+t) - \mathcal{F}_\mu(x)|dt\right).$$

The following tail bound for the Stieltjes transforms is a consequence of the intermediate result [BM20, Lemma 5.5] :

Lemma 2.8.2. *If μ and ν have finite second moments, smaller than $\sigma \geq 1$, then for any $y \in (0,1)$ and $A > 0$ we have :*

$$\int_{(\pm A)^c} |g_\nu - g_\mu|(t+iy)dt \leq \frac{4\sigma^3}{y^2 A}.$$

Finally the last lemma is a simple consequence of the Hölder property :

Lemma 2.8.3. *If $\mathcal{F}_\mu$ is Hölder continuous with constant C and exponent $\beta \in (0,1]$, then for any $y > 0$:*

$$\frac{1}{y}\sup_{x\in\mathbb{R}}\int_{[\pm 2y\mathfrak{a}]} |\mathcal{F}_\mu(x+t) - \mathcal{F}_\mu(x)|dt \leq 50Cy^\beta.$$

Cutting the first integral of Lemma 2.8.1 in $\pm A$ and using the lemmas gives the following estimate for the Kolmogorov distance between two fixed measures :

Proposition 2.8.4. *Suppose that μ and ν have finite second moments, smaller than $\sigma \geq 1$, that $\int_{\mathbb{R}} |\mathcal{F}_\nu(t) - \mathcal{F}_\mu(t)|dt < \infty$, and that $\mathcal{F}_\mu$ is Hölder continuous with constant C and exponent $\beta \in (0,1]$.*

There is a new constant $C' > 0$, only depending on C, such that for any $y \in (0,1)$ and $A > 0$:

$$D(\mu,\nu) \leq C'\left(A \max_{t\in(\pm A)} |g_\nu - g_\mu|(t+iy) + \frac{\sigma^3}{y^2 A} + y^\beta\right).$$

2.8.2 Asymptotic bound

Let us consider μ_n and ν_n two sequences of probability measures on $\mathbb{R}^+$ satisfying :

Assumptions 2.8.5. (1) $\int_{\mathbb{R}} |\mathcal{F}_{\nu_n}(t) - \mathcal{F}_{\mu_n}(t)|dt < \infty$ for all n (a uniform bound is not required).

(2) All CDFs $\mathcal{F}_{\mu_n}$ are uniformly Hölder continuous with the same exponent $\beta \in (0,1]$ and constant $C > 0$.

(3) μ_n and ν_n have finite second moments bounded by a sequence $(\sigma_n) \geq 1$.

(4) There are exponents $k, l \in (0,\infty)$ and a sequence $(\epsilon_n) > 0$ with $(\sigma_n^{3(1+l)}\epsilon_n) \to 0$ such that, for any $z \in \mathbb{C}^+$:

$$|g_{\mu_n}(z) - g_{\nu_n}(z)| \leq \epsilon_n \frac{|z|^l}{(\Im z)^k}.$$

This last assumption may be replaced by the following weaker version :

(4') There are exponents $k, l \in (0,\infty)$, and a sequence $(\epsilon_n) > 0$ with $(\sigma_n^{3(1+l)}\epsilon_n) \to 0$ such that, setting $y_n = \left(\sigma_n^{3(1+l)}\epsilon_n\right)^{\frac{1}{2+2l+k+(2+l)\beta}}$ and $A_n = \left(\frac{\sigma_n^3}{\epsilon_n}\right)^{\frac{1}{2+l}} y_n^{\frac{k-2}{2+l}}$, the following inequality holds :

$$\sup_{z=t+iy_n,\, t\in(\pm A_n)} |g_{\mu_n}(z) - g_{\nu_n}(z)| \leq \epsilon_n \frac{|A_n + iy_n|^l}{y_n^k}.$$

Theorem 2.8.6. *Under the Assumptions 2.8.5 :*

$$D(\mu_n,\nu_n) \leq O\left(\left(\sigma_n^{3(1+l)}\epsilon_n\right)^{\frac{\beta}{2+2l+k+(2+l)\beta}}\right).$$

Proof. The idea is to optimize the inequality given by Proposition 3.8.2 in both y and A. We pass on the details leading to the aforementioned choices of y_n and A_n. We claim that $\epsilon_n \frac{A_n^{1+l}}{y_n^k} = \frac{\sigma_n^3}{y_n^2 A_n} = y_n^\beta$. This fact is related to the properties of optimizing sequences and can also be checked manually. Indeed $A_n^{2+l} = \frac{\sigma_n^3}{\epsilon_n} y_n^{k-2} \iff \epsilon_n \frac{A_n^{1+l}}{y_n^k} = \frac{\sigma_n^3}{y_n^2 A_n}$, and :

$$\frac{\sigma_n^3}{y_n^2 A_n} = \sigma_n^3 y_n^{-2} \left(\frac{\sigma_n^3}{\epsilon_n}\right)^{\frac{-1}{2+l}} y_n^{\frac{2-k}{2+l}} = \left(\sigma_n^{3(1+l)} \epsilon_n\right)^{\frac{1}{2+l}} y_n^{-\frac{2+2l+k}{2+l}} = y_n^\beta.$$

Also note that $y_n \to 0$, and $A_n \geq y_n^{-2-\beta} \to \infty$, in particular $(y_n) \in (0,1)$ and $A_n \geq 1$ for n large enough. From Assumption 2.8.5 (4') :

$$\max_{t \in (\pm A_n)} |g_{\nu_n}(t + iy_n) - g_{\mu_n}(t + iy_n)| \leq \epsilon_n \frac{|A_n + iy_n|^l}{y_n^k} \leq O\left(\epsilon_n \frac{A_n^l}{y_n^k}\right).$$

We finally apply Proposition 3.8.2 :

$$\begin{aligned} D_(\mu_n, \nu_n) &\leq C' \left(A_n \max_{t \in (\pm A_n)} |g_{\nu_n} - g_{\mu_n}|(t + iy_n) + \frac{\sigma_n^3}{y_n^2 A_n} + y_n^\beta \right) \\ &\leq C' \left(O\left(\epsilon_n \frac{A_n^{1+l}}{y_n^k}\right) + \frac{\sigma_n^3}{y_n^2 A_n} + y_n^\beta \right) \\ &\leq O\left(y_n^\beta\right) \leq O\left(\left(\sigma_n^{3(1+l)} \epsilon_n\right)^{\frac{\beta}{2+2l+k+(2+l)\beta}} \right). \end{aligned}$$

□

2.8.3 Application to the empirical spectral distribution of sample covariance matrices

Let us go back to the setting of our main result. To apply Theorem 2.8.6 given the bound of Corollary 2.2.5, we need to address two difficulties : the Hölder property for the CDF of $\nu_n = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$, and a control on the second moment of μ_K.

Lemma 2.8.7. *If Σ is invertible and $\|\|\Sigma^{-1}\|\|$ is bounded, then all CDF of ν_n are uniformly $1/2$-Hölder continuous on $\mathbb{R}^+$. If moreover $\gamma_n \leq 1$, the same property holds on $\mathbb{R}$.*

Proof. $\nu_n = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$ is supported on $\left[0, (1 + \gamma_n^{\frac{1}{2}})^2 \lambda_{\max}(\Sigma)\right]$ and admits a continuous density on f_ν on $(0, \infty)$ that satisfies $f_\nu(t) \leq \pi^{-1} (\lambda_{\min}(\Sigma) t \gamma_n)^{-\frac{1}{2}}$ ([BHZ12, Lemma 2.3]). Given the bounds on γ_n and the eigenvalues of Σ, we deduce that all $\mathcal{F}_{\nu_n}$ are uniformly $1/2$-Hölder continuous on $\mathbb{R}^+$. In the case $\gamma_n \leq 1$, all $\mathcal{F}_{\nu_n}$ are identically equal to 0 on $(-\infty, 0]$, hence the Hölder continuity on $\mathbb{R}$. □

Lemma 2.8.8. *$|||K||| \leq O(1)$ almost surely eventually (a.s.e.). In particular μ_K have uniformly bounded support a.s.e. and the second moments of μ_K are uniformly bounded a.s.e.*

Proof. From Proposition 2.3.4 we have $|||X - \mathbb{E}[X]||| \leq O\left(n^{\frac{1}{2}}\right)$ a.s.e. Given that $|||\mathbb{E}[X]||| = n^{\frac{1}{2}}\|\mathbb{E}[x]\| = O(n^{\frac{1}{2}})$, we also have $|||X||| \leq O\left(n^{\frac{1}{2}}\right)$ a.s.e. As a consequence, $|||K||| = n^{-1}|||X|||^2 \leq O(1)$ a.s.e. which in turn implies that μ_K have uniformly bounded support and uniformly bounded second moments a.s.e. □

Proof of Theorem 2.2.7. Let us first treat the case where all $\gamma_n \leq 1$. We want to apply Theorem 2.8.6 with the measures μ_K and ν.

The Assumptions (1), (2) and (3), with $\beta = 1/2$ and $\sigma_n = 1$, are consequences of the previous lemmas. Let us explain why the Assumption (4'), with $l = k = 9$ and $\epsilon_n = n^{-1}(\log n)^{\frac{1}{2}}$ holds true. $y_n = \epsilon_n^{\frac{1}{2+2l+k+(2+l)\beta}} = \epsilon_n^{2/69}$ is bounded and $A_n = y_n^{-2-\beta} = \epsilon_n^{-5/69}$, hence $\frac{|A_n+iy_n|^7}{y_n^{16}} \leq O\left(\epsilon_n^{-67/69}\right) = o(n)$. All $z = t + iy_n$, where $t \in (\pm A_n)$, satisfy the Assumptions of Corollary 2.2.5, which is why :

$$\max_{t\in(\pm A_n)} |g_{\nu_n}(t+iy_n) - g_{\mu_n}(t+iy_n)| \leq \max_{t\in(\pm A_n)} \epsilon_n \frac{|t+iy_n|^{\frac{5}{2}}}{y_n^9} \leq O\left(\epsilon_n \frac{A_n^9}{y_n^9}\right).$$

We conclude that$D(\mu_K, \nu) \leq O\left(\epsilon_n^{1/69}\right) \leq O\left(n^{-1/70}\right)$.

In the case where all $\gamma_n \geq 1$, ν has an atom at 0 (see Lemma 2.6.1) and $\mathcal{F}_\nu$ is even discontinuous. We work instead on $\tilde{\nu}$, thanks to the following formulas :

$$\begin{aligned} \mu_{\check{K}} &= (1-\gamma_n)\delta_0 + \gamma_n \mu_K, & \check{\nu} &= (1-\gamma_n)\delta_0 + \gamma_n \nu, \\ g_{\check{K}} &= \frac{\gamma_n - 1}{z} + \gamma_n g_K, & g_{\check{\nu}} &= \frac{\gamma_n - 1}{z} + \gamma_n g_\nu, \\ \mathcal{F}_{\mu_{\check{K}}} &= (1-\gamma_n)\mathbf{1}_{\mathbb{R}^+} + \gamma_n \mathcal{F}_\mu, & \mathcal{F}_{\check{\nu}} &= (1-\gamma_n)\mathbf{1}_{\mathbb{R}^+} + \gamma_n \mathcal{F}_\nu. \end{aligned}$$

$\check{\nu}$ has no atom in 0, hence all $\mathcal{F}_{\check{\nu}}$ are identically equal to 0 on $(-\infty, 0]$. Given that they also are uniformly $\frac{1}{2}$-Hölder continuous on $\mathbb{R}^+$, they are uniformly $\frac{1}{2}$-Hölder continuous on $\mathbb{R}$. Since $|g_{\check{K}} - g_{\check{\nu}}| = \gamma_n |g_K - g_\nu|$, we can apply Theorem 2.8.6 with the measures $\mu_{\check{K}}$ and $\tilde{\nu}$. We conclude that $D(\mu_K, \nu) = \gamma_n^{-1} D(\mu_{\check{K}}, \check{\nu}) \leq O\left(n^{-1/70}\right)$.

In all generality, we apply the arguments on the subsequences corresponding to $\gamma_n \leq 1$ and $\gamma_n > 1$ and we still retrieve the same result. □

Remark 2.8.9. In the application of Theorem 2.8.6, choosing indices $l = 5/2$ and $k = 9$ as in Corollary 2.2.5, would lead to $\frac{A_n^7}{y_n^{16}} \leq O(\epsilon_n)^{-143/73} \not\leq o(n)$ which is not sufficient to apply Corollary 2.2.5. We could have chosen however any $l = (43+\epsilon)/5$, which yields $D(\mu_K, \nu) \leq O(n^{-1/67-\epsilon'})$. Given that our quantitative bound is not optimal in the first place, we choose not to include this technical detail in the result.

To prove Corollary 2.2.8 we need a bound on the Kolmogorov distance between Marčenko-Pastur distributions with different shape parameters. Surprisingly enough, we could not find any references for this question, which is why we propose our own proof below.

Lemma 2.8.10. *If γ and $\gamma' > 0$, then $D(\mathrm{MP}(\gamma), \mathrm{MP}(\gamma')) \leq \frac{|\gamma-\gamma'|}{\max(\gamma,\gamma')}$.*

Proof. Let n and $p \geq p'$ be fixed integers, and consider a random matrix $Y \in \mathbb{R}^{p\times n}$ filled with i.i.d. $\mathcal{N}$ entries. We set $Y' = (Y_{ij})_{1\leq i\leq p', 1\leq j\leq n}$ to be the first $p' \times n$ block matrix of Y, and we define the sample covariance matrices $C = n^{-1}YY^\top \in \mathbb{R}^{p\times p}$ and $C' = n^{-1}Y'Y'^\top \in \mathbb{R}^{p'\times p'}$. We denote by $\mathcal{F} = \mathcal{F}_{\mu_C}$ and $\mathcal{F}' = \mathcal{F}_{\mu_{C'}}$ the CDF corresponding to the ESD of C and C'. If $\lambda_1 \leq \cdots \leq \lambda_p$ are the ordered eigenvalues of C and $\lambda'_1 \leq \cdots \leq \lambda'_{p'}$ those of C', then $\mathcal{F}(t) = (1/p)\cdot\sum_{k=1}^{p} \mathbf{1}_{\lambda_k\leq t}$, and similarly $\mathcal{F}'(t) = (1/p)\cdot\sum_{k=1}^{p} \mathbf{1}_{\lambda'_k\leq t}$.

C' being equal to the first $p'\times p'$ block matrix of C, the interlacing eigenvalues theorem states that $\lambda_k \leq \lambda'_k \leq \lambda_{k+p-p'}$ for any $k \in [\![1, p']\!]$. Depending on the values of $t \in \mathbb{R}$, we can distinguish between three cases :

1. If $\mathcal{F}'(t) = 0$, then $t < \lambda'_1 \leq \lambda_{1+p-p'}$, hence $\mathcal{F}(t) \leq \frac{p-p'}{p}$ and $|\mathcal{F}(t) - \mathcal{F}'(t)| \leq \frac{p-p'}{p}$.
2. If $\mathcal{F}'(t) = 1$, then $t \geq \lambda'_{p'} \geq \lambda_{p'}$, hence $\mathcal{F}(t) \geq \frac{p'}{p}$ and $|\mathcal{F}(t) - \mathcal{F}'(t)| \leq \frac{p-p'}{p}$.
3. If $\mathcal{F}'(t) = k/p'$ for some $k \in [\![1, p'-1]\!]$, then $\lambda_k \leq \lambda'_k \leq t < \lambda'_{k+1} \leq \lambda_{k+1+p-p'}$, in particular $\frac{k}{p} \leq \mathcal{F}(t) \leq \frac{k+p-p'}{p}$. Given that :

$$\left|\frac{k}{p} - \frac{k}{p'}\right| = \frac{k(p-p')}{pp'} \leq \frac{p-p'}{p} \quad ,$$
$$\left|\frac{k+p-p'}{p} - \frac{k}{p'}\right| = \frac{(p-p')(p'-k)}{pp'} \leq \frac{p-p'}{p} \quad ,$$

in this case we also have $|\mathcal{F}(t) - \mathcal{F}'(t)| \leq \frac{p-p'}{p}$.

We have therefore obtained : $D(\mu_C, \mu_{C'}) = \sup_{t\in\mathbb{R}} |\mathcal{F}(t) - \mathcal{F}'(t)| \leq \frac{p-p'}{p}$.

Consider now parameters $\gamma \geq \gamma' > 0$, and $p(n)$ and $p'(n)$ sequences of integers such that $p/n \to \gamma$ and $p'/n \to \gamma'$ respectively. In this case $\limsup_{n\to\infty} D(\mu_C, \mu_{C'}) \leq \lim_{n\to\infty} \frac{p-p'}{p} = \frac{\gamma-\gamma'}{\gamma}$. On the other hand μ_C and $\mu_{C'}$ converge respectively to $\mathrm{MP}(\gamma)$ and $\mathrm{MP}(\gamma')$ in Kolmogorov distance ([BS10]). The lemma is proven after taking the limit in the following triangular inequality :

$$D(\mathrm{MP}(\gamma), \mathrm{MP}(\gamma')) \leq D(\mathrm{MP}(\gamma), \mu_C) + D(\mu_C, \mu_{C'}) + D(\mu_{C'}, \mathrm{MP}(\gamma')).$$

□

Proof of Corollary 2.2.8. The almost surely weak convergence of μ_K towards ν_∞ is a consequence of the regularity of the free convolution with respect to the weak convergence of measures. Moreover from [BV93, Proposition 4.13] we have the quantitative bound : $D(\nu_n, \nu_\infty) \leq D(\mathrm{MP}(\gamma_n), \mathrm{MP}(\gamma_\infty)) + D(\mu_\Sigma, \mu_\infty)$. Also note that $\frac{|\gamma_n - \gamma_\infty|}{\max(\gamma_n, \gamma_\infty)} \leq O(|\gamma_n - \gamma_\infty|)$. Using the preceding lemma, we get :

$$D(\mu_K, \nu_\infty) \leq O\Big(n^{-\frac{1}{70}}\Big) + O(|\gamma_n - \gamma_\infty|) + D(\mu_\Sigma, \mu_\infty).$$

□

2.9 Application to kernel methods

This section deals with the regularization of random feature models, using our main result to obtain Theorem 2.9.2. Let us introduce this method, closely following the approach of [JSS+20].

2.9.1 Kernel ridge regression

In kernel ridge regression (KRR) we are considering a training dataset made of N distincts vectors $\mathcal{X}_1, \dots, \mathcal{X}_N \in \mathbb{R}^D$, and associated real numbers labels $\mathcal{Y}_1, \dots, \mathcal{Y}_N$, grouped as a vector $\mathcal{Y} \in \mathbb{R}^N$. A kernel $\mathcal{K} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is a function such that the matrix $\mathcal{K}_\mathcal{X} = (\mathcal{K}(\mathcal{X}_i, \mathcal{X}_j))_{1 \leq i,j \leq N}$ is definite positive. We define the functions $\mathcal{K}(\cdot, \mathcal{X}_j) : \mathbb{R}^D \to \mathbb{R}$, $x \mapsto \mathcal{K}(x, \mathcal{X}_j)$ and $\mathcal{K}(\cdot, \mathcal{X}) : \mathbb{R}^D \to \mathbb{R}^N, x \mapsto (\mathcal{K}(x, \mathcal{X}_j))_{1 \leq j \leq N}$.

The goal of KRR is to find in the linear span of $\{\mathcal{K}(\cdot, \mathcal{X}_j), 1 \leq j \leq N\}$ a function f such that $f(\mathcal{X}_i) \approx \mathcal{Y}_i$ for all $1 \leq i \leq N$. More precisely, given a ridge parameter $\lambda > 0$, we want to minimize in $\theta \in \mathbb{R}^N$ the mean square error with a λ ridge penalization term :

$$\frac{1}{N} \sum_{1 \leq i \leq N} \left(\sum_{1 \leq j \leq N} \theta_j \mathcal{K}(\mathcal{X}_i, \mathcal{X}_j) - \mathcal{Y}_i \right)^2 + \frac{\lambda}{N} \theta^\top \mathcal{K}(\mathcal{X}, \mathcal{X}) \theta.$$

Applying basic linear algebra techniques to the above minimization problem, it can be shown that $\hat{\theta} = (\mathcal{K}_\mathcal{X} + \lambda I_N)^{-1} \mathcal{Y}$ is the optimal value for θ, and that :

$$\begin{aligned} \hat{f}_\lambda^{(K)} : \mathbb{R}^D &\to \mathbb{R}, \\ x &\mapsto \mathcal{K}(x, \mathcal{X})(\mathcal{K}_\mathcal{X} + \lambda I_N)^{-1} \mathcal{Y} \end{aligned}$$

is the optimal function, thus called KRR predictor with ridge λ.

2.9.2 Random features method

A set of random features (RF) associated with the kernel $\mathcal{K}$ is a collection $\Phi = \left(\Phi^{(j)}\right)_{1 \leq j \leq P}$ of P random i.i.d. processes $\Phi^{(j)} : \mathbb{R}^D \to \mathbb{R}$, chosen such that they are centered and admit $\mathcal{K}$ as covariance function : for all $x, x' \in \mathbb{R}^D$, $\mathbb{E}\left[\Phi^{(j)}(x)\right] = 0$ and $\mathbb{E}\left[\Phi^{(j)}(x)\Phi^{(j)}(x')\right] = \mathcal{K}(x, x')$.

In the RF method, we would like to find in the linear span of Φ a function f such that $f(\mathcal{X}_i) \approx \mathcal{Y}_i$ for all $1 \leq i \leq N$. More precisely, given a ridge parameter $\lambda > 0$, we want to minimize in $\theta \in \mathbb{R}^P$ the mean square error with a λ ridge penalization term :

$$\frac{1}{N} \sum_{1 \leq i \leq N} \left(\sum_{1 \leq j \leq N} P^{-\frac{1}{2}} \theta_j \Phi^{(j)}(\mathcal{X}_i) - \mathcal{Y}_i \right)^2 + \frac{\lambda}{N} \theta^\top \theta.$$

Let us define the data matrix $F = P^{-\frac{1}{2}}\left(\Phi^{(j)}(\mathcal{X}_i)\right)_{1\leq i\leq N, 1\leq j\leq P} \in \mathbb{R}^{N\times P}$. Again using linear algebra techniques, it can be shown that :

$$\hat{\theta} = F^\top\left(FF^\top + \lambda I_N\right)^{-1}\mathcal{Y}$$

is the optimal value for θ, and that :

$$\begin{aligned}\hat{f}_\lambda^{(RF)} : \mathbb{R}^D &\to \mathbb{R}\\ x &\mapsto P^{-\frac{1}{2}}\Phi(x)F^\top\left(FF^\top + \lambda I_N\right)^{-1}\mathcal{Y}\end{aligned}$$

is the optimal function, thus called RF predictor with ridge λ.

2.9.3 Effective ridge parameter

The RF predictor is a good approximation of the KRR predictor in the over-parametrized regime $P \gg N$. Indeed :

$$\hat{f}_\lambda^{(RF)}(x) = (\Psi(x,\mathcal{X}_j))_{1\leq j\leq N}\left[(\Psi(\mathcal{X}_i,\mathcal{X}_j))_{1\leq i,j\leq N} + \lambda I_N\right]^{-1}\mathcal{Y},$$

where Ψ is the random function $(x,x') \mapsto \frac{1}{P}\sum_{1\leq k\leq P}\Phi^{(k)}(x)\Phi^{(k)}(x')$. When N is fixed and $P\to\infty$, Ψ converges to $\mathcal{K}$ by the law of large numbers, hence :

$$\hat{f}_\lambda^{(RF)}(x) \to (\mathcal{K}(x,\mathcal{X}_j))_{1\leq j\leq N}\left[(\mathcal{K}(\mathcal{X}_i,\mathcal{X}_j))_{1\leq i,j\leq N} + \lambda I_N\right]^{-1}\mathcal{Y} = \hat{f}_\lambda^{(K)}(x).$$

The under-parametrized regime $P < N$ is more interesting for practical purposes, since computing the RF predictor requires to invert a $P\times P$ matrix instead of a $N\times N$ matrix for the KRR predictor. In this regime unfortunately, a systematic bias appears and the RF predictor is not a good estimator of KRR predictor. The authors of [JSS+20] proved however that the average RF predictor $\mathbb{E}\left[\hat{f}_\lambda^{(RF)}\right]$ is close to the KRR predictor $\hat{f}_{\tilde{\lambda}}^{(K)}$ with a different parameter $\tilde{\lambda}$, which they called effective ridge parameter. In order to state their result, let us introduce $\gamma = \frac{P}{N}$ the ratio of dimensions, d_i the eigenvalues of $\mathcal{K}_\mathcal{X}$, and the inverse kernel norm $\|\cdot\|_{\mathcal{K}_\mathcal{X}^{-1}}$ defined as $\|\mathbf{v}\|^2_{\mathcal{K}_\mathcal{X}^{-1}} = \mathbf{v}^\top\mathcal{K}_\mathcal{X}^{-1}\mathbf{v}$.

Theorem 2.9.1 ([JSS+20], Theorem 4.1)**.** *If $\Phi^{(j)}$ are Gaussian processes, for $N,P>0$ and $\lambda>0$, we have :*

$$\left|\mathbb{E}\left[\hat{f}_\lambda^{(RF)}(x)\right] - \hat{f}_{\tilde{\lambda}}^{(K)}(x)\right| \leq \frac{c\sqrt{\mathcal{K}(x,x)}\|\mathcal{Y}\|_{\mathcal{K}_\mathcal{X}^{-1}}}{P},$$

where the effective ridge $\tilde{\lambda} > \lambda$ is the unique positive number satisfying :

$$\tilde{\lambda} = \lambda + \frac{\tilde{\lambda}}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i}.$$

The constant $c>0$ only depends on λ,γ, and $\frac{1}{N}\operatorname{Tr}\mathcal{K}_\mathcal{X}$.

Let us define the measure $\check{\nu}$ as :

$$\check{\nu} = \Big(1 - \frac{N}{P}\Big)\delta_0 + \frac{N}{P}\Big(\mathrm{MP}\Big(\frac{N}{P}\Big) \boxtimes \mu_{\mathcal{K}_\mathcal{X}}\Big).$$

We will now see how our main result implies that $\mathbb{E}\Big[\hat{f}_\lambda^{(RF)}\Big] \approx \hat{f}_{\tilde{\lambda}}^{(K)}$ with quantitative bounds, where $\tilde{\lambda} = g_{\check{\nu}}(\lambda)^{-1}$.

Theorem 2.9.2. *Consider $\Phi^{(j)}$ random processes (not necessarily Gaussian), such that the data matrix F is $\propto \mathcal{E}_2(N^{-\frac{1}{2}})$ concentrated with respect to the Frobenius norm. Assume that $|||\mathcal{K}_\mathcal{X}|||$ is bounded, and that γ is bounded from above and from below.*

Then uniformly in any $x \in \mathbb{R}^D$ and any bounded $\lambda > 0$ such that $\lambda^{-7} \leq O(N)$, with $\tilde{\lambda} = g_{\check{\nu}}(\lambda)^{-1} > \lambda$, as N and P tend to ∞ :

$$\Big|\mathbb{E}\Big[\hat{f}_\lambda^{(RF)}(x)\Big] - \hat{f}_{\tilde{\lambda}}^{(K)}(x)\Big| \leq O\left(\frac{\mathcal{K}(x,x)^{\frac{1}{2}}\|\mathcal{Y}\|_{\mathcal{K}_\mathcal{X}^{-1}}}{N^{\frac{1}{2}}\lambda^{\frac{9}{2}}}\right).$$

The implicit constant in the $O(\cdot)$ notation only depends on the concentration constant of F, and the bounds on γ, $\mathcal{K}_\mathcal{X}$ and λ.

Remark 2.9.3. Both descriptions of the effective ridge parameter $\tilde{\lambda}$ in the Theorems 2.9.1 and 2.9.2 are of course equivalent, a fact that is best seen using Proposition 2.6.2. As a consequence, one can easily retrieve all properties of $\tilde{\lambda}$ contained in [JSS+20, Proposition 4.2] using general results on Stieltjes transforms and free multiplicative convolution.

When γ is bounded from above and from below, that is when the dimensions grow at the same rate (up to a multiplicative factor), we can compare Theorems 2.9.1 and 2.9.2, and remark that the convergence speed of the original result is faster than ours. The authors relied on a similar deterministic equivalent result, in operator norm but with a better exponent than the one we obtained in Frobenius norm. This is made posible by reducing the problem to diagonal kernel matrices only, and we believe these arguments to be Gaussian specific.

Proof. By assumption $P^{\frac{1}{2}}F \propto \mathcal{E}_2(1)$, and the columns of $P^{\frac{1}{2}}F$ are i.i.d., centered, and admit $\mathcal{K}_\mathcal{X}$ as covariance matrix. We can apply Theorem 2.2.3 with $z = -\lambda \in \mathbb{R}^{*-}$:

$$\left\|\mathbb{E}\Big[\Big(FF^\top + \lambda I_N\Big)^{-1}\Big] - \Big(\frac{\lambda}{\tilde{\lambda}}\mathcal{K}_\mathcal{X} + \lambda I_N\Big)^{-1}\right\|_F \leq O\Big(N^{-\frac{1}{2}}\lambda^{-\frac{11}{2}}\Big).$$

Given the identities $F\Big(F^\top F + \lambda I_P\Big)^{-1}F^\top = I_N - \lambda\Big(FF^\top + \lambda I_N\Big)^{-1}$ and $I_N - \lambda\Big(\frac{\lambda}{\tilde{\lambda}}\mathcal{K}_\mathcal{X} + \lambda I_N\Big)^{-1} = \mathcal{K}_\mathcal{X}\Big(\mathcal{K}_\mathcal{X} + \tilde{\lambda} I_N\Big)^{-1}$, we have :

$$\left\|\mathbb{E}\Big[F\Big(F^\top F + \lambda I_P\Big)^{-1}F^\top\Big] - \mathcal{K}_\mathcal{X}\Big(\mathcal{K}_\mathcal{X} + \tilde{\lambda} I_N\Big)^{-1}\right\|_F \leq O\Big(N^{-\frac{1}{2}}\lambda^{-\frac{9}{2}}\Big).$$

We deduce that :

$$\begin{aligned}
&\Big|\mathbb{E}\big[\hat{f}_\lambda^{(RF)}(x)\big] - \hat{f}_{\tilde{\lambda}}^{(K)}(x)\Big| \\
&= \Big|\mathcal{K}(x,\mathcal{X})\Big(\mathcal{K}_{\mathcal{X}}^{-1}\mathbb{E}\Big[F\Big(F^\top F + \lambda I_P\Big)^{-1}F^\top\Big] - \Big(\mathcal{K}_{\mathcal{X}} + \tilde{\lambda} I_N\Big)^{-1}\Big)\mathcal{Y}\Big| \\
&\leq O\Big(N^{-\frac{1}{2}}\lambda^{-\frac{9}{2}}\Big)\Big|\mathcal{K}(x,\mathcal{X})\mathcal{K}_{\mathcal{X}}^{-1}\mathcal{Y}\Big| \\
&\leq O\Big(N^{-\frac{1}{2}}\lambda^{-\frac{9}{2}}\Big)\|\mathcal{K}(x,\mathcal{X})\|_{\mathcal{K}_{\mathcal{X}}^{-1}}\|\mathcal{Y}\|_{\mathcal{K}_{\mathcal{X}}^{-1}},
\end{aligned}$$

where we use the Cauchy-Schwarz inequality on the inverse kernel norm to obtain the last estimate. As a general property of inverse kernel norms (see [JSS+20, Theorem C.8]), it can be shown that $\|\mathcal{K}(x,\mathcal{X})\|_{\mathcal{K}_{\mathcal{X}}^{-1}} \leq \mathcal{K}(x,x)^{\frac{1}{2}}$, which concludes the proof. □

Chapitre 3

Deterministic equivalent of the conjugate kernel matrix associated to artificial neural networks

This self-contained chapter has been taken from the article [Cho23], to be submitted soon, and available on arXiv.

Abstract

We study the conjugate kernel associated to a multi-layer linear-width feed-forward neural network with random weights, biases and data. We show that the empirical spectral distribution of the conjugate kernel converges to a deterministic limit. More precisely we obtain a deterministic equivalent for its Stieltjes transform and its resolvent, with quantitative bounds involving both the dimension and the spectral parameter. The limiting equivalent objects are described by iterating free convolution of measures and classical matrix operations involving the parameters of the model.

Contents

3.1 Introduction **81**
3.1.1 Overview of the article 83
3.1.2 General notations and definitions 83
3.2 Technical tools **85**
3.2.1 Concentration framework 85
3.2.2 Polynomial bounds in z and notation $O_z(\epsilon_n)$ 86
3.3 Covariance matrices of coordinate-wise functions of Gaussian vectors **89**
3.3.1 Hermite polynomials 89
3.3.2 Iterated Hadamard products 91
3.3.3 General expansion of the covariance matrix Σ 92
3.3.4 Approximation of Σ for weakly correlated Gaussian vectors 92
3.3.5 Linearization of Σ in specific settings 94
3.4 Deterministic equivalent of sample covariance matrices with a general dependence structure **96**
3.4.1 General results 97
3.4.2 Regularity of the Stieltjes transform with respect to the free convolution 98
3.4.3 Approximation of the deterministic equivalent built from deterministic matrices 102
3.4.4 Concentration of the deterministic equivalent built from random matrices 103
3.5 Single-layer neural network with deterministic data **105**
3.5.1 Setting 105
3.5.2 Technicalities and linearization of Σ 106
3.5.3 Propagation of the approximate orthogonality 107
3.5.4 Deterministic equivalent and consequences 108
3.5.5 Application to another model involving entry-wise operations 111
3.6 Single-layer neural network with random data **113**
3.6.1 Setting 113
3.6.2 Deterministic equivalent and consequences 114
3.6.3 Application to data matrices with i.i.d. columns . . . 117
3.7 Multi-layer neural network model **119**
3.8 Appendix : Bounds on Kolmogorov distances between empirical measures **123**

3.1 Introduction

Artificial neural networks are computing systems inspired by a simplified model of interconnected neurons, storing and exchanging information inside a biological brain. Theorized in the late fifties [Ros58], this class of algorithms only established oneself as a cutting-edge technology in the two thousands, thanks to the ever-increasing computing power, and perhaps even more thanks to the accessibility of huge databases in the modern-age internet. Compared to the delicate craftsmanship of artificial intelligence specialists, the mathematical understanding of these networks remained at a very basic level until recently. The classical tools of random matrix theory in particular were helpless regarding the intrinsic non-linear connections between the artificial neurons.

This article is concerned about the conjugate kernel model, first introduced from the point of view of random matrices in [PW17]. This model is a random matrix ensemble that corresponds to a multi-layer feed-forward neural network with weights initialized at random, in the asymptotic regime where the network width grows linearly with the sample size. In this setting, the conjugate kernel matrix is simply the sample covariance matrix of the output vectors produced by the final layer of the network. It can be shown that the conjugate kernel governs the training and generalization properties of the underlying network ([ASS20], [YS19]). Its spectral properties are thus of great theoretical and practical interest.

The limiting spectral distribution of the conjugate kernel associated to a single-layer network was first characterized in [PW17] under the form of a quartic polynomial equation satisfied by the Stieltjes transform of the limiting measure. This was proven rigorously in [BP21] under more general assumptions, using advanced combinatorics. Some results on the extremal eigenvalues of the model were even obtained in [BP22]. [Pé19] first noticed a connection between these equations and the measures obtained from a free multiplicative convolution with some Marčenko-Pastur distributions. Let us also mention the work of [PS21], where a variant of model including a rank-one additive perturbation is examined, also by means of combinatorial techniques.

A first step towards a deeper understanding of the model using analytical techniques was done in [LC18], in the case of a single-layer network with deterministic data. [FW20] later generalized this result to random data and multi-layer networks, and was the first paper to convincingly explain the appearance of free probability convolutions in the limiting spectral distribution of the conjugate kernel matrix. A similar analysis was done in [WZ23], in the regime where the network width is much larger than the sample size.

The aforementioned results may be classified as global laws, in the sense that they establish the existence of a limiting spectral distribution, a problem that is directly linked to the convergence of the Stieltjes transform via the

classical Frobenius-Perron inversion formula. The next natural question that arises in random matrix theory is to look for a deterministic equivalent of the resolvent of the conjugate kernel matrix in the sense of [HLN07]. This consists in finding a deterministic matrix that is asymptotically close to the expected resolvent of the matrix, and thus close to the random resolvent itself provided enough concentration in the random matrices. Establishing a deterministic equivalent result allows for instance to approximate linear statistics of the eigenvectors [Yan20], to examine the convergence of the eigenvectors empirical spectral distributions, and possibly to study the outliers of spiked models [Noi21]. Regarding classical sample covariance random ensembles, this task was done a decade ago in the series of articles [PY14], [BEK+14], [KY17]. Such results were recently partially extended to models with a general dependence structure in [LC21], [Cho22].

The most important contribution of this article is Theorem 3.7.2, that provides a deterministic equivalent of the conjugate kernel of a multi-layer neural network model. This theorem extends previously known results in various directions. First we include models with non-differentiable activation functions, as well as potential biases inside or outside the activation function. Secondly we give a quantitative estimate for the convergence of the Stieltjes transforms, which translates into a quantitative convergence of the measures in Kolmogorov distance. Finally and most importantly, we obtain local results, taking the form of a deterministic equivalent for the resolvent matrix of the conjugate kernel, quantitatively on both the dimension and the spectral parameter.

In the rest of the paper, we also establish intermediary results that may be interesting by themselves. In particular we show that the free convolution of measures with a Marčenko-Pastur distribution is regular with respect to the Stieltjes transform of these distributions (Theorem 3.4.7). We obtain a similar property for the deterministic equivalent matrices that appear in our results (Proposition 3.4.11). We also show how our framework applies to other models involving entry-wise operations on random matrices in Section 3.5.5.

This paper relies mostly on analytical methods that study spectral functions of random matrices. We make great use of the theory of Stieltjes transforms and resolvent matrices, particularly its recent developments towards the deterministic equivalent of general sample covariance matrices ([LC21], [Cho22]). We introduce a new notion of asymptotic equivalence for objects that depend both on a dimension parameter, and on a complex spectral parameter. Our definition is convenient to work with, whilst keeping enough precision to imply some quantitative results like the convergence of measures in Kolmogorov distance.

This analytical toolbox works in conjunction with concentration of measure principles. We follow the framework of [LC20], in particular we use the notion of Lipschitz concentration that is remarkably compatible with entry-wise operations on random matrices. We also use a linearization argument to

study the covariance matrices of functions of weakly correlated Gaussian vectors. Using the theory of Hermite polynomials, we provide estimates for this approximation, not only on an entry-wise basis like it was done in [FW20], but also in spectral norm which is new to our knowledge.

After we finished a first version of this manuscript, we came across the preprint [SCDL23], which we were not aware of. A future comparison of the two different perspectives would be certainly interesting, as [SCDL23] studies similar objects and seems to describe related phenomena to those considered here.

3.1.1 Overview of the article

In Section 3.2, we remind the notations and basic properties of concentrated random vectors and matrices, following the framework of [LC20]. We also introduce the new notion of asymptotic equivalence that will be key in the rest of the article. In Section 3.3, we address the problem of the approximating the covariance matrices of functions of weakly correlated Gaussian vectors. In Section 3.4, we remind the general deterministic equivalent results on which this article is based, and we study thoroughly the properties of the matrices appearing in these equivalents. In Section 3.5, we study a first artificial neural network model with a single layer and deterministic data. We also explain how our framework applies to other models involving entry-wise operations on random matrices (3.5.5). In Section 3.6, we analyze a second artificial neural network model, still consisting of a single layer but with random data instead. In Section 3.7, we explain how to study multi-layer networks by induction on the model with one layer. In the Appendix 3.8, we prove an independent result about the convergence in Kolmogorov distance of some empirical spectral measures.

3.1.2 General notations and definitions

The set of matrices with d lines, n columns, and entries belonging to a set $\mathbb{K}$ is denoted as $\mathbb{K}^{d\times n}$. We use the following norms for vectors and matrices : $\|\cdot\|$ the Euclidean norm, $\|\cdot\|_F$ the Frobenius norm, $|||\cdot|||$ the spectral norm, and $\|\cdot\|_{\max}$ the entry-wise maximum norm.

Given $M \in \mathbb{C}^{n\times n}$, we denote by $M^\top$ its transpose, and by $\mathrm{Tr}(M) = \sum_{i=1}^n M_{ii}$ its trace. If M is real and diagonalizable we denote by $\mathrm{Sp}M$ its spectrum and by $\mu_M = \frac{1}{n}\sum_{\lambda\in \mathrm{Sp}M}\delta_\lambda$ its empirical spectral distribution. If M is symmetric positive semi-definite, then $\mathrm{Sp}M \subset \mathbb{R}^+$, and given a spectral parameter $z \in \mathbb{C}^+ = \{z \in \mathbb{C} \text{ such that } \Im(z) > 0\}$, we define its resolvent $\mathcal{G}_M(z) = (M - zI_n)^{-1}$ and its Stieltjes transform $g_M(z) = (1/n)\mathrm{Tr}\mathcal{G}_M(z)$.

If μ is a real probability distribution, its Stieltjes transform is $g_\mu(z) = \int_{\mathbb{R}} \frac{\mu(dt)}{t-z}$, well defined for $z \in \mathbb{C}^+$. It is easy to see that the Stieltjes transform

of a matrix is the same as the Stieltjes transform of its empirical spectral distribution, that is $g_{\mu_M}(z) = g_M(z)$. We denote by $\mathcal{F}_\mu$ be the cumulative distribution function of the measure μ, and by $D(\mu,\nu) = \sup_{t\in\mathbb{R}} |\mathcal{F}_\mu(t) - \mathcal{F}_\nu(t)|$ the Kolmogorov distance between μ and ν.

Let us also introduce two operations on measures : if $\gamma \in \mathbb{R}$, we denote by $\gamma \cdot \mu + (1-\gamma)\cdot\nu$ the new measure such that $(\gamma\cdot\mu + (1-\gamma)\cdot\nu)(B) = \gamma\mu(B) + (1-\gamma)\nu(B)$ for any measurable set B (it may be a signed measure). If $a, b \in \mathbb{R}$, we denote by $a + b\mu$ the distribution of the random variable $a + bX$, where X is a random variable distributed according to μ.

We denote by $\mathcal{N}$ a standard (centered and reduced) Gaussian random variable, or the law of this random variable depending on the context. We denote by δ_a the Dirac delta distribution centered at $a \in \mathbb{R}$, and by $\mathrm{MP}(\gamma)$ the Marčenko-Pastur distribution with shape parameter $\gamma > 0$. The operator $\boxtimes$ denotes the multiplicative free convolution between measures ([BV93]). The distribution $\mathrm{MP}(\gamma) \boxtimes \mu$ may be also defined by its Stieltjes transform $g(z)$, which is the unique Stieltjes transform function that solves the self-consistent equation [MP67] :

$$g(z) = \int_{\mathbb{R}} \frac{1}{(1-\gamma-\gamma z g(z))t - z}\mu(dt).$$

Throughout this paper, we say that a property occurs almost surely eventually (a.s.e.) if there is a full measure set on which the property holds true for all but finitely many n. Equivalenty, we ask that there is a full measure set on which the property holds true for all integers n greater than a (possibly random) rank n_0. Such statements typically result from the application of Borel-Cantelli lemma.

For a better readability, we will sometimes omit to mention indices n and spectral parameters z in our notations, especially in the course of technical proofs, even if we are dealing implicitly with sequences and complex functions.

3.2 Technical tools

3.2.1 Concentration framework

Definition 3.2.1. Let X_n be a sequence of random vectors in finite dimensional normed vector spaces $(E_n, \|\cdot\|)$, and let $\sigma_n > 0$ be a sequence.

1. We say that X is Lipschitz concentrated if there is a constant $C > 0$ such that for any sequence of 1-Lipschitz maps $f_n : E_n \to \mathbb{C}$, for any $n \in \mathbb{N}$ and $t \geq 0$:

$$\mathbb{P}(|f_n(X_n) - \mathbb{E}[f_n(X_n)]| \geq t) \leq Ce^{-\frac{1}{C}\left(\frac{t}{\sigma_n}\right)^2}.$$

   We note this Lipschitz concentration property $X \propto_{\|\cdot\|} \mathcal{E}(\sigma_n)$, or simply $X \propto \mathcal{E}(\sigma_n)$ when there is no ambiguity on the chosen norm.

2. If $Z_n \in E_n$ is a sequence of deterministic vectors, we say that X is linearly concentrated around Z if there is a constant $C > 0$ such that for any sequence of 1-Lipschitz linear maps $f_n : E_n \to \mathbb{C}$, for any $n \in \mathbb{N}$ and $t \geq 0$:

$$\mathbb{P}(|f_n(X_n) - f_n(Z_n)| \geq t) \leq Ce^{-\frac{1}{C}\left(\frac{t}{\sigma_n}\right)^2}.$$

   We note this linear concentration property $X \in_{\|\cdot\|} Z \pm \mathcal{E}(\sigma_n)$, or simply $X \in Z \pm \mathcal{E}(\sigma_n)$.

We refer to [LC21] for a comprehensive study of these notions of concentration. Let us enumerate the main properties that will be used throughout this article :

Proposition 3.2.2. *1. A random vector (respectively a matrix) with i.i.d. $\mathcal{N}$ entries is $\propto \mathcal{E}(1)$ concentrated with respect to the Euclidean norm (respectively the Frobenius norm). Other examples are listed in [LC21, Theorem 1].*

2. A Lipschitz transformation of a Lipschitz concentrated vector still verifies a Lipschitz concentration property [LC21, Proposition 1].

3. If $X \propto \mathcal{E}(\sigma_n)$, $Y \propto \mathcal{E}(\sigma'_n)$, and X and Y are independent, then $(X, Y) \propto \mathcal{E}(\sigma_n + \sigma'_n)$ ([LC21, Proposition 7]).

4. Lipschitz concentration implies linear concentration : if $X \propto \mathcal{E}(\sigma_n)$, then $X \in \mathbb{E}[X] \pm \mathcal{E}(\sigma_n)$([LC21, Proposition 4]). Both definitions are equivalent in one-dimensional spaces.

5. If $X \in Z \pm \mathcal{E}(\sigma_n)$ and $\tilde{Z}$ is another deterministic vector, we have the equivalence : $X \in \tilde{Z} \pm \mathcal{E}(\sigma_n) \iff \left\|Z - \tilde{Z}\right\| \leq O(\sigma_n)$ ([LC21, Lemma 3]).

6. *If* $X \in Z \pm \mathcal{E}(\sigma_n)$ *and* X' *is another random vector such that* $\|X'\| \leq O(\sigma_n)$ *a.s., then* $X + X' \in Z \pm \mathcal{E}(\sigma_n)$.
7. *Since* $|||\cdot||| \leq \|\cdot\|_F$ *on matrices, the concentration with respect to the Frobenius norm is a stronger property than with the spectral norm.*
8. *The map* $Y \mapsto \left(Y^\top Y/n - zI_p\right)^{-1}$ *is* $\frac{2^{3/2}|z|^{1/2}}{\Im(z)^2\sqrt{n}}$*-Lipschitz with respect to the Frobenius norm ([Cho22, Proposition 4.3]). In particular, if* $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$, *then* $\mathcal{G}_{Y^\top Y/n}(z) \propto_{\|\cdot\|_F} \mathcal{E}\left(\frac{|z|^{1/2}}{\Im(z)^2\sqrt{n}}\right)$.
9. *The map* $M \mapsto \frac{1}{n}\mathrm{Tr}(M)$ *is* $1/\sqrt{n}$*-Lipschitz in Frobenius norm. In particular, if* $\mathcal{G}_K(z) \propto_{\|\cdot\|_F} \mathcal{E}(\sigma_n)$ *for some random matrix* $K \in \mathbb{R}^{n\times n}$, *then* $g_K(z) \propto \mathcal{E}(\sigma_n/\sqrt{n})$.
10. *If a random variable* X *is* $\propto \mathcal{E}(\sigma_n)$ *concentrated, then* $|X - \mathbb{E}[X]| \leq O(\sqrt{\log n}\,\sigma_n)$ *almost surely eventually ([Cho22, Proposition 3.3]).*

We will also need a result for a product of matrices, and introduce for this reason a notion of conditional concentration. If $\mathcal{B}$ is a measurable subset of the universe Ω with $\mathbb{P}(\mathcal{B}) > 0$, the random matrix X conditioned with the event $\mathcal{B}$, denoted as $(X|\mathcal{B})$, designates the measurable mapping $(\mathcal{B}, \mathcal{F}_\mathcal{B}, \mathbb{P}/\mathbb{P}(\mathcal{B})) \mapsto E_n$ satisfying : $\forall \omega \in \mathcal{B}$, $(X|\mathcal{B})(\omega) = X(\omega)$.

Proposition 3.2.3 (Proposition 9 in [LC20])**.** *If* X *and* Y *are sequences of random matrices with dimensions bounded by* $O(n)$, *such that* $(X, Y) \propto_{\|\cdot\|_F} \mathcal{E}(1)$, $\mathbb{E}[|||X|||] \leq O(\sqrt{n})$ *and* $\mathbb{E}[|||Y|||] \leq O(\sqrt{n})$, *then there exists a constant* $c > 0$ *such that :*

1. $\mathbb{E}[|||X|||] \leq c\sqrt{n}$ *and* $\mathbb{E}[|||Y|||] \leq c\sqrt{n}$.
2. *With* $\mathcal{B} = \{|||X||| \text{ and } |||Y||| \leq 2c\sqrt{n}\})$, $\mathbb{P}(\mathcal{B}^c) \leq ce^{-n/c}$.
3. $(XY|\mathcal{B}) \propto_{\|\cdot\|_F} \mathcal{E}(\sqrt{n})$.

3.2.2 Polynomial bounds in z and notation $O_z(\epsilon_n)$

Throughout this article we will deal with quantitative estimates involving both a dimension parameter $n \in \mathbb{N}$ and a spectral parameter $z \in \mathbb{C}^+$. As it turns out, for many applications it is not useful to track the exact dependence in z, which motivates the following definition :

Definition 3.2.4. Let $\zeta : \mathbb{N} \times \mathbb{C}^+ \to \mathbb{R}^+$ be a function and $\epsilon_n > 0$ a sequence. We say that ζ is bounded by ϵ_n in n and polynomially in z, and we note $\zeta(n, z) \leq O_z(\epsilon_n)$, if there exists $\alpha \geq 0$ such that, uniformly in $n \in \mathbb{N}$ and $z \in \mathbb{C}^+$ with bounded $\Im(z)$:

$$\zeta(n, z) \leq O\left(\epsilon_n \frac{|z|^\alpha}{\Im(z)^{2\alpha}}\right).$$

We say that a family of functions is uniformly $O_z(\epsilon_n)$ bounded if the above bound holds uniformly for any function of the family.

Remark 3.2.5. By "uniformly in $n \in \mathbb{N}$ and $z \in \mathbb{C}^+$ with bounded $\Im(z)$", we mean that for any fixed $\tau > 0$, the bound holds uniformly in the parameters n and z belonging to the domain $\{z \in \mathbb{C} \text{ such that } 0 < \Im(z) \leq \tau\}$. We allow in particular z to vary with n. The implicit constant in the $O(\cdot)$ notation may however depend on τ.

By the way, in such a domain, we have $\frac{|z|}{\Im(z)} \geq 1$ and $\frac{1}{\Im(z)} \geq 1/\tau$. As a consequence, if $\alpha' \geq \alpha \geq 0$, then $\frac{|z|^\alpha}{\Im(z)^{2\alpha}} \leq \tau^{\alpha'-\alpha} \frac{|z|^{\alpha'}}{\Im(z)^{2\alpha'}}$. The classical rules of calculus thus apply to our notation : $O_z(\epsilon_n) + O_z(\epsilon'_n) = O_z(\epsilon_n + \epsilon'_n)$ and $O_z(\epsilon_n)O_z(\epsilon'_n) = O_z(\epsilon_n \epsilon'_n)$.

The next technical lemma will be key to translate results like the deterministic equivalent Theorem 3.4.3 into simplified versions using our $O_z(\epsilon_n)$ notations.

Lemma 3.2.6. *If for some constants $\alpha_0, \alpha_1, \ldots, \alpha_6 > 0$, the function ζ satisfies an* a priori *inequality $|\zeta(n,z)| \leq O\Big(\frac{|z|^{\alpha_1}}{\epsilon_n^{\alpha_0} \Im(z)^{\alpha_1+\alpha_2}}\Big)$, and if the bound :*

$$\zeta(n,z) \leq O\left(\epsilon_n \frac{|z|^{\alpha_3}}{\Im(z)^{\alpha_3+\alpha_4}}\right)$$

holds uniformly for $z \in \mathbb{C}^+$ with bounded $\Im(z)$ and satisfying $\epsilon_n \frac{|z|^{\alpha_5}}{\Im(z)^{\alpha_5+\alpha_6}} \leq c$ for some constant $c > 0$, then for some exponent $\alpha > 0$, the bound :

$$\zeta(n,z) \leq O\left(\epsilon_n \frac{|z|^{\alpha}}{\Im(z)^{2\alpha}}\right)$$

holds uniformly in $z \in \mathbb{C}^+$ with bounded $\Im(z)$, that is $\zeta(n,z) \leq O_z(\epsilon_n)$.

Proof. Let us choose $\alpha = \max\left(\alpha_5(\alpha_0+1) + \alpha_1, \alpha_6(\alpha_0+1) + \alpha_2, \alpha_3, \alpha_4\right)$. If $\epsilon_n \frac{|z|^{\alpha_5}}{\Im(z)^{\alpha_5+\alpha_6}} \leq c$, we use the main inequality on ζ :

$$\zeta(n,z) \leq O\left(\epsilon_n \frac{|z|^{\alpha_3}}{\Im(z)^{\alpha_3+\alpha_4}}\right) \leq O\left(\epsilon_n \frac{|z|^{\alpha}}{\Im(z)^{2\alpha}}\right).$$

In the case where $\epsilon_n \frac{|z|^{\alpha_5}}{\Im(z)^{\alpha_5+\alpha_6}} \geq c$, we use the *a priori* inequality on ζ :

$$\begin{aligned}
|\zeta(n,z)| &\leq O\left(\frac{|z|^{\alpha_1}}{\epsilon_n^{\alpha_0} \Im(z)^{\alpha_1+\alpha_2}}\right) \\
&\leq O\left(\epsilon_n \left(\frac{|z|^{\alpha_5}}{c\, \Im(z)^{\alpha_5+\alpha_6}}\right)^{\alpha_0+1} \frac{|z|^{\alpha_1}}{\Im(z)^{\alpha_1+\alpha_2}}\right) \\
&\leq O\left(\epsilon_n \frac{|z|^{\alpha}}{\Im(z)^{2\alpha}}\right).
\end{aligned}$$

□

Remark 3.2.7. As we will see later in this article, the lemma is particularly useful to simplify some quantitative statements that require the spectral parameter z to be away from the real axis. Indeed the Stieltjes tranforms and resolvent matrices satisfy the classical bounds $|g(z)| \leq 1/\Im(z)$ and $\|\mathcal{G}(z)\|_F \leq \sqrt{n}\|\mathcal{G}(z)\| \leq \sqrt{n}/\Im(z)$, which translate into *a priori* inequalities for any difference of such objects.

Let us wrap this subsection by giving a general setting on which $O_z(\epsilon_n)$ bounds between Stieltjes transforms imply polynomial bounds in Kolmogorov distance. We remind our reader that for measures ν and μ with cumulative distribution functions $\mathcal{F}_\nu$ and $\mathcal{F}_\mu$, the Kolmogorov distance is defined as $D(\mu,\nu) = \sup_{t\in\mathbb{R}} |F_\nu(t) - \mathcal{F}_\mu(t)|$. The following result is an immediate consequence of [Cho22, Theorem 3.6].

Proposition 3.2.8. *Let μ_n and ν_n be sequences of probability measures on $\mathbb{R}^+$ such that :*

1. $\int_{\mathbb{R}} |\mathcal{F}_{\mu_n}(t) - \mathcal{F}_{\nu_n}(t)|dt < \infty$.
2. $\mathcal{F}_{\nu_n}$ *are uniformly Hölder continuous with exponent* $\beta \in (0,1]$.
3. μ_n *and* ν_n *have uniformly bounded second moments.*
4. $|g_{\mu_n}(z) - g_{\nu_n}(z)| \leq O_z(\epsilon_n)$ *for some sequence* $\epsilon_n \to 0$.

Then there exists $\theta > 0$ such that $D(\mu_n,\nu_n) \leq O(\epsilon_n^\theta)$.

Remark 3.2.9. In some cases, the uniform Hölder continuity hypothesis may be adapted using a simple change of measures. For instance, given shape parameters $\gamma_n > 1$ and measures τ_n supported on the same compact of $(0,\infty)$, the cumulative distribution function of $\nu_n = \mathrm{MP}(\gamma_n) \boxtimes \tau_n$ are not even continuous at 0. However we can consider the new measures $\check{\nu}_n = (1-\gamma_n)\cdot\delta_0 + \gamma_n \cdot \nu_n$, and use the fact that $\mathcal{F}_{\check{\nu}_n}$ are uniformly 1/2-Hölder continuous to deduce the same result (see [Cho22, Section 8.3]).

A variant of this result for empirical spectral distributions may also be adapted, see Proposition 3.8.1 in the Appendix.

3.3 Covariance matrices of coordinate-wise functions of Gaussian vectors

In this section we consider the random vector $y = f(u) \in \mathbb{R}^n$, obtained by applying a real function on each of the coordinates of a centered Gaussian vector $u \in \mathbb{R}^n$. We are particularly interested in the case where the entries of u are weakly correlated. In this instance, we will give approximations of the covariance matrix $\Sigma = \mathbb{E}\left[yy^\top\right]$, with a control of the error either entry-wise, or in spectral norm, or for their respective Stieltjes transforms. We first present the tools making these approximations possible.

3.3.1 Hermite polynomials

Let us remind the notation $\mathcal{N}$ for a standard Gaussian random variable. We denote by $\mathcal{H}$ the Hilbert space of real Borel functions such that $\mathbb{E}[f(\mathcal{N})^2] < \infty$, endowed with the Gaussian inner product :

$$< f, g >= \mathbb{E}[f(\mathcal{N})g(\mathcal{N})] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t)g(t)e^{-t^2/2}dt.$$

We can remark that all Lipschitz continuous functions automatically belong to $\mathcal{H}$. The r-th non-normalized Hermite polynomial is by definition :

$$h_r(t) = (-1)^r e^{t^2/2} \frac{\mathrm{d}^r}{\mathrm{d}t^r}\left(e^{-t^2/2}\right).$$

The first Hermite polynomials are : $h_0 = 1$, $h_0 = \mathbf{X}$, $h_2 = \mathbf{X}^2 - 1$, and $h_3 = \mathbf{X}^3 - 3\mathbf{X}$. The r-th normalized Hermite polynomial is $\hat{h}_r = \frac{h_r}{\sqrt{r!}}$, and the r-th Hermite coefficient of a function $f \in \mathcal{H}$ is $\zeta_r(f) =< f, \hat{h}_r >$. Hereafter we remind some classical properties of Hermite polynomials without proofs. For more details, we invite our reader to consult [San59, Chapter 4].

Proposition 3.3.1. *(i) h_r are monic polynomials of degree r. $(\hat{h}_r)_{r\in\mathbb{N}}$ form a complete orthonormal basis of $\mathcal{H}$.*

(ii) Every function of $\mathcal{H}$ can be expanded as the converging sum in $\mathcal{H}$: $f = \sum_{r\geq 0} \zeta_r(f)\hat{h}_r$. In particular $\|f\|^2_{\mathcal{H}} = \mathbb{E}[f(\mathcal{N})^2] = \sum_{r\geq 0} \zeta_r(f)^2$.

(iii) The Hermite polynomials satisfy the relations :

$$h'_r(t) = rh_{r-1}(t),$$
$$h_{r+1}(t) = th_r(t) - h'_r(t).$$

We move on to more specific properties that shall be used in the course of this section.

Lemma 3.3.2 (Lemma D.2 in [NM20])**.** *If z_1 and z_2 are standard Gaussian random variables, such that (z_1, z_2) forms a Gaussian vector, then :*

$$\mathbb{E}\Big[\hat{h}_r(z_1)\hat{h}_s(z_2)\Big] = \mathbf{1}_{r=s}\,\mathrm{Cov}[z_1, z_2]^r.$$

Proposition 3.3.3. *Let us fix $f \in \mathcal{H}$, and for $r \in \mathbb{N}$ let $\Psi_r(\sigma) = \sigma^{-r}\mathbb{E}[f(\sigma\mathcal{N})h_r(\mathcal{N})]$. Then Ψ_r is $\mathcal{C}^\infty$ on $(0,\infty)$ and $\Psi_r'(\sigma) = \sigma\Psi_{r+2}(\sigma)$.*

Proof. From the identity $\mathbb{E}[f(\sigma\mathcal{N})h_r(\mathcal{N})] = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} f(t)h_r(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}dt$ and the Leibniz integral rule, we deduce that Ψ_r is $\mathcal{C}^\infty$ and that :

$$\begin{aligned}\Psi_r'(\sigma) &= \frac{d}{d\sigma}\left(\frac{1}{\sigma^{r+1}\sqrt{2\pi}}\int_{\mathbb{R}} f(t)h_r(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}dt\right)\\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(t)\frac{d}{d\sigma}\left(\frac{1}{\sigma^{r+1}}h_r(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}\right)dt.\end{aligned}$$

The derivative inside the integral can be computed in a few steps. Firstly, from the definition $\frac{d}{dt}\Big(h_r(t)e^{-t^2/2}\Big) = (-1)^r\frac{d^{r+1}}{dt^{r+1}}\Big(e^{-t^2/2}\Big) = -h_{r+1}(t)e^{-t^2/2}$. Then applying the identity $th_{r+1}(t) = h_{r+2}(t) + (r+1)h_r(t)$, which can be deduced from Proposition 3.3.1, leads to :

$$\begin{aligned}\frac{d}{d\sigma}\left(h_r(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}\right) &= \frac{t}{\sigma^2}h_{r+1}(t/\sigma)e^{-t^2/2\sigma^2}\\ &= \frac{1}{\sigma}(h_{r+2}(t/\sigma) + (r+1)h_r(t/\sigma))e^{-\frac{t^2}{2\sigma^2}},\\ \frac{d}{d\sigma}\left(\frac{1}{\sigma^{r+1}}h_r(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}\right) &= \frac{1}{\sigma^{r+2}}(h_{r+2}(t/\sigma) + (r+1)h_r(t/\sigma))e^{-\frac{t^2}{2\sigma^2}}\\ &\quad - \frac{r+1}{\sigma^{r+2}}h_r(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}\\ &= \frac{1}{\sigma^{r+2}}h_{r+2}(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}.\end{aligned}$$

We obtain finally :

$$\begin{aligned}\Psi_r'(\sigma) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(t)\frac{1}{\sigma^{r+2}}h_{r+2}(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}dt\\ &= \sigma\cdot\left(\frac{1}{\sigma^{r+3}\sqrt{2\pi}}\int_{\mathbb{R}} f(t)h_{r+2}(t/\sigma)e^{-\frac{t^2}{2\sigma^2}}dt\right)\\ &= \sigma\cdot\Psi_{r+2}(\sigma).\end{aligned}$$

□

Corollary 3.3.4. *Let $f \in \mathcal{H}$, and for $\sigma > 0$ let $f_\sigma : t \mapsto f(\sigma t)$. Then for any integer $r \geq 0$:*

$$\begin{aligned}\zeta_r(f_\sigma) &= \sigma^r(\zeta_r(f) + O(\sigma - 1))\\ &= \sigma^r\Big(\zeta_r(f) + \sqrt{(r+1)(r+2)}(\sigma-1)\zeta_{r+2}(f) + O((\sigma-1)^2)\Big).\end{aligned}$$

Proof. These expressions are straightforward consequences of first and second order Taylor expansions of Ψ_r. For instance, regarding the second formula :

$$\begin{aligned}\zeta_r(f_\sigma) &= \frac{\sigma^r}{\sqrt{r!}}\Psi_r(\sigma)\\ &= \frac{\sigma^r}{\sqrt{r!}}\Big(\Psi_r(1) + (\sigma-1)\Psi_{r+2}(1) + O((\sigma-1)^2)\Big)\\ &= \sigma^r\Big(\zeta_r(f) + \sqrt{(r+1)(r+2)}(\sigma-1)\zeta_{r+2}(f) + O((\sigma-1)^2)\Big).\end{aligned}$$

□

3.3.2 Iterated Hadamard products

In what follows, $M^{\circ r}$ denotes the Hadamard product of r copies of a matrix $M \in \mathbb{R}^{n\times n}$. We denote by $\mathrm{diag}(M)$ and $\mathrm{off}(M)$ respectively the diagonal and off-diagonal sub-matrices of M. $\vec{\mathrm{diag}}(M) \in \mathbb{R}^n$ denotes the diagonal elements of M reshaped as a vector in $\mathbb{R}^n$.

Lemma 3.3.5. *If $\Delta \in \mathbb{R}^{n\times n}$ is a symmetric matrix such that $I_n+\Delta$ is positive semi-definite, then for any integer $r \geq 1$:*

$$\begin{aligned}\left|\!\left|\!\left|\Delta^{\circ r+1}\right|\!\right|\!\right| &\leq (1+\|\Delta\|_{\max})|||\Delta^{\circ r}||| + \|\Delta\|_{\max}^r,\\ \left|\!\left|\!\left|\Delta^{\circ 2r}\right|\!\right|\!\right| &\leq \|\Delta\|_{\max}^{2r-2}|||\Delta\circ\Delta|||.\end{aligned}$$

Proof. For any matrices A and $B \in \mathbb{R}^{n\times n}$ with A positive semi-definite, the inequality $|||A\circ B||| \leq \|A\|_{\max}|||B|||$ holds true ([Joh90, Proposition 3.7.9.]). As a consequence :

$$\begin{aligned}\left|\!\left|\!\left|\Delta^{\circ r+1}\right|\!\right|\!\right| &\leq |||(I_n+\Delta)\circ\Delta^{\circ r}||| + |||I_n\circ\Delta^{\circ r}|||\\ &\leq \|I_n+\Delta\|_{\max}|||\Delta^{\circ r}||| + |||\mathrm{diag}(\Delta)^r|||\\ &\leq (1+\|\Delta\|_{\max})|||\Delta^{\circ r}||| + \|\Delta\|_{\max}^r.\end{aligned}$$

For the second inequality, since $\Delta^{\circ 2r}$ has non negative entries, there is $\alpha \in (\mathbb{R}^+)^n$ such that $\|\alpha\| = 1$ and $|||\Delta^{\circ 2r}||| = \|\Delta^{\circ 2r}\alpha\|$. We deduce that :

$$\begin{aligned}\left|\!\left|\!\left|\Delta^{\circ 2r}\right|\!\right|\!\right| &= \left(\sum_{i=1}^n\left(\sum_{j=1}^n \Delta_{ij}^{2r}\alpha_j\right)^2\right)^{1/2}\\ &\leq \|\Delta\|_{\max}^{2r-2}\left(\sum_{i=1}^n\left(\sum_{j=1}^n \Delta_{ij}^{2}\alpha_j\right)^2\right)^{1/2}\\ &\leq \|\Delta\|_{\max}^{2r-2}|||\Delta\circ\Delta|||.\end{aligned}$$

□

Proposition 3.3.6. *If $\Delta \in \mathbb{R}^{n\times n}$ is a sequence of symmetric matrices such that $I_n + \Delta$ is positive semi-definite, $\|\Delta\|_{\max}$ converges to 0 and $|||\Delta|||$ is bounded, then for any $r \geq 1$ the quantities $|||\Delta^{\circ 2r}|||$, $|||\mathrm{off}(\Delta^{\circ 2r})|||$, $|||\Delta^{\circ 2r+1}|||$ and $|||\mathrm{off}(\Delta^{\circ 2r+1})|||$ are all bounded by $O\Big(\|\Delta\|_{\max}^{2r-2}\Big)$ when n goes to ∞.*

Proof. According to the preceding lemma, $|||\Delta \circ \Delta||| \leq (1 + \|\Delta\|_{\max})|||\Delta||| + \|\Delta\|_{\max} \leq O(1)$. This implies that : $|||\Delta^{\circ 2r}||| \leq \|\Delta\|_{\max}^{2r-2}|||\Delta \circ \Delta||| \leq O\Big(\|\Delta\|_{\max}^{2r-2}\Big)$, and $|||\Delta^{\circ 2r+1}||| \leq (1 + \|\Delta\|_{\max})|||\Delta^{\circ 2r}|||+\|\Delta\|_{\max}^{2r} \leq O\Big(\|\Delta\|_{\max}^{2r-2}\Big)$. The remaining results follow after remarking that $|||\mathrm{diag}(\Delta^{\circ r})||| \leq O(\|\Delta\|_{\max}^{r})$. □

3.3.3 General expansion of the covariance matrix Σ

In the following paragraphs we are considering a function $f \in \mathcal{H}$, and $u \in \mathbb{R}^n$ a centered Gaussian vector with covariance matrix $S \in \mathbb{R}^{n\times n}$. The random vector $y = f(u)$ is obtained by applying the function f entry-wise on u, and we let $\Sigma = \mathbb{E}\Big[yy^\top\Big]$. For $1 \leq i \leq n$, we denote by f_i the function $f_i : t \mapsto f\Big(S_{ii}^{1/2}\, t\Big)$, and by $D_r \in \mathbb{R}^{n\times n}$ the diagonal matrix with entries ${S_{ii}}^{-r/2}\zeta_r(f_i)$.

Proposition 3.3.7. *The convergence of the following sum holds entry-wise :*

$$\Sigma = \sum_{r\geq 0} D_r\, S^{\circ r} D_r.$$

Proof. The random variables $\tilde{u}_i = S_{ii}^{-1/2}u_i$ are standard Gaussian random variables, and they form a Gaussian vector. Using Lemma 3.3.2, we have :

$$\begin{aligned}
\Sigma_{ij} &= \mathbb{E}[f_i(\tilde{u}_i)f_j(\tilde{u}_j)] \\
&= \sum_{r,s\geq 0} \zeta_r(f_i)\zeta_s(f_j)\mathbb{E}\Big[\hat{h}_r(\tilde{u}_i)\hat{h}_s(\tilde{u}_j)\Big] \\
&= \sum_{r\geq 0} \zeta_r(f_i)\zeta_r(f_j)\,\mathrm{Cov}[\tilde{u}_i, \tilde{u}_j]^r \\
&= \sum_{r\geq 0} \zeta_r(f_i)\zeta_r(f_j){S_{ii}}^{-r/2}{S_{jj}}^{-r/2}{S_{ij}}^{r} \\
&= \sum_{r\geq 0} (D_r)_i(D_r)_j(S^{\circ r})_{ij},
\end{aligned}$$

which proves the proposition. □

3.3.4 Approximation of Σ for weakly correlated Gaussian vectors

We will now explain how the expansion of Σ given by Proposition 3.3.7 can be simplified when u has weakly correlated entries, in the sense that its

covariance matrix S is close to I_n. To this end we let $\Delta = S - I_n$, and :

$$\Sigma_{\text{approx}} = \|f\|_{\mathcal{H}}^2 I_n + \frac{\zeta_2(f)^2}{2}\vec{\text{diag}}(\Delta)\vec{\text{diag}}(\Delta)^\top + \sum_{r=1,2,3}\zeta_r(f)^2\Delta^{\circ r}$$

Assumptions 3.3.8.
1. f is Lipschitz and Gaussian centered, i.e. $\zeta_0(f) = \mathbb{E}[f(\mathcal{N})] = 0$.
2. $|||\Delta|||$ and $\left\|\vec{\text{diag}}(\Delta)\right\|$ are bounded.
3. $\|\Delta\|_{\max}$ converges to 0.

Theorem 3.3.9. *Under Assumptions 3.3.8,* $|||\Sigma - \Sigma_{\text{approx}}||| \leq O(\|\Delta\|_{\max})$.

Proof. Let $\epsilon_n = \|\Delta\|_{\max}$. Since f is Lipschitz, $\|f - f_i\|_{L^2} = O\left(\left|1 - S_{ii}^{1/2}\right|\right) = O(\epsilon_n)$. We deduce that $|||\text{diag}(\Sigma) - \|f\|I_n||| = \max_{1\leq i\leq n}|\|f_i\| - \|f\|_{\mathcal{H}}| \leq O(\epsilon_n)$. For the off-diagonal terms, let ν be the vector in $\mathbb{R}^n$ defined by its coordinates $\nu_i = \zeta_0(f_i)$. Let also $\mu = \frac{\zeta_2(f)^2}{2}\vec{\text{diag}}(\Delta)$. The decomposition :

$$\text{off}(\Sigma) = \text{off}\left(\nu\nu^\top\right) + \sum_{r\geq 1} D_r\,\text{off}(\Delta^{\circ r})\,D_r$$

follows easily from the expansion of Σ given by Proposition 3.3.7. Since $\zeta_0(f) = 0$ and $S_{ii}^{1/2} = 1 + \frac{\Delta_{ii}}{2} + O(\Delta_{ii}^2)$, using the second order expansion given by Corollary 3.3.4 applied to the function f, uniformly in $i \in [\![1, n]\!]$ we have :

$$\begin{aligned}\nu_i = \zeta_0(f_i) &= \sqrt{2}\left(S_{ii}^{1/2} - 1\right)\zeta_2(f) + O\left(\left(S_{ii}^{1/2} - 1\right)^2\right)\\ &= \frac{\zeta_2(f)}{\sqrt{2}}\Delta_{ii} + O(\Delta_{ii}^2)\\ &= \mu_i + O(\Delta_{ii}^2).\end{aligned}$$

Therefore $\|\nu - \mu\|^2 = O(\sum_{i=1}^n \Delta_{ii}^4) \leq O(\max_{1\leq i\leq n}\Delta_{ii}^2)\left\|\vec{\text{diag}}(\Delta)\right\|^2 \leq O(\epsilon_n^2)$, and :

$$\left\|\text{off}\left(\nu\nu^\top\right) - \mu\mu^\top\right\| \leq \|\nu - \mu\|(\|\nu\| + \|\mu\|) + \left\|\text{diag}\left(\nu\nu^\top\right)\right\| \leq O(\epsilon_n)$$

For the remaining sum $\sum_{r\geq 1} D_r\,\text{off}(\Delta^{\circ r})\,D_r$, we use again that $|\zeta_r(f_i) - \zeta_r(f)| \leq \|f_i - f\|_{\mathcal{H}} \leq O(\epsilon_n)$, hence $|||D_r - \zeta_r(f)I_n||| \leq O(\epsilon_n)$, and in particular $|||D_r||| \leq O(1)$. Using Proposition 3.3.6, for $r = 1, 2$ or 3 :

$$\begin{aligned}\left\|D_r\,\text{off}(\Delta^{\circ r})\,D_r - \zeta_r(f)^2\Delta^{\circ r}\right\| &\leq |||D_r|||^2|||\text{diag}(\Delta^{\circ r})|||\\ &\quad + |||D_r - \zeta_r(f)I_n|||\,|||\Delta^{\circ r}|||\,(|||D_r||| + |\zeta_r(f)|)\\ &\leq O(\epsilon_n).\end{aligned}$$

Finally for the remaining terms, using Proposition 3.3.6 for $r \geq 4$:

$$\left\|\sum_{r\geq 4} D_r\,\text{off}(\Delta^{\circ r})\,D_r\right\| \leq \sum_{r\geq 2} O\left(\epsilon_n^{2r-2}\right) \leq O\left(\epsilon_n^2\right).$$

□

3.3.5 Linearization of Σ in specific settings

We remind our reader the definition of $\Delta = S - I_n$, and we let :

$$\Sigma_{\text{lin}} = \|f\|_{\mathcal{H}}^2 I_n + \zeta_1(f)^2 \Delta$$

The matrix Σ_{lin} is obtained by means of usual matrix operations, easy to handle with classical random matrix theory tools, in contrast to Σ_{approx} which involves Hadamard products. Unfortunately, $|||\Sigma_{\text{approx}} - \Sigma_{\text{lin}}|||$ may not converge to 0 under the mere Assumptions 3.3.8. We can nonetheless identify conditions involving Δ and f, under which Σ_{lin} is a good approximation of Σ, either entry-wise, or in spectral norm, or for their respective Stieltjes transforms.

Proposition 3.3.10. *Under Assumptions 3.3.8, $|||\Sigma - \Sigma_{\text{lin}}|||$ is bounded, and with $\epsilon_n = \zeta_2(f)^2 \|\Delta\|_{\max}^2 + \zeta_3(f)^2 \|\Delta\|_{\max}^3$:*

$$\begin{aligned} \|\Sigma - \Sigma_{\text{lin}}\|_{\max} &\leq O(\|\Delta\|_{\max}), \\ |||\Sigma - \Sigma_{\text{lin}}||| &\leq O(\|\Delta\|_{\max} + n\epsilon_n), \\ |g_\Sigma(z) - g_{\Sigma_{\text{lin}}}(z)| &\leq \frac{1}{\Im(z)^2} O\Big(\|\Delta\|_{\max} + \sqrt{n}\epsilon_n\Big). \end{aligned}$$

Proof. We start from the expression :

$$\Sigma_{\text{approx}} - \Sigma_{\text{lin}} = \frac{\zeta_2(f)^2}{2} \vec{\text{diag}}(\Delta) \vec{\text{diag}}(\Delta)^\top + \zeta_2(f)^2 \Delta^{\circ 2} + \zeta_3(\tilde{f})^2 \Delta^{\circ 3}$$

The matrices $\vec{\text{diag}}(\Delta)\vec{\text{diag}}(\Delta)^\top$, $\Delta^{\circ 2}$ and $\Delta^{\circ 3}$ are all bounded in spectral norm, thus $|||\Sigma_{\text{approx}} - \Sigma_{\text{lin}}|||$ is bounded and from Theorem 3.3.9 $|||\Sigma - \Sigma_{\text{lin}}|||$ is always bounded. The bound $\|\Sigma_{\text{approx}} - \Sigma_{\text{lin}}\|_{\max} \leq O(\epsilon_n)$ is immediate given the above expression for the difference of these matrices, and from Theorem 3.3.9 :

$$\begin{aligned} \|\Sigma - \Sigma_{\text{lin}}\|_{\max} &\leq |||\Sigma - \Sigma_{\text{approx}}||| + \|\Sigma_{\text{approx}} - \Sigma_{\text{lin}}\|_{\max} \\ &\leq O(\|\Delta\|_{\max} + \epsilon_n) \leq O(\|\Delta\|_{\max}). \end{aligned}$$

The other bounds follow from classical matrix inequalities. Indeed :

$$\begin{aligned} |||\Sigma - \Sigma_{\text{lin}}||| &\leq |||\Sigma - \Sigma_{\text{approx}}||| + |||\Sigma_{\text{approx}} - \Sigma_{\text{lin}}||| \\ &\leq O(\|\Delta\|_{\max}) + n\|\Sigma_{\text{approx}} - \Sigma_{\text{lin}}\|_{\max} \\ &\leq O(\|\Delta\|_{\max} + n\epsilon_n). \end{aligned}$$

Finally for the inequality on Stieltjes transforms, for any $A, B \in \mathbb{R}^{n \times n}$ and $z \in \mathbb{C}^+$:

$$\begin{aligned} |g_A(z) - g_B(z)| &= \left| \frac{1}{n} \text{Tr}(\mathcal{G}_A(z)(B - A)\mathcal{G}_B(z)) \right| \\ &\leq \frac{1}{n} |||\mathcal{G}_A(z)||| \|I_n\|_F \|B - A\|_F |||\mathcal{G}_B(z)||| \\ &\leq \frac{\|B - A\|_F}{\sqrt{n}\Im(z)^2}. \end{aligned}$$

We deduce that :

$$\begin{aligned}
|g_\Sigma(z) - g_{\Sigma_{\text{lin}}}(z)| &\leq \Big|g_\Sigma(z) - g_{\Sigma_{\text{approx}}}(z)\Big| + \Big|g_{\Sigma_{\text{approx}}}(z) - g_{\Sigma_{\text{lin}}}(z)\Big| \\
&\leq \frac{\|\Sigma - \Sigma_{\text{lin}}\|_F}{\sqrt{n}\Im(z)^2} + \frac{\|\Sigma_{\text{approx}} - \Sigma_{\text{lin}}\|_F}{\sqrt{n}\Im(z)^2} \\
&\leq \frac{|||\Sigma - \Sigma_{\text{lin}}|||}{\Im(z)^2} + \sqrt{n}\frac{\|\Sigma_{\text{approx}} - \Sigma_{\text{lin}}\|_{\max}}{\Im(z)^2} \\
&\leq \frac{1}{\Im(z)^2} O\Big(\|\Delta\|_{\max} + \sqrt{n}\epsilon_n\Big).
\end{aligned}$$

□

Remark 3.3.11. In practice, $|||\Sigma - \Sigma_{\text{lin}}|||$ converges to 0, provided one of the following additional hypothesis holds true :

— $\|\Delta\|_{\max} = o\Big(n^{-1/2}\Big)$,

— $\zeta_2(f) = 0$ and $\|\Delta\|_{\max} = o\Big(n^{-1/3}\Big)$,

— $\zeta_2(f) = \zeta_3(f) = 0$.

For the Stieltjes transforms, $g_\Sigma(z) - g_{\Sigma_{\text{lin}}}(z)$ converges to 0 pointwise, and thus Σ and Σ_{lin} have the same limiting spectral distribution if it exists, provided one of the following additional hypothesis holds true :

— $\|\Delta\|_{\max} = o\Big(n^{-1/4}\Big)$,

— $\zeta_2(f) = 0$ and $\|\Delta\|_{\max} = o\Big(n^{-1/6}\Big)$,

— $\zeta_2(f) = \zeta_3(f) = 0$.

3.4 Deterministic equivalent of sample covariance matrices with a general dependence structure

In this section we recall the latest results about the deterministic equivalent of the resolvents of sample covariance matrices on which this article is based. These estimates were first established in [LC21] with a convergence speed in n only, and later complemented with quantitative estimates in the spectral parameter z in [Cho22]. In a second step, we will thoroughly study the properties of the deterministic equivalent matrices appearing in these results.

Given $\Sigma \in \mathbb{R}^{n\times n}$ a positive semi-definite matrix and a sequence of shape parameters $\gamma_n > 0$, we build the matrix function $\mathbf{G}^{\Sigma}_{\boxtimes} : \mathbb{C}^+ \to \mathbb{C}^{n\times n}$ from the following objects :

$$\begin{aligned}
\nu^{\Sigma} &= \mathrm{MP}(\gamma_n) \boxtimes \mu_{\Sigma}, \\
\check{\nu}^{\Sigma} &= (1-\gamma_n)\cdot\delta_0 + \gamma_n \cdot \nu^{\Sigma}, \\
l_{\check{\nu}^{\Sigma}}(z) &= -1/g_{\check{\nu}^{\Sigma}}(z), \\
\mathbf{G}^{\Sigma}_{\boxtimes}(z) &= (-z g_{\check{\nu}^{\Sigma}}(z)\Sigma - zI_n)^{-1}, \\
&= z^{-1} l_{\check{\nu}^{\Sigma}}(z) \mathcal{G}_L(l_{\check{\nu}^{\Sigma}}(z)).
\end{aligned}$$

Let us state without proofs some useful properties of these objects. $\check{\nu}^{\Sigma}$ is always a true probability measure ([Cho22, Lemma 6.1]). The usual resolvent inequality $\left\|\mathbf{G}^{\Sigma}_{\boxtimes}(z)\right\| \leq 1/\Im(z)$ holds true ([Cho22, Lemma 5.1]), and $\mathrm{Tr}\mathbf{G}^{\Sigma}_{\boxtimes}(z) = n g_{\nu^{\Sigma}}(z)$ ([Cho22, Proposition 6.2]).

Remark 3.4.1. The notation $\mathbf{G}^{\Sigma}_{\boxtimes}(z)$ is inspired from the free probability theory, which appears as the profound canvas hidden within the above definitions. Indeed it is not difficult to see that the pair $\left(g_{\check{\nu}^{\Sigma}}(z), \mathbf{G}^{\Sigma}_{\boxtimes}(z)\right)$ satisfies the system of self-consistent equations :

$$\begin{aligned}
g_{\check{\nu}^{\Sigma}}(z) &= \left(-z - z\frac{\gamma_n}{n}\mathrm{Tr}\left(\mathbf{G}^{\Sigma}_{\boxtimes}(z)\Sigma\right)\right)^{-1}, \\
\mathbf{G}^{\Sigma}_{\boxtimes}(z) &= (-zI_n - z g_{\check{\nu}^{\Sigma}}(z)\Sigma)^{-1}.
\end{aligned}$$

This may be rewritten as the operator-valued self-consistent equation :

$$z\mathcal{H}(z) = I_{n+1} + z\eta(\mathcal{H}(z))\mathcal{H}(z),$$

where $\mathcal{H}(z) = \begin{pmatrix} g_{\check{\nu}^{\Sigma}}(z) & 0 \\ 0 & \mathbf{G}^{\Sigma}_{\boxtimes}(z) \end{pmatrix}$, and :

$$\begin{aligned}
\eta : \mathbb{C}\oplus\mathbb{C}^{n\times n} &\to \mathbb{C}\oplus\mathbb{C}^{n\times n}, \\
\begin{pmatrix} g & 0 \\ 0 & G \end{pmatrix} &\mapsto \begin{pmatrix} \frac{\gamma_n}{n}\mathrm{Tr}(G\Sigma) & 0 \\ 0 & g\Sigma \end{pmatrix}.
\end{aligned}$$

$z\mathcal{H}(z^2)$ thus corresponds to the resolvent of an operator-valued free semi-circular variable with covariance η (see [FOBS06, Section 3.3]). In this sense the definition of $\mathbf{G}_{\boxtimes}^{\Sigma}(z)$ extends the notion of free convolution, and it should come as no surprise that such objects appear as deterministic equivalents of sample covariance matrices.

3.4.1 General results

Let $Y \in \mathbb{R}^{d\times n}$ be a sequence of random matrices. The associated sample covariance matrix is $K = Y^\top Y/d$, and for $z \in \mathbb{C}^+$ we define its resolvent $\mathcal{G}_K(z) = (K - zI_n)^{-1}$ and Stieltjes transform $g_K(z) = (1/n)\mathrm{Tr}\mathcal{G}_K(z)$. From the expected covariance matrix $\Sigma = \mathbb{E}[K]$, we follow the above procedure to build the measure ν^Σ and the matrix function $\mathbf{G}_{\boxtimes}^{\Sigma}(z)$.

Assumptions 3.4.2.
1. $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$, the rows of Y are i.i.d. sampled from the distribution of a random vector y, with $\|\mathbb{E}[y]\|$ and $\left\|\left|\mathbb{E}[yy^\top]\right|\right\| = |||\Sigma|||$ bounded.
2. The ratio of dimensions $\gamma_n = \frac{n}{d}$ is bounded from above and away from 0.

Theorem 3.4.3. *[Cho22, Theorem 2.3] Uniformly under Assumptions 3.4.2, there exists $C > 0$ such that uniformly in $z \in \mathbb{C}^+$ with $\Im(z)$ bounded and $\frac{|z|^7}{\Im(z)^{16}} \leq Cn$, the following inequality holds :*

$$\left\|\mathbb{E}[\mathcal{G}_K(z)] - \mathbf{G}_{\boxtimes}^{\Sigma}(z)\right\|_F \leq O\left(\frac{|z|^{5/2}}{\Im(z)^9 n^{1/2}}\right).$$

Remark 3.4.4. By "uniformly under Assumptions 3.4.2", we mean that the implicit constants in the result only depend on the constants chosen in the assumptions. Said otherwise, if a family of random matrices satisfies Assumptions 3.4.2 uniformly, then the result holds uniformly for any matrix of this family.

Keeping track of the exponents appearing in this theorem will become more and more difficult as we go further into this article. We choose for this reason to work with a weaker notion of approximation using our concept of $O_z(\epsilon_n)$ polynomial bounds. As a side effect, we will see that we do not need to specify conditions on z anymore for the approximations to hold true.

Theorem 3.4.5 (Simplified version of the deterministic equivalent)**.** *Uniformly under Assumptions 3.4.2, the following concentration properties hold true :*

1. $g_K(z) \propto \mathcal{E}(O_z(1/n))$ *and* $\mathcal{G}_K(z) \propto_{\|\cdot\|_F} \mathcal{E}(O_z(1/\sqrt{n}))$.
2. $g_K(z) \in g_{\nu^\Sigma}(z) \pm \mathcal{E}(O_z(1/n))$ *and* $\mathcal{G}_K(z) \in_{\|\cdot\|_F} \mathbf{G}_{\boxtimes}^{\Sigma}(z) \pm \mathcal{E}(O_z(1/\sqrt{n}))$.

Proof. From Theorem 3.4.3, Lemma 3.2.6 and the bound $\left\|\mathbb{E}[\mathcal{G}_K(z)] - \mathbf{G}^{\Sigma}_{\boxtimes}(z)\right\|_F \leq 2\sqrt{n}/\Im(z)$, we have $\left\|\mathbb{E}[\mathcal{G}_K(z)] - \mathbf{G}^{\Sigma}_{\boxtimes}(z)\right\|_F \leq O_z(1/\sqrt{n})$. The general properties of concentration recalled in Proposition 3.2.2 prove the theorem. □

Similarly to [Cho22, Corolaries 2.5 and 2.8], we may deduce the following spectral properties of K. We will not prove this result immediately, but rather prompt our reader to consult the proof of Corollary 3.5.9 which is extremely similar.

Corollary 3.4.6. *Uniformly under Assumptions 3.4.2 :*

1. $|g_K(z) - g_{\nu^\Sigma}(z)| \leq O_z\left(\sqrt{\log n}/n\right)$ *almost surely eventually (a.s.e.), and* $\left\|\mathcal{G}_K(z) - \mathbf{G}^{\Sigma}_{\boxtimes}(z)\right\|_{\max} \leq O_z\left(\sqrt{\log n/n}\right)$ *a.s.e.*
2. *If the eigenvalues of* Σ *are bounded from below, there exists* $\theta > 0$ *such that* $D(\mu_K, \nu^\Sigma) \leq O(n^{-\theta})$ *a.s.e.*
3. *If additionally* $\mu_\Sigma \to \mu_\infty$ *weakly and* $\gamma_n \to \gamma_\infty$*, then* $\mu_K \to \nu_\infty = \mathrm{MP}(\gamma_\infty) \boxtimes \mu_\infty$ *weakly a.s., and more precisely :*

$$D(\mu_K, \nu_\infty) \leq D(\mu_\Sigma, \mu_\infty) + O(|\gamma_n - \gamma_\infty|) + O(n^{-\theta}) \quad \text{a.s.e.}$$

3.4.2 Regularity of the Stieltjes transform with respect to the free convolution

Let $\Sigma \in \mathbb{R}^{n\times n}$ be a sequence of positive semi-definite matrices, and let τ be probability measures such that $g_\Sigma(z) - g_\tau(z)$ converges to 0 pointwise in $\mathbb{C}^+$. Then μ_Σ and τ converge weakly to the same limit if it exists. The free multiplicative convolution measures $\nu^\Sigma = \mathrm{MP}(\gamma_n) \boxtimes \mu_\Sigma$ and $\chi = \mathrm{MP}(\gamma_n) \boxtimes \tau$ also converge weakly to the same limit, which is equivalent to their Stieltjes transforms $g_{\nu^\Sigma}(z)$ and $g_\chi(z)$ having the same limit pointwise. In this section we would like to refine this result by quantifying the convergence of $g_{\nu^\Sigma}(z) - g_\chi(z)$ to 0.

In the upcoming paragraphs, we always consider shape parameters γ_n that are bounded from above and away from 0, like in Assumptions 3.4.2. We remind our reader that we are dealing with sequences of matrices, measures, and complex functions, even if we sometimes omit the indices n and z for a better readability.

Theorem 3.4.7. *Let* $\Sigma \in \mathbb{R}^{n\times n}$ *be deterministic positive semi-definite matrices, and* τ *deterministic probability measures supported on* $\mathbb{R}^+$*.*
If $|g_\Sigma(z) - g_\tau(z)| \leq O_z(\epsilon_n)$*, then* $|g_{\nu^\Sigma}(z) - g_\chi(z)| \leq O_z(\epsilon_n)$*.*

The proof of this result may be decomposed in several steps. First we translate the definition of the measures ν and χ into appropriate self-consistent equations on their reciprocal Cauchy transforms (Proposition 3.4.8). Then we

show that χ is an approximate fixed point of the equation corresponding to ν (Proposition 3.4.10). Finally we use the stability of the self-consistent equation and the tools developed in [LC21] and [Cho22] to conclude.

The key function in the upcoming paragraphs $l_{\check{\nu}^\Sigma}(z) = -1/g_{\check{\nu}}(z)$ is known as the reciprocal Cauchy transform of the measure $\tilde{\nu}^\Sigma$. As such some classical properties of this function may be found in the seminal book [MS17, Section 3.4]. We will nonetheless provide short proofs for all the properties we need in this article.

Proposition 3.4.8 (Self-consistent equation for reciprocal Cauchy transforms)**.** *If μ is a measure supported on $\mathbb{R}^+$, we let $\nu = \mathrm{MP}(\gamma_n) \boxtimes \mu$, $\check{\nu} = (1-\gamma_n)\cdot\delta_0 + \gamma_n \cdot \nu$, and $l_{\check{\nu}}(z) = -1/g_{\check{\nu}}(z)$.*

Then $l_{\check{\nu}}(z)$ is the only solution on $\mathbb{C}^+$ of the self consistent equation in l :

$$l = z + \gamma_n l + \gamma_n l^2 g_\mu(l) = z + \gamma_n z \int_{\mathbb{R}} \frac{t}{\frac{zt}{l} - z} \mu(dt).$$

Proof. The two right hand side terms are always equal since :

$$\int_{\mathbb{R}} \frac{zt}{\frac{zt}{l} - z} \mu(dt) = l \int_{\mathbb{R}} \frac{t}{t-l} \mu(dt) = l \int_{\mathbb{R}} \left(1 + \frac{l}{t-l}\right) \mu(dt) = l + l^2 g_\mu(l).$$

Let us work with the first formulation. It is a classical property of Stieltjes transforms that $g_{\check{\nu}}(z) \in \mathbb{C}^+$ when $z \in \mathbb{C}^+$, hence $l_{\check{\nu}}(z) \in \mathbb{C}^+$. Since $g_{\check{\nu}} = \frac{\gamma_n - 1}{z} + \gamma_n g_\nu$, we have the identity $1 - \gamma_n - \gamma_n z g_\nu = \frac{z}{l_{\check{\nu}}}$. By definition of the free multiplicative convolution with a Marčenko-Pastur distribution, g_ν is the only solution on $\mathbb{C}^+$ of :

$$\begin{aligned} g_\nu &= \int_{\mathbb{R}} \frac{1}{(1-\gamma_n-\gamma_n z g_\nu)t - z} \mu(dt) \\ &= \int_{\mathbb{R}} \frac{1}{zt/l_{\check{\nu}} - z} \mu(dt) \\ &= \frac{l_{\check{\nu}}}{z} \int_{\mathbb{R}} \frac{1}{t - l_{\check{\nu}}} \mu(dt) \\ &= \frac{l_{\check{\nu}}}{z} g_\mu(l_{\check{\nu}}). \end{aligned}$$

Using again the identity $1 - \gamma_n - \gamma_n z g_\nu = \frac{z}{l_{\check{\nu}}}$, the self-consistent equation characterizing g_ν is equivalent for $l_{\check{\nu}}$ to satisfy :

$$\gamma_n l_{\check{\nu}} + \gamma_n l_{\check{\nu}}^2 g_\mu(l_{\check{\nu}}) = \gamma_n l_{\check{\nu}}(1 + z g_\nu) = l_{\check{\nu}} - z.$$

□

Lemma 3.4.9. *With the same notations and hypothesis as in Proposition 3.4.8 :*

$$\begin{aligned} 0 &\leq \Im\Big(z^{-1} l_{\check{\nu}}(z)\Big), \\ \Im(z) &\leq \Im(l_{\check{\nu}}(z)) \leq O(1+\Im(z)), \\ |l_{\check{\nu}}(z)| &\leq O\left(\frac{|z|}{\Im(z)}\right). \end{aligned}$$

In particular if a function $\zeta : \mathbb{N} \times \mathbb{C}^+ \to \mathbb{R}$ satisfies $\zeta(n,z) \leq O_z(\epsilon_n)$, then $\zeta(n, l_{\check{\nu}}(z)) \leq O_z(\epsilon_n)$.

Proof. $\nu = \mathrm{MP}(\gamma_n) \boxtimes \mu$ is supported on $\mathbb{R}^+$, thus :

$$\begin{aligned} -\Im\left(\frac{z}{l_{\check{\nu}}}\right) &= -\Im(1-\gamma_n - \gamma_n z g_\nu) \\ &= \gamma_n \int_{\mathbb{R}^+} \Im\left(\frac{z}{t-z}\right)\nu(dt) \\ &= \gamma_n \Im(z) \int_{\mathbb{R}^+} \frac{t}{|t-z|^2}\nu(dt) \geq 0, \end{aligned}$$

which proves that $\Im(z^{-1} l_{\check{\nu}}) \geq 0$. Secondly :

$$\begin{aligned} \Im(l_{\check{\nu}} - z) &= \Im\left(\gamma_n z \int_{\mathbb{R}} \frac{t}{\frac{zt}{l_{\check{\nu}}} - z}\mu(dt)\right) \\ &= \gamma_n \int_{\mathbb{R}^+} \Im\left(\frac{1}{1/l_{\check{\nu}} - 1/t}\right)\mu(dt) \\ &= \gamma_n \int_{\mathbb{R}^+} \frac{\Im(l_{\check{\nu}})t^2}{\Re(l_{\check{\nu}})^2 + \Im(l_{\check{\nu}})^2 + t^2 - 2\Re(l_{\check{\nu}})t}\mu(dt) \geq 0. \end{aligned}$$

The right hand side integral is bounded from above by $\frac{1}{\Im(l_{\check{\nu}})}\int_{\mathbb{R}^+} t^2\mu(dt)$, hence $\Im(l_{\check{\nu}})^2 \leq \Im(z)\Im(l_{\check{\nu}}) + \gamma_n \int_{\mathbb{R}^+} t^2\mu(dt)$. Solving this second order polynomial inequality gives $\Im(l_{\check{\nu}}) \in \left[\Im(z)/2 \pm (\gamma_n \int_{\mathbb{R}^+} t^2\mu(dt) + \Im(z)^2/4)^{1/2}\right]$ and $\Im(l_{\check{\nu}}) \leq O(1+\Im(z))$. We also have $\Im\left(\frac{zt}{l_{\check{\nu}}}\right) \leq 0$ for any $t \geq 0$, hence :

$$\begin{aligned} |l_{\check{\nu}}| &\leq |z| + \gamma_n |z| \int_{\mathbb{R}^+} \frac{t}{\left|\Im\left(\frac{zt}{l_{\check{\nu}}} - z\right)\right|}\mu(dt) \\ &\leq |z| + \gamma_n \frac{|z|}{\Im(z)} \int_{\mathbb{R}^+} t\,\mu(dt) \leq O\left(\frac{|z|}{\Im(z)}\right). \end{aligned}$$

For the last statement, let ζ be a function such that $\zeta(n,z) \leq O_z(\epsilon_n)$. If $\Im(z)$ is bounded, so is $\Im(l_{\check{\nu}}(z))$, hence there exists $\alpha > 0$ such that :

$$\zeta(n, l_{\check{\nu}}(z)) \leq O\left(\epsilon_n \frac{|l_{\check{\nu}}(z)|^\alpha}{\Im(l_{\check{\nu}}(z))^{2\alpha}}\right) \leq O\left(\epsilon_n \frac{|z|^\alpha}{\Im(z)^{3\alpha}}\right) \leq O\left(\epsilon_n \frac{|z|^{2\alpha}}{\Im(z)^{4\alpha}}\right),$$

from which we deduce that $\zeta(n, l_{\check{\nu}}(z)) \leq O_z(\epsilon_n)$. □

Let us now move on to the second part of the proof of Theorem 3.4.7. We define the mapping :

$$\mathcal{F} : l \in \mathbb{C}^+ \mapsto z + \gamma_n \frac{z}{n} \mathrm{Tr}\left(\left(\frac{z}{l}\Sigma - zI_n \right)^{-1} \Sigma \right).$$

As shown in Proposition 3.4.8, the definition of $\mathcal{F}$ is equivalent to :

$$\mathcal{F}(l) = z + \gamma_n z \int_{\mathbb{R}} \frac{t}{\frac{zt}{l} - z} \mu_\Sigma(dt) = z + \gamma_n l + \gamma_n l^2 g_\Sigma(l).$$

We set $l_{\check{\nu}^\Sigma}(z) = -1/g_{\check{\nu}^\Sigma}(z)$ and $l_{\check{\chi}}(z) = -1/g_{\check{\chi}}(z)$. In Proposition 3.4.8 we have proved that $l_{\check{\nu}^\Sigma}(z)$ is a fixed point of $\mathcal{F}$. We will see in the next proposition that $l_{\check{\chi}}(z)$ is almost a fixed point of $\mathcal{F}$.

Proposition 3.4.10. $|\mathcal{F}(l_{\check{\chi}}(z)) - l_{\check{\chi}}(z)| \leq O_z(\epsilon_n)$.

Proof. Using the first formulation of the self-consistent equation of Proposition 3.4.8, we have $l_{\check{\chi}} = z + \gamma_n l_{\check{\chi}} + \gamma_n l_{\check{\chi}}^2 g_\nu(l_{\check{\chi}})$, thus :

$$\mathcal{F}(l_{\check{\chi}}) - l_{\check{\chi}} = \gamma_n l_{\check{\chi}}^2 (g_{\mu_\Sigma}(l_{\check{\chi}}) - g_\tau(l_{\check{\chi}})).$$

As seen in Lemma 3.4.9, since $|g_{\mu_\Sigma}(z) - g_\tau(z)| \leq O_z(\epsilon_n)$, we also have $|g_{\mu_\Sigma}(l_{\check{\chi}}(z)) - g_\tau(l_{\check{\chi}}(z))| \leq O_z(\epsilon_n)$, and we obtain :

$$\begin{aligned} |\mathcal{F}(l_{\check{\chi}}) - l_{\check{\chi}}| &\leq |\gamma_n| |l_{\check{\chi}}|^2 |g_{\mu_\Sigma}(l_{\check{\chi}}) - g_\tau(l_{\check{\chi}})| \\ &\leq O(1) O\left(\frac{|z|}{\Im(z)} \right)^2 O_z(\epsilon_n) \leq O_z(\epsilon_n). \end{aligned}$$

□

The last step to prove Theorem 3.4.7 is to use the stability of the self-consistent equation $l = \mathcal{F}(l)$. Let us recall the tools and results established in [Cho22, Section 6]. For a fixed $z \in \mathbb{C}^+$, we introduce the domain $\mathbf{D} = \{\omega \in \mathbb{C}$ such that $\Im(\omega) \geq \Im(z)$ and $\Im(z^{-1}\omega) \geq 0\}$, and the semi-metric on $\mathbb{C}^+$:

$$d(\omega_1, \omega_2) = \frac{|\omega_1 - \omega_2|}{\Im(\omega_1)^{1/2} \Im(\omega_2)^{1/2}}.$$

$\mathcal{F}$ is a contraction mapping on $\mathbf{D}$ with respect to d. More precisely, $\mathcal{F}$ is $k_{\mathcal{F}}$-Lipschitz with $k_{\mathcal{F}} = \frac{\frac{|z|}{\Im(z)^2}}{1 + \frac{|z|}{\Im(z)^2}}$ ([Cho22, Proposition 6.11]). Moreover, if $c \in \mathbf{D}$ is a fixed point of $\mathcal{F}$ and $b \in \mathbf{D}$ any other point, provided $k_{\mathcal{F}}(1 + d(b, \mathcal{F}(b))) < 1$, the following inequality holds true ([Cho22, Lemma 6.14]) :

$$|c - b| \leq \frac{|\mathcal{F}(b) - b|}{1 - k_{\mathcal{F}}(1 + d(b, \mathcal{F}(b)))}.$$

We have now collected all the arguments required to compare the Stieltjes transforms of ν^Σ and χ.

Proof of Theorem 3.4.7. If $\Im(z)$ is bounded, using Proposition 3.4.10 there exists α and $C > 0$ such that $|\mathcal{F}(l_{\check{\chi}}) - l_{\check{\chi}}| \leq C\epsilon_n \frac{|z|^\alpha}{\Im(z)^{2\alpha}}$. For values of z such that $\epsilon_n \frac{|z|^{\alpha+1}}{\Im(z)^{2\alpha+3}} \leq 1/2C$, $\frac{|z|}{\Im(z)^2} d(\mathcal{F}(l_{\check{\chi}}), l_{\check{\chi}}) \leq 1/2$, thus :

$$\begin{aligned} k_{\mathcal{F}}(1 + d(\mathcal{F}(l_{\check{\chi}}), l_{\check{\chi}})) &= \frac{\frac{|z|}{\Im(z)^2} + \frac{|z|}{\Im(z)^2} d(\mathcal{F}(l_{\check{\chi}}), l_{\check{\chi}})}{1 + \frac{|z|}{\Im(z)^2}} \\ &\leq 1 - \frac{1}{2\Big(1 + \frac{|z|}{\Im(z)^2}\Big)}. \end{aligned}$$

In particular $k_{\mathcal{F}}(1 + d(\mathcal{F}(l_{\check{\chi}}), l_{\check{\chi}})) < 1$, and from [Cho22, Lemma 6.14] :

$$\begin{aligned} |l_{\check{\chi}} - l_{\check{\nu^\Sigma}}| &\leq \frac{|\mathcal{F}(l_{\check{\chi}}) - l_{\check{\chi}}|}{1 - k_{\mathcal{F}}(1 + d(\mathcal{F}(l_{\check{\chi}}), l_{\check{\chi}}))} \\ &\leq 2\left(1 + \frac{|z|}{\Im(z)^2}\right) C\epsilon_n \frac{|z|^\alpha}{\Im(z)^{2\alpha}} \\ &\leq O\left(\epsilon_n \frac{|z|^{\alpha+1}}{\Im(z)^{2\alpha+2}}\right). \end{aligned}$$

To conclude :

$$|g_{\nu^\Sigma} - g_\chi| = \frac{|g_{\check{\nu^\Sigma}} - g_{\check{\chi}}|}{\gamma} = \frac{|g_{\check{\nu^\Sigma}}||g_{\check{\chi}}||l_{\check{\nu^\Sigma}} - l_{\check{\chi}}|}{\gamma_n} \leq O\left(\epsilon_n \frac{|z|^{\alpha+1}}{\Im(z)^{2\alpha+2}}\right),$$

for values of z such that $\epsilon_n \frac{|z|^{\alpha+1}}{\Im(z)^{2\alpha+3}} \leq \frac{1}{2C}$. Given the *a priori* bound $|g_{\nu^\Sigma}(z) - g_\chi(z)| \leq \frac{2}{\Im(z)}$ and Lemma 3.2.6, the bound $|g_{\nu^\Sigma}(z) - g_\chi(z)| \leq O_z(\epsilon_n)$ holds true. □

3.4.3 Approximation of the deterministic equivalent built from deterministic matrices

Given an approximation for the Stieltjes transform $g_\Sigma(z) \approx g_\tau(z)$, and another approximation for the resolvent $\mathcal{G}_\Sigma(z) \approx \mathbf{H}(z)$, can we find an approximation for the matrix $\mathbf{G}^\Sigma_{\boxtimes}(z)$?

To this end we build a matrix function $\mathbf{K}$ using the same procedure we mentioned earlier to build $\mathbf{G}^\Sigma_{\boxtimes}$ from μ_Σ and $\mathcal{G}_\Sigma$:

$$\begin{aligned} \chi &= \mathrm{MP}(\gamma_n) \boxtimes \tau, \\ \check{\chi} &= (1 - \gamma_n) \cdot \delta_0 + \gamma_n \cdot \chi, \\ l_{\check{\chi}}(z) &= -1/g_{\check{\chi}}(z), \\ \mathbf{K}(z) &= z^{-1} l_{\check{\chi}}(z) \mathbf{H}(l_{\check{\chi}}(z)). \end{aligned}$$

Proposition 3.4.11. *Let $\Sigma \in \mathbb{R}^{n\times n}$ be deterministic positive semi-definite matrices, τ deterministic probability measures supported on $\mathbb{R}^+$, and $\mathbf{H} : \mathbb{C}^+ \to \mathbb{C}^{n\times n}$ deterministic complex functions.*

If $|g_\Sigma(z) - g_\tau(z)| \leq O_z(\epsilon_n)$ and $\|\|\mathcal{G}_\Sigma(z) - \mathbf{H}(z)\|\| \leq O_z(\epsilon'_n)$, then $\left\|\left\|\mathbf{G}^\Sigma_\boxtimes(z) - \mathbf{K}(z)\right\|\right\| \leq O_z(\epsilon_n + \epsilon'_n)$.

Proof. We use a triangular inequality in the following decomposition :

$$\begin{aligned}\mathbf{G}^\Sigma_\boxtimes(z) - \mathbf{K}(z) = z^{-1}(l_{\check\nu} - l_{\check\chi})\mathcal{G}_\Sigma(l_{\check\nu}) \\ + z^{-1}l_{\check\chi}(\mathcal{G}_\Sigma(l_{\check\nu}) - \mathcal{G}_\Sigma(l_{\check\chi})) \\ + z^{-1}l_{\check\chi}(\mathcal{G}_\Sigma(l_{\check\chi}) - \mathbf{H}(l_{\check\chi})).\end{aligned}$$

For the first term, $|l_{\check\nu} - l_{\check\chi}| \leq O_z(\epsilon_n)$ and $\|\|\mathcal{G}_\Sigma(l_{\check\nu})\|\| \leq \frac{1}{\Im(l_{\check\nu})} \leq \frac{1}{\Im(z)}$, thus $\|\|z^{-1}(l_{\check\nu} - l_{\check\chi})\mathcal{G}_\Sigma(l_{\check\nu})\|\| \leq O_z(\epsilon_n)$. For the second term, using a resolvent identity :

$$\begin{aligned}\left\|\left\|z^{-1}l_{\check\chi}(\mathcal{G}_\Sigma(l_{\check\nu}) - \mathcal{G}_\Sigma(l_{\check\chi}))\right\|\right\| &\leq \left|z^{-1}l_{\check\chi}(z)\right|\|\|\mathcal{G}_\Sigma(l_{\check\nu})\|\|l_{\check\nu} - l_{\check\chi}|\|\|\mathcal{G}_\Sigma(l_{\check\chi})\|\| \\ &\leq \frac{1}{|z|}O\left(\frac{|z|}{\Im(z)}\right)\frac{O_z(\epsilon_n)}{\Im(z)^2} \leq O_z(\epsilon_n).\end{aligned}$$

Finally for the third term, $\|\|\mathcal{G}_\Sigma(l_{\check\chi}) - \mathbf{H}(l_{\check\chi})\|\| \leq O_z(\epsilon'_n)$ using Lemma 3.4.9, thus :

$$\left\|\left\|z^{-1}l_{\check\chi}(\mathcal{G}_\Sigma(l_{\check\chi}) - \mathbf{H}(l_{\check\chi}))\right\|\right\| \leq \frac{1}{|z|}O\left(\frac{|z|}{\Im(z)}\right)O_z(\epsilon'_n) \leq O_z(\epsilon'_n).$$

□

Corollary 3.4.12. *The map $\Sigma \mapsto \mathbf{G}^\Sigma_\boxtimes(z)$ is $O_z(1)$ Lipschitz with respect to the spectral norm.*

Proof. Using a resolvent identity we have :

$$\|\|\mathcal{G}_\Sigma(z) - \mathcal{G}_{\Sigma'}(z)\|\| \leq \|\|\mathcal{G}_\Sigma(z)\|\|\|\Sigma' - \Sigma\|\|\|\mathcal{G}_\Sigma(z)\|\| \leq \frac{\|\|\Sigma - \Sigma'\|\|}{\Im(z)^2}.$$

In particular $\left|g_\Sigma(z) - g_{\Sigma'(z)}\right| \leq \|\|\mathcal{G}_\Sigma(z) - \mathcal{G}_{\Sigma'}(z)\|\| \leq O_z(\|\|\Sigma - \Sigma'\|\|)$. The result follows from Theorem 3.4.7 and Proposition 3.4.11. □

3.4.4 Concentration of the deterministic equivalent built from random matrices

If Σ is random and satisfies a typical $\mathcal{E}(1/\sqrt{n})$ Lipschitz concentration property, from the approximations $\mathbb{E}[g_\Sigma(z)] \approx g_\tau(z)$ and $\mathbb{E}[\mathcal{G}_\Sigma(z)] \approx \mathbf{H}(z)$, we may deduce that $g_\Sigma(z) \approx g_\tau(z)$ a.s., but we cannot expect that $\mathcal{G}_\Sigma(z) \approx \mathbf{H}(z)$ since $\mathcal{G}_\Sigma(z)$ and $\mathbb{E}[\mathcal{G}_\Sigma(z)]$ are not necessarily close in spectral norm. We can however prove that $\mathbf{G}^\Sigma_\boxtimes(z)$ is linearly concentrated around $\mathbf{K}(z)$.

Proposition 3.4.13. *Let $\Sigma \in \mathbb{R}^{n\times n}$ be random positive semi-definite matrices such that $\Sigma \propto_{\|\cdot\|_F} \mathcal{E}(1/\sqrt{n})$. Then :*

1. *$g_\nu(z) \propto \mathcal{E}(O_z(1/n))$ and $\mathbf{G}^\Sigma_\boxtimes(z) \propto_{\|\cdot\|} \mathcal{E}(O_z(1/\sqrt{n}))$.*
2. *If τ are deterministic probability measures supported on $\mathbb{R}^+$ such that $|\mathbb{E}[g_\Sigma(z)] - g_\tau(z)| \leq O_z(\epsilon_n)$, then $|g_{\nu^\Sigma}(z) - g_\chi(z)| \leq O_z(\epsilon_n + \sqrt{\log n}/n)$ almost surely eventually (a.s.e.).*
3. *If in addition $\mathbf{H} : \mathbb{C}^+ \to \mathbb{C}^{n\times n}$ are deterministic complex functions such that $\|\|\mathbb{E}[\mathcal{G}_\Sigma(z)] - \mathbf{H}(z)\|\| \leq O_z(\epsilon'_n)$, then $\mathbf{G}^\Sigma_\boxtimes(z) \in_{\|\cdot\|} \mathbf{K}(z) \pm \mathcal{E}(O_z(\epsilon_n + \epsilon'_n + 1/\sqrt{n}))$.*

Proof. We refer to Proposition 3.2.2 for the properties of Lipschitz and linear concentrations used in this proof. The map $\Sigma \mapsto \mathbf{G}^\Sigma_\boxtimes$ is $O_z(1)$ Lipschitz with respect to the spectral norm, thus $\mathbf{G}^\Sigma_\boxtimes \propto_{\|\cdot\|} \mathcal{E}(O_z(1/\sqrt{n}))$. Remembering the identity $\mathrm{Tr}\mathbf{G}^\Sigma_\boxtimes = n g_{\nu^\Sigma}$, we also deduce that $g_{\nu^\Sigma} \propto \mathcal{E}(O_z(1/\sqrt{n}))$.

The map $\Sigma \mapsto \mathcal{G}_\Sigma$ is $1/\Im(z)^2$ Lipschitz, thus $\mathcal{G}_\Sigma \propto_{\|\cdot\|_F} \mathcal{E}(O_z(1/\sqrt{n}))$ and $g_\Sigma \propto \mathcal{E}(O_z(1/n))$. We deduce that $|\mathbb{E}[g_\Sigma] - g_\Sigma| \leq O_z(\sqrt{\log n}/n)$ a.s.e., hence $|g_\Sigma - g_\tau| \leq O_z(\epsilon_n + \sqrt{\log n}/n)$ a.s.e. We can apply Theorem 3.4.7 uniformly on this set of full measure, and we obtain that $|g_{\nu^\Sigma} - g_\chi| \leq O_z(\epsilon_n + \sqrt{\log n}/n)$ a.s.e.

In the decomposition :

$$\begin{aligned}\mathbf{G}^\Sigma_\boxtimes(z) - \mathbf{K}(z) = {} & z^{-1}(l_{\check{\nu}^\Sigma} - l_{\check{\chi}})\mathcal{G}_\Sigma(l_{\check{\nu}^\Sigma}) \\ & + z^{-1} l_{\check{\chi}}(\mathcal{G}_\Sigma(l_{\check{\nu}^\Sigma}) - \mathcal{G}_\Sigma(l_{\check{\chi}})) \\ & + z^{-1} l_{\check{\chi}}(\mathcal{G}_\Sigma(l_{\check{\chi}}) - \mathbf{H}(l_{\check{\chi}})),\end{aligned}$$

the first two terms are bounded by $O_z(\epsilon_n + \sqrt{\log n}/n) \leq O_z(\epsilon_n + 1/\sqrt{n})$ a.s.e. in spectral norm (see the proof of Proposition 3.4.11). For the third term, $\mathcal{G}_\Sigma(z) \propto_{\|\cdot\|_F} \mathcal{E}(O_z(1/\sqrt{n}))$ and $\|\|\mathbb{E}[\mathcal{G}_\Sigma(z)] - \mathbf{H}(z)\|\| \leq O_z(\epsilon'_n)$, thus $\mathcal{G}_\Sigma(z) \in_{\|\cdot\|} \mathbf{H}(z) \pm \mathcal{E}(O_z(\epsilon'_n + 1/\sqrt{n}))$. From Lemma 3.4.9, we also have $\mathcal{G}_\Sigma(l_{\check{\chi}}) \in_{\|\cdot\|} \mathbf{H}(l_{\check{\chi}}) \pm \mathcal{E}(O_z(\epsilon'_n + 1/\sqrt{n}))$. Finally $|z^{-1} l_{\check{\chi}}| \leq 1/\Im(z)$, and combining the above estimates and concentration properties leads to $\mathbf{G}^\Sigma_\boxtimes(z) \in_{\|\cdot\|} \mathbf{K}(z) \pm \mathcal{E}(O_z(\epsilon_n + \epsilon'_n + 1/\sqrt{n}))$. □

3.5 Single-layer neural network with deterministic data

3.5.1 Setting

In this section we consider the conjugate kernel matrix associated to a single-layer artificial neural network with deterministic input. The model is made of :

— a random weight matrix $W \in \mathbb{R}^{d\times d_0}$,

— a deterministic data matrix $X \in \mathbb{R}^{d_0\times n}$,

— two random biases matrices B and $D \in \mathbb{R}^{d\times n}$,

— an activation function $f : \mathbb{R} \to \mathbb{R}$,

— and real parameters $\sigma_W^2, \sigma_X^2 > 0$ and $\sigma_B^2, \sigma_D^2 \geq 0$ that will appear later in Assumptions 3.5.2.

As output of the neuron, we set $Y = f(WX/\sqrt{d_0} + B) + D \in \mathbb{R}^{d\times n}$, where the function f is applied entry-wise. Our goal is to investigate the spectral properties of the conjugate kernel matrix $K = Y^\top Y/d$. For $z \in \mathbb{C}^+$ we define its resolvent $\mathcal{G}_K(z) = (K - zI_n)^{-1}$ and Stieltjes transform $g_K(z) = (1/n)\mathrm{Tr}\mathcal{G}_K(z)$. We also define the following objects :

$$\begin{aligned}
\tilde{\sigma}^2 &= \sigma_W^2\sigma_X^2 + \sigma_B^2,\\
\tilde{f}(t) &= f(\tilde{\sigma}t),\\
\mathfrak{a} &= \left\|\tilde{f}\right\|_{\mathcal{H}}^2 - \frac{\sigma_W^2\sigma_X^2}{\tilde{\sigma}^2}\zeta_1(\tilde{f})^2 + \sigma_D^2,\\
\mathfrak{b} &= \zeta_1(\tilde{f})^2\frac{\sigma_W^2}{\tilde{\sigma}^2},\\
K_X &= X^\top X/d_0,\\
\Delta_X &= K_X - \sigma_X^2 I_n,\\
\Sigma &= \mathbb{E}[K],\\
\Sigma_{\mathrm{lin}} &= \mathfrak{a}I_n + \mathfrak{b}K_X.
\end{aligned}$$

Remark 3.5.1. We always have $\mathfrak{a} \geq 0$ since $\zeta_1(f)^2 \leq \left\|\tilde{f}\right\|_{\mathcal{H}}^2$ and $\sigma_W^2\sigma_X^2 \leq \tilde{\sigma}^2$. Moreover $\mathfrak{a} = 0$ if and only if f is a linear function and there is no bias in the model (that is $\sigma_B^2 = \sigma_D^2 = 0$).

Like in the rest of this article, we sometimes omit the indices n and z for a better readability, even if we are implicitly dealing with sequences of matrices, measures, and complex functions.

Assumptions 3.5.2. 1. W, B and D are random, independent, with i.i.d. $\mathcal{N}(\sigma_W^2)$, $\mathcal{N}(\sigma_B^2)$ and $\mathcal{N}(\sigma_D^2)$ entries respectively.

2. $\tilde{f}$ is Lipschitz continuous and Gaussian centered, that is $\mathbb{E}\left[\tilde{f}(\mathcal{N})\right] = \mathbb{E}[f(\tilde{\sigma}\mathcal{N})] = 0$.
3. X is deterministic and $|||K_X|||$ is bounded.
4. $\left\|\vec{\text{diag}}(\Delta_X)\right\|$ is bounded and $\|\Delta_X\|_{\max}$ converges to 0.
5. The ratio $\gamma_n = \frac{n}{d}$ is bounded from above and away from 0.

We refer to the Remark 3.5.8 for a detailed discussion about the assumption (4). The main result of this section is Theorem 3.5.7 that gives a deterministic equivalent for $\mathcal{G}_K(z)$ and $g_K(z)$. To prove this result, we will combine the general results on resolvent matrices recalled in Section 3.4, with the linearization techniques of Section 3.3. We will wrap up this section by applying our framework to a simple yet original model having weakly correlated entries.

3.5.2 Technicalities and linearization of Σ

Proposition 3.5.3. *Under Assumptions 3.5.2, $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$. The rows of Y are i.i.d. sampled from the distribution of a random vector $y = f(X^\top w/\sqrt{d_0} + b) + \tilde{b}$, where $w \in \mathbb{R}^{d_0}$ and $b, \tilde{b} \in \mathbb{R}^n$ are independent Gaussian vectors with i.i.d. $\mathcal{N}(\sigma_W^2)$, $\mathcal{N}(\sigma_B^2)$ and $\mathcal{N}(\sigma_D^2)$ coordinates respectively. $\|\mathbb{E}[y]\|$ is moreover bounded.*

Proof. The map $(W, B, D) \mapsto f(WX/\sqrt{d_0} + B) + D$ is Lipschitz with respect to the product Frobenius norm since f is Lipschitz and $\left|\left|\left|X/\sqrt{d_0}\right|\right|\right|$ is bounded. Given the Gaussian concentration $(W, B, D) \propto_{\|\cdot\|_F} \mathcal{E}(1)$, we immediately obtain that $Y \propto_{\|\cdot\|_F} \mathcal{E}(1)$. From the expression :

$$Y_{ij} = f\left(\sum_{k=1}^{d_0} W_{ik}X_{kj}/\sqrt{d_0} + B_{ij}\right) + D_{ij},$$

we see that the rows of Y are independent and have the same distribution as $y = f\left(X^\top w/\sqrt{d_0} + b\right) + \tilde{b}$. For the last statement, we may write $y = \tilde{f}(u) + \tilde{b}$, where u is a centered Gaussian vector with covariance matrix $S = (\sigma_W^2 K_X + \sigma_B^2 I_n)/\tilde{\sigma}^2$. We have $S - I_n = (\sigma_W^2/\tilde{\sigma}^2)\Delta_X$, hence the random variables u_i are centered Gaussian with covariance $1 + O((\Delta_X)_{ii})$. Since $\zeta_0(\tilde{f}) = \mathbb{E}[f(\tilde{\sigma}\mathcal{N})] = 0$, using the first order expansion given by Corollary 3.3.4 applied to the function $\tilde{f}$, we have $\mathbb{E}\left[\tilde{f}(u_i)\right] = O((\Delta_X)_{ii})$ uniformly on $i \in [\![1, n]\!]$. We deduce that $\|\mathbb{E}[y]\| = \left\|\mathbb{E}\left[\tilde{f}(u)\right]\right\| = O\left(\left\|\vec{\text{diag}}(\Delta_X)\right\|\right) = O(1)$. □

Corollary 3.5.4. *Under Assumptions 3.5.2, $|||\Sigma|||$ is bounded, and moreover :*

$$|||\Sigma - \Sigma_{\text{lin}}||| \leq O(\|\Delta_X\|_{\max} + n\zeta_2(\tilde{f})^2\|\Delta_X\|_{\max}^2 + n\zeta_3(\tilde{f})^2\|\Delta_X\|_{\max}^3),$$
$$|g_\Sigma(z) - g_{\Sigma_{\text{lin}}}(z)| \leq O(\|\Delta_X\|_{\max} + \sqrt{n}\zeta_2(\tilde{f})^2\|\Delta_X\|_{\max}^2 + \sqrt{n}\zeta_3(\tilde{f})^2\|\Delta_X\|_{\max}^3).$$

Proof. As seen in the last proposition, the rows of Y are i.i.d. sampled from the distribution of a vector $y = \tilde{f}(u) + \tilde{b}$ where u is a centered Gaussian vector with covariance matrix $S = I_n + \Delta$, and $\Delta = (\sigma_W^2/\tilde{\sigma}^2)\Delta_X$. The Assumptions 3.3.8 are satisfied, and we deduce from Proposition 3.3.10 that $|||\Sigma - \Sigma_{\text{lin}}|||$ is bounded. Since $|||K_X|||$ is bounded and $\Sigma_{\text{lin}} = \mathfrak{a} I_n + \mathfrak{b} K_X$, $|||\Sigma|||$ is also bounded. From the same proposition, we have the following estimates in spectral norm, with $\epsilon_n = \|\Delta_X\|_{\max} + n\zeta_2(\tilde{f})^2\|\Delta_X\|_{\max}^2 + n\zeta_3(\tilde{f})^2\|\Delta_X\|_{\max}^3$:

$$\begin{aligned}
\mathbb{E}\big[\tilde{f}(u)\tilde{f}(u)^\top\big] &= \big\|\tilde{f}\big\|^2 I_n + \zeta_1(\tilde{f})^2\Delta + O(\epsilon_n) \\
&= \big\|\tilde{f}\big\|^2 I_n + \zeta_1(\tilde{f})^2(\sigma_W^2/\tilde{\sigma}^2)\Big(K_X - \sigma_X^2 I_n\Big) + O(\epsilon_n) \\
\Sigma = \mathbb{E}\big[yy^\top\big] &= \mathbb{E}\big[\tilde{f}(u)\tilde{f}(u)^\top\big] + \sigma_D^2 I_n \\
&= \left(\big\|\tilde{f}\big\|^2 - \frac{\sigma_W^2\sigma_X^2}{\tilde{\sigma}^2}\zeta_1(\tilde{f})^2 + \sigma_D^2\right) I_n + \zeta_1(\tilde{f})^2\frac{\sigma_W^2}{\tilde{\sigma}^2}K_X + O(\epsilon_n) \\
&= \Sigma_{\text{lin}} + O(\epsilon_n).
\end{aligned}$$

The proof for the Stieltjes transforms is similar. □

3.5.3 Propagation of the approximate orthogonality

The contents of this paragraph will not be useful right away in this section, but rather in later stages of the article to study multi-layer networks by induction. Similar results may be found in [FW20, Section D] under the name of propagation of approximate orthogonality (see Remark 3.5.8 for more details on this concept), and proved by slightly different means.

Lemma 3.5.5. *Let $\sigma_Y^2 = \big\|\tilde{f}\big\|_{\mathcal{H}}^2 + \sigma_D^2$, and $\Delta_Y = K - \sigma_Y^2 I_n = Y^\top Y/d - \sigma_Y^2 I_n$. Under Assumptions 3.5.2, there exists an event $\mathcal{B}$ with $\mathbb{P}(\mathcal{B}^c) \leq ce^{-n/c}$ for some constant $c > 0$, such that uniformly on $\mathcal{B}$, $\big\|\vec{\text{diag}}(\Delta_Y)\big\|$ and $|||K|||$ are bounded, and $\|\Delta_Y\|_{\max} \leq O\Big(\|\Delta_X\|_{\max} + \sqrt{\log n/n}\Big)$.*

Proof. We have $|||\mathbb{E}[Y]|||^2 \leq \|\mathbb{E}[Y]\|_F^2 \leq d\|\mathbb{E}[y]\|^2 \leq O(n)$. From Proposition 3.2.3, there exists a constant $c > 0$ and an event $\mathcal{B}$ with $\mathbb{P}(\mathcal{B}^c) \leq ce^{-n/c}$, such that $(K|\mathcal{B}) \propto_{\|\cdot\|_F} \mathcal{E}(1/\sqrt{n})$, and $|||Y||| \leq 2c\sqrt{n}$ on $\mathcal{B}$, thus $|||K|||$ is bounded on $\mathcal{B}$. We will check the other statements in expectation first, and in a second stage use concentration to obtain bounds on the random objects.

For any $i \in [\![1, n]\!]$, Y_{ii} has the same distribution as $\tilde{f}(u_i) + \tilde{b}_i$, where u_i and $\tilde{b}_i$ are centered Gaussian variables, independent, with variances $1 + O((\Delta_X)_{ii})$ and σ_D^2 respectively. Using the first order expansion given by Corollary 3.3.4

applied to the function $\tilde{f}^2$, uniformly on $i \in [\![1, n]\!]$ we have :

$$\begin{aligned}\mathbb{E}\Big[Y_{ii}^2\Big] &= \mathbb{E}\Big[\Big(\tilde{f}(u_i) + b_i\Big)^2\Big] \\ &= \zeta_0(\tilde{f}^2) + O((\Delta_X)_{ii}) + \sigma_D^2 \\ &= \Big\|\tilde{f}\Big\|_{\mathcal{H}}^2 + \sigma_D^2 + O((\Delta_X)_{ii}) \\ &= \sigma_Y^2 + O((\Delta_X)_{ii}).\end{aligned}$$

We deduce that $\mathbb{E}\Big[\Big\|\vec{\mathrm{diag}}(\Delta_Y)\Big\|^2\Big] \leq \sum_{i=1}^n O((\Delta_X)_{ii}^2) \leq O\Big(\Big\|\vec{\mathrm{diag}}(\Delta_X)\Big\|^2\Big) \leq O(1)$. Since $\Big(\vec{\mathrm{diag}}(\Delta_Y)|\mathcal{B}\Big) \propto_{\|\cdot\|} \mathcal{E}(1/\sqrt{n})$, $\Big\|\vec{\mathrm{diag}}(\Delta_Y) - \mathbb{E}\Big[\vec{\mathrm{diag}}(\Delta_Y)\Big]\Big\| \leq O(1)$ a.s. on $\mathcal{B}$, and $\Big\|\vec{\mathrm{diag}}(\Delta_Y)\Big\| \leq O(1)$ a.s. on $\mathcal{B}$.

Finally $\|K - \Sigma\|_{\max} \leq O\Big(\sqrt{\log n/n}\Big)$ a.s. on $\mathcal{B}$ using the general properties of concentration listed in Proposition 3.2.2, and from Proposition 3.3.10 :

$$\Big\|\Sigma - \sigma_Y^2 I_n\Big\|_{\max} \leq \|\Sigma - \Sigma_{\mathrm{lin}}\|_{\max} + O(\|\Delta_X\|_{\max}) \leq O(\|\Delta_X\|_{\max}),$$

which proves the result after a final triangular inequality. □

3.5.4 Deterministic equivalent and consequences

We let $\nu^{\Sigma_{\mathrm{lin}}} = \mathrm{MP}(\gamma_n) \boxtimes \mu_{\Sigma_{\mathrm{lin}}}$, and we refer to the beginning of Section 3.4 for the definition of the deterministic equivalent matrix function $\mathbf{G}_{\boxtimes}^{\Sigma_{\mathrm{lin}}}(z)$.

Remark 3.5.6. If $\mathfrak{b} = 0$, then Σ_{lin} does not depend on X and the objects $\nu^{\Sigma_{\mathrm{lin}}}$ and $\mathbf{G}_{\boxtimes}^{\Sigma_{\mathrm{lin}}}(z)$ are fully explicit. Indeed $\Sigma_{\mathrm{lin}} = \mathfrak{a} I_n$, $\mu_{\Sigma_{\mathrm{lin}}} = \delta_{\mathfrak{a}}$, and $\nu^{\Sigma_{\mathrm{lin}}} = \mathrm{MP}(\gamma_n) \boxtimes \delta_{\mathfrak{a}} = \mathfrak{a}\mathrm{MP}(\gamma_n)$. Since Σ_{lin} is a multiple of the identity matrix, so are $\mathcal{G}_{\Sigma_{\mathrm{lin}}}(z)$ and $\mathbf{G}_{\boxtimes}^{\Sigma_{\mathrm{lin}}}$, hence :

$$\mathbf{G}_{\boxtimes}^{\Sigma_{\mathrm{lin}}}(z) = \Big(\frac{1}{n}\mathrm{Tr}\mathbf{G}_{\boxtimes}^{\Sigma_{\mathrm{lin}}}(z)\Big) I_n = g_{\mathfrak{a}\mathrm{MP}(\gamma_n)}(z) I_n.$$

Note that if $f \neq 0$, then $\mathfrak{a} \neq 0$ and $g_{\mathfrak{a}\mathrm{MP}(\gamma_n)}(z) = \mathfrak{a} g_{\mathrm{MP}(\gamma_n)}(z/\mathfrak{a})$.

In the case where $\mathfrak{b} \neq 0$, we can describe the deterministic equivalents as functions of X. Indeed $\mu_{\Sigma_{\mathrm{lin}}} = \mathfrak{a} + \mathfrak{b}\mu_{K_X}$, $\nu^{\Sigma_{\mathrm{lin}}} = \mathrm{MP}(\gamma_n) \boxtimes (\mathfrak{a} + \mathfrak{b}\mu_{K_X})$, and since $\mathcal{G}_{\Sigma_{\mathrm{lin}}}(z) = \frac{1}{\mathfrak{b}}\mathcal{G}_{K_X}\Big(\frac{z-\mathfrak{a}}{\mathfrak{b}}\Big)$:

$$\begin{aligned}\mathbf{G}_{\boxtimes}^{\Sigma_{\mathrm{lin}}}(z) &= z^{-1} l_{\check{\nu}^{\Sigma_{\mathrm{lin}}}}(z) \mathcal{G}_L(l_{\check{\nu}^{\Sigma_{\mathrm{lin}}}}(z)) \\ &= \frac{l_{\check{\nu}^{\Sigma_{\mathrm{lin}}}}(z)}{\mathfrak{b} z} \mathcal{G}_{K_X}\left(\frac{l_{\check{\nu}^L}(z) - \mathfrak{a}}{\mathfrak{b}}\right),\end{aligned}$$

where $\check{\nu}^{\Sigma_{\mathrm{lin}}} = (1 - \gamma_n) \cdot \delta_0 + \gamma_n \cdot \nu^{\Sigma_{\mathrm{lin}}}$ and $l_{\check{\nu}^{\Sigma_{\mathrm{lin}}}}(z) = -1/g_{\check{\nu}^{\Sigma_{\mathrm{lin}}}}(z)$.

For the next results, let us denote :

$$\epsilon_n = \frac{1}{n} + \|\Delta_X\|_{\max} + \sqrt{n}\zeta_2(\tilde{f})^2\|\Delta_X\|_{\max}^2 + \sqrt{n}\zeta_3(\tilde{f})^2\|\Delta_X\|_{\max}^3,$$
$$\epsilon'_n = \frac{1}{\sqrt{n}} + \|\Delta_X\|_{\max} + n\zeta_2(\tilde{f})^2\|\Delta_X\|_{\max}^2 + n\zeta_3(\tilde{f})^2\|\Delta_X\|_{\max}^3.$$

Theorem 3.5.7. *Uniformly under Assumptions 3.5.2, the following concentration properties hold true :*

1. $g_K(z) \propto \mathcal{E}(O_z(1/n))$ *and* $\mathcal{G}_K(z) \propto_{\|\cdot\|_F} \mathcal{E}(O_z(1/\sqrt{n}))$.
2. $g_K(z) \in g_{\nu^{\Sigma_{\text{lin}}}}(z) \pm \mathcal{E}(O_z(\epsilon_n))$ *and* $\mathcal{G}_K(z) \in_{\|\cdot\|} \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z) \pm \mathcal{E}(O_z(\epsilon'_n))$.

Remark 3.5.8. Assumption 3.5.2(4) states roughly that the data matrix is close to being an orthogonal matrix (up to rescaling), a notion that was first introduced in [FW20] under the name of approximate orthogonality. If this holds true, the conjugate kernel model may be compared to a classical model derived from matrices with i.i.d. entries. To clarify this point, the assumption on $\|\Delta_X\|_{\max}$ measures how the entries of Y are dependent and how far they are from being standard Gaussian random variables individually. The additional bound on $\left\|\vec{\text{diag}}(\Delta_X)\right\|$ ensures that the latter phenomenon occurs somewhat uniformly.

In order for the deterministic equivalents to be meaningful, we need that ϵ_n converges to 0 for the Stieltjes transforms, and that ϵ'_n converges to 0 for the resolvent matrices. This is a stronger assumption, that depends both on the data matrix X, and on the activation function f. More precisely, K_X should not be too far from I_n entry-wise, to an extent that also depends on how far f is from acting linearly on Gaussians.

In practice, for the Stieltjes transforms and a general activation function f we need $\|\Delta_X\|_{\max} = o\left(n^{-1/4}\right)$. If $\zeta_2(\tilde{f}) = 0$, which happens for instance if f is odd symmetric, this convergence can be relaxed to $\|\Delta_X\|_{\max} = o\left(n^{-1/6}\right)$. If $\zeta_2(\tilde{f}) = \zeta_3(\tilde{f}) = 0$, no additional hypothesis is required as $\epsilon_n = \|\Delta_X\|_{\max}$ converges already to 0. For the resolvent matrices, the equivalent statements boil down to $\|\Delta_X\|_{\max} = o\left(n^{-1/2}\right)$ in general, $\|\Delta_X\|_{\max} = o\left(n^{-1/3}\right)$ if $\zeta_2(\tilde{f}) = 0$, and no additional hypothesis if $\zeta_2(\tilde{f}) = \zeta_3(\tilde{f}) = 0$.

Proof of Theorem 3.5.7. Y satisfies the Assumptions 3.4.2 as seen in Proposition 3.5.3. We deduce from Theorem 3.4.5 the Lipschitz concentration properties (1) and the linear concentration properties $\mathcal{G}_K(z) \in_{\|\cdot\|_F} \mathbf{G}_{\boxtimes}^{\Sigma}(z) \pm \mathcal{E}(O_z(1/\sqrt{n}))$ and $g_K(z) \in g_\nu(z) \pm \mathcal{E}(O_z(1/n))$. Moreover from Theorem 3.4.7

and Corollaries 3.5.4 and 3.4.12 :

$$\begin{aligned} |g_{\nu^{\Sigma}}(z) - g_{\nu^{\Sigma_{\text{lin}}}}(z)| &\leq O_z(|g_{\Sigma}(z) - g_{\Sigma_{\text{lin}}}(z)|) \\ &\leq O_z(\epsilon_n), \\ \left\|\left|\mathbf{G}_{\boxtimes}^{\Sigma}(z) - \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z)\right|\right\| &\leq O_z(|||\Sigma - \Sigma_{\text{lin}}|||) \\ &\leq O_z(\epsilon'_n), \end{aligned}$$

which imply the linear concentration properties (2). □

Similarly to Corollary 3.4.6, we may deduce from the deterministic equivalents the following spectral properties :

Corollary 3.5.9. *Uniformly under Assumptions 3.4.2 :*

1. $|g_K(z) - g_{\nu^{\Sigma_{\text{lin}}}}(z)| \leq \sqrt{\log n}\, O_z(\epsilon_n)$ *almost surely eventually (a.s.e.), and* $\left\|\mathcal{G}_K(z) - \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z)\right\|_{\max} \leq \sqrt{\log n}\, O_z(\epsilon'_n)$ *a.s.e.*
2. *If f is not linear, or if is f linear and the eigenvalues of K_X are bounded from below, there exists $\theta > 0$ such that $D(\mu_K, \nu^{\Sigma_{\text{lin}}}) \leq O(\epsilon_n^{\theta})$ a.s.e.*
3. *If moreover μ_{K_X} converges weakly to a measure μ_∞ and if $\gamma_n \to \gamma_\infty$, then μ_K converges weakly a.s. to $\nu_\infty = \mathrm{MP}(\gamma_\infty) \boxtimes (\mathfrak{a} + \mathfrak{b}\mu_\infty)$, and more precisely :*

$$D(\mu_K, \nu_\infty) \leq O\Big(D(\mu_{K_X}, \mu_\infty) + |\gamma_n - \gamma_\infty| + \epsilon_n^{\theta}\Big) \quad \text{a.s.e.}$$

Proof. By definition of the $O_z(\epsilon'_n)$ notation, we may find $\alpha > 0$ such that uniformly in $z \in \mathbb{C}^+$ with bounded $\Im(z)$, $\mathcal{G}_K(z) \in_{\|\cdot\|} \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z) \pm \mathcal{E}(O(\epsilon'_n(z)))$ where $\epsilon'_n(z) = \epsilon'_n \frac{|z|^{\alpha}}{\Im(z)^{2\alpha}}$. Let us now fix $z \in \mathbb{C}^+$. The maps $M \mapsto M_{ij}$ are linear and 1-Lipschitz with respect to the spectral norm. By definition of the linear concentration, there are constants $C > 0$ such that for any n, t, i and j :

$$\mathbb{P}\left(\left|\left(\mathcal{G}_K(z) - \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z)\right)_{ij}\right| \geq t\right) \leq Ce^{-\frac{t^2}{C\epsilon'_n(z)^2}}.$$

We choose $t_n = \epsilon'_n(z)\sqrt{4C\log n}$, and we use a union bound :

$$\mathbb{P}\left(\left\|\mathcal{G}_K(z) - \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z)\right\|_{\max} \geq t_n\right) \leq n^2 Ce^{-\frac{t_n^2}{C\epsilon'_n(z)^2}} = Ce^{2\log n - 4\log n} = C/n^2.$$

These probabilities are summable, so using Borel-Cantelli lemma :

$$\left\|\mathcal{G}_K(z) - \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z)\right\|_{\max} \leq t_n \leq O\left(\sqrt{\log n}\,\epsilon'_n(z)\right) \quad \text{a.s.e.}$$

We deduce that $\left\|\mathcal{G}_K(z) - \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z)\right\|_{\max} \leq \sqrt{\log n}\, O_z(\epsilon'_n)$ a.s.e. The proof for the Stieltjes transforms is similar.

If f is not linear, or if f is linear and the eigenvalues of K_X are bounded from below, then the eigenvalues of Σ_{lin} are bounded from below. In this case, for values of $\gamma_n \leq 1$, the cumulative distribution functions $\mathcal{F}_{\nu^{\Sigma_{\text{lin}}}}$ are uniformly Hölder continuous with exponent $\beta = 1/2$. For values of $\gamma_n \geq 1$, the cumulative distribution functions $\mathcal{F}_{\check{\nu}^{\Sigma_{\text{lin}}}}$ are uniformly Hölder continuous with exponent $\beta = 1/2$ (see [Cho22, Section 8.3] and Remark 3.2.9). In both cases, from Proposition 3.2.8 we may deduce the second assertion.

Finally for the third assertion, using the properties of the free multiplicative convolution [BV93, Proposition 4.13] we deduce that $\nu^{\Sigma_{\text{lin}}}$ converges weakly to ν_∞, and that :

$$\begin{aligned} D(\nu^{\Sigma_{\text{lin}}}, \nu^\infty) &\leq D(\text{MP}(\gamma_n), \text{MP}(\gamma_\infty)) + D(\mu_{\Sigma_{\text{lin}}}, \mathfrak{a} + \mathfrak{b}\mu_\infty) \\ &\leq O(|\gamma_n - \gamma_\infty| + D(\mu_{K_X}, \mu_\infty)). \end{aligned}$$

□

3.5.5 Application to another model involving entry-wise operations

In this paragraph we show how our framework applies to other models of random matrices, not strictly related to artificial neural networks.

We consider $U \in \mathbb{R}^{n\times n}$ a Gaussian random matrix filled with $\mathcal{N}$ random variables, i.i.d. within the columns and weakly correlated within the rows. More precisely, we consider $u \in \mathbb{R}^n$ a Gaussian random vector, centered, with covariance matrix :

$$S = \begin{pmatrix} 1 & 1/n & \dots & & 1/n \\ 1/n & 1 & 1/n & & \vdots \\ \vdots & 1/n & \ddots & & \\ & & & \ddots & 1/n \\ 1/n & \dots & & 1/n & 1 \end{pmatrix}.$$

We let U be the random matrix with i.i.d. sampled columns from the distribution of u. Let $f = \tanh$ be the hyperbolic tangent function, B and $D \in \mathbb{R}^{n\times n}$ independent random matrices filled with i.i.d. $\mathcal{N}$ random variables, and $Y = f(U+B)+D$. We want to apply the contents of this section to study the spectral properties of the sample covariance matrix $K = Y^\top Y/n$, its resolvent $\mathcal{G}_K(z) = (K - zI_n)^{-1}$, and its Stieltjes transform $g_K(z) = (1/n)\text{Tr}\mathcal{G}_K(z)$.

Let us denote by $J \in \mathbb{R}^{n\times n}$ the matrix whose entries are all equal to 1, so that $S = I_n + (J - I_n)/n$. In law $U = XW$, where $W \in \mathbb{R}^{n\times n}$ is a matrix filled with i.i.d. $\mathcal{N}$ entries, independent from the other sources of randomness, and $X = S^{1/2}$. Up to a transposition in the independence structure, the model Y is equal in law as $f(XW + B) + D$, and satisfies the Assumptions 3.5.2,

with $\Delta_X = (J - I_n)/n$ and $\sigma_W^2 = \sigma_X^2 = \sigma_B^2 = \sigma_D^2 = 1$. Indeed $\tilde{\sigma}^2 = 2$, $\tilde{f}(t) = \tanh(\sqrt{2}t)$, and by odd symmetry of tanh it is clear that $\zeta_0(\tilde{f}) = \zeta_2(\tilde{f}) = 0$. We also have $|||\Delta_X||| = 1 - 1/n$, $\vec{\text{diag}}(\Delta_X) = 0$, $\|\Delta_X\|_{\max} = 1/n$ and $\epsilon_n = O(1/n)$.

The constants $\mathfrak{a}$ and $\mathfrak{b}$ are linked to Gaussian moments of the hyperbolic tangent function and may be numerically approximated. The matrix $\Sigma_{\text{lin}} = \mathfrak{a} I_p + \mathfrak{b} S = (\mathfrak{a} + \mathfrak{b}) I_p + \mathfrak{b}(J - I_n)/n$ is explicitly diagonalizable and $\mu_{\Sigma_{\text{lin}}} = \frac{n-1}{n} \cdot \delta_{\mathfrak{a}+\mathfrak{b}-\mathfrak{b}/n} + \frac{1}{n} \cdot \delta_{\mathfrak{a}+2\mathfrak{b}-\mathfrak{b}/n}$. The measure $\nu^{\Sigma_{\text{lin}}} = \tilde{\nu}^{\Sigma_{\text{lin}}} = \text{MP}(1) \boxtimes \mu_L$ is characterized by its Stieltjes transform $g_{\nu^{\Sigma_{\text{lin}}}}(z)$, which is the only solution $g \in \mathbb{C}^+$ of the self-consistent equation :

$$\begin{aligned} g &= \int_{\mathbb{R}} \frac{1}{-zgt - z} \mu_{\Sigma_{\text{lin}}}(dt) \\ &= \frac{1}{-zg(\mathfrak{a} + \mathfrak{b} - \mathfrak{b}/n) - z} + \frac{1/n}{-zg(\mathfrak{a} + 2\mathfrak{b} - \mathfrak{b}/n) - z} - \frac{1/n}{-zg(\mathfrak{a} + \mathfrak{b} - \mathfrak{b}/n) - z}. \end{aligned}$$

This equation may be rewritten as a cubic polynomial equation and solved explicitly. The deterministic equivalent matrix $\mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z)$ may also be explicitly computed. The interested reader can check that :

$$\begin{aligned} \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z) = &- \frac{g_{\nu^{\Sigma_{\text{lin}}}}(z)\mathfrak{b}}{(g_{\nu^{\Sigma_{\text{lin}}}}(z)(\mathfrak{a} + \mathfrak{b} - \mathfrak{b}/n) + 1)(g_{\nu^{\Sigma_{\text{lin}}}}(z)(\mathfrak{a} + 2\mathfrak{b} - \mathfrak{b}/n) + 1)} \frac{J}{n} \\ &- \frac{1}{(g_{\nu^{\Sigma_{\text{lin}}}}(z)(\mathfrak{a} + \mathfrak{b} - \mathfrak{b}/n) + 1)} \frac{I_n}{z}. \end{aligned}$$

Theorem 3.5.7 applied to this model provides the following deterministic equivalents : $g_K(z) \in g_{\nu^{\Sigma_{\text{lin}}}}(z) \pm \mathcal{E}(O_z(1/n))$, and $\mathcal{G}_K(z) \in_{\|\cdot\|} \mathbf{G}_{\boxtimes}^{\Sigma_{\text{lin}}}(z) \pm \mathcal{E}(O_z(1/\sqrt{n}))$.

Since $\mu_{\Sigma_{\text{lin}}}$ converges weakly to $\delta_{\mathfrak{a}+\mathfrak{b}}$, ν^L and μ_K converge weakly to $\nu_\infty = \text{MP}(1) \boxtimes \delta_{\mathfrak{a}+\mathfrak{b}} = (\mathfrak{a} + \mathfrak{b})\text{MP}(1)$, and $D(\mu_K, \nu^{\Sigma_{\text{lin}}}) \leq O(n^{-\theta})$ for some $\theta > 0$. To go further, Corollary 3.5.9(3) is not helpful since the measures are discrete. As a matter of fact, $D(\mu_{\Sigma_{\text{lin}}}, \delta_{\mathfrak{a}+\mathfrak{b}}) = 1 - 1/n$ does not vanish. However we can directly compare the Stieltjes transforms of Σ_{lin} and $\delta_{\mathfrak{a}+\mathfrak{b}}$:

$$\begin{aligned} \left| g_{\Sigma_{\text{lin}}}(z) - g_{\delta_{\mathfrak{a}+\mathfrak{b}}}(z) \right| &= \left| \frac{1 - 1/n}{\mathfrak{a} + \mathfrak{b} - \mathfrak{b}/n - z} + \frac{1/n}{\mathfrak{a} + 2\mathfrak{b} - \mathfrak{b}/n - z} - \frac{1}{\mathfrak{a} + \mathfrak{b} - z} \right| \\ &\leq O_z(1/n). \end{aligned}$$

From Theorem 3.4.7, we have $\left| g_{\nu^{\Sigma_{\text{lin}}(z)}} - g_{\nu^\infty}(z) \right| \leq O_z(1/n)$, hence $D(\nu^{\Sigma_{\text{lin}}}, \nu^\infty) \leq O(n^{-\theta})$ for some $\theta > 0$ from Proposition 3.2.8. We conclude that μ_K converges to ν_∞ as speed $O(n^{-\theta})$ in Kolmogorov distance.

3.6 Single-layer neural network with random data

3.6.1 Setting

In this section we study the conjugate kernel matrix associated to a single-layer artificial neural network with random input. The hypothesis are thus the same as in the last section, excepted for the random data matrix. We consider :

— a random weight matrix $W \in \mathbb{R}^{d\times d_0}$,

— a random data matrix $X \in \mathbb{R}^{d_0\times n}$,

— two random biases matrices B and $D \in \mathbb{R}^{d\times n}$,

— an activation function $f : \mathbb{R} \to \mathbb{R}$,

— and real parameters $\sigma_W^2, \sigma_X^2 > 0$ and $\sigma_B^2, \sigma_D^2 \geq 0$ that will appear later in Assumptions 3.6.1.

As output of the neuron, we set $Y = f(WX/\sqrt{d_0} + B) + D \in \mathbb{R}^{d\times n}$, where the function f is applied entry-wise. Our goal is to investigate the spectral properties of the conjugate kernel matrix $K = Y^\top Y/d$. For $z \in \mathbb{C}^+$ we define its resolvent $\mathcal{G}_K(z) = (K - zI_n)^{-1}$ and Stieltjes transform $g_K(z) = (1/n)\mathrm{Tr}\mathcal{G}_K(z)$. We also define the following objects :

$$\begin{aligned}
\tilde{\sigma}^2 &= \sigma_W^2\sigma_X^2 + \sigma_B^2,\\
\tilde{f}(t) &= f(\tilde{\sigma}t),\\
\mathfrak{a} &= \left\|\tilde{f}\right\|_{\mathcal{H}}^2 - \frac{\sigma_W^2\sigma_X^2}{\tilde{\sigma}^2}\zeta_1(\tilde{f})^2 + \sigma_D^2,\\
\mathfrak{b} &= \zeta_1(\tilde{f})^2\frac{\sigma_W^2}{\tilde{\sigma}^2},\\
K_X &= X^\top X/d_0,\\
\Delta_X &= K_X - \sigma_X^2 I_n,\\
\Sigma_X &= \mathfrak{a}I_n + \mathfrak{b}K_X.
\end{aligned}$$

Like in the rest of this article, we sometimes omit the indices n and z for a better readability, even if we are implicitly dealing with sequences of matrices, measures, and complex functions. For reasons that we will understand later, we do not make hypothesis on the random matrix X itself, but rather on X conditioned with a high probability event (see the definition above Proposition 3.2.3).

Assumptions 3.6.1.

1. W, B and D are random, independent, with i.i.d. $\mathcal{N}(\sigma_W^2)$, $\mathcal{N}(\sigma_B^2)$ and $\mathcal{N}(\sigma_D^2)$ entries respectively.
2. $\tilde{f}$ is Lipschitz continuous and Gaussian centered, that is $\mathbb{E}\left[\tilde{f}(\mathcal{N})\right] = \mathbb{E}[f(\tilde{\sigma}\mathcal{N})] = 0$.

3. X is random, independent from W, B and D. There is an event $\mathcal{B}$ with $\mathbb{P}(\mathcal{B}^c) \leq O(\sqrt{\log n}/n)$, such that $(X|\mathcal{B}) \propto_{\|\cdot\|_F} \mathcal{E}(1)$.
4. There is a sequence ϵ_n converging to 0, such that uniformly in $\omega \in \mathcal{B}$, $|||K_X|||$ and $\left\|\vec{\text{diag}}(\Delta_X)\right\|$ are bounded, and $\|\Delta_X\|_{\max} \leq O(\epsilon_n)$.
5. The ratio $\gamma_n = \frac{n}{d}$ is bounded from above and away from 0.
6. There is a sequence $\hat{\epsilon}_n \geq 0$ such that $|\mathbb{E}[g_{K_X}(z)] - g_\tau(z)| \leq O_z(\hat{\epsilon}_n)$ for some sequence of measures τ supported on $\mathbb{R}^+$.
7. There is a sequence $\hat{\epsilon}'_n \geq \hat{\epsilon}_n$ such that $|||\mathbb{E}[\mathcal{G}_{K_X}(z)] - \mathbf{H}(z)||| \leq O_z(\hat{\epsilon}'_n)$ for some sequence of matrix functions $\mathbf{H} : \mathbb{C}^+ \to \mathbb{C}^{n\times n}$ satisfying $|||\mathbf{H}||| \leq 1/\Im(z)$.

Remark 3.6.2. Compared to Assumptions 3.5.2, in the assertions (3) and (4) we ask the data matrix X to be independent from the other random matrices, well concentrated, and uniformly approximately orthogonal on an event of high probability.

Let us explain why $\mathbb{P}(\mathcal{B}^c) \leq O(\sqrt{\log n}/n)$ is a convenient choice to simplify our results. We claim that, up to a $O_z(\sqrt{\log n}/n)$ error that will blend into similar terms, it will be equivalent for us to compute expectations on the whole probability set or on $\mathcal{B}$. Indeed, if a random function $\zeta : \mathbb{N} \times \mathbb{C}^+ \to \mathbb{C}$ satisfies the *a priori* bound $\zeta(n, z) \leq 1/\Im(z)$, then we have :

$$\begin{aligned}
|\mathbb{E}_{\mathcal{B}}[\zeta] - \mathbb{E}[\zeta]| &= |\mathbb{E}[\mathbf{1}_{\mathcal{B}}\zeta]/\mathbb{P}(\mathcal{B}) - \mathbb{E}[\zeta]| \\
&= \frac{|\mathbb{E}[\mathbf{1}_{\mathcal{B}^c}\zeta] + \mathbb{E}[\zeta]\mathbb{P}(\mathcal{B}^c)|}{\mathbb{P}(\mathcal{B})} \\
&\leq \frac{2\mathbb{P}(\mathcal{B}^c)}{1 - \mathbb{P}(\mathcal{B}^c)} \frac{1}{\Im(z)} \\
&\leq O_z(\mathbb{P}(\mathcal{B}^c)) \leq O_z(\sqrt{\log n}/n).
\end{aligned}$$

The assertions (6) and (7) correspond to deterministic equivalents for the Stieltjes transform and the resolvent of K_X. If $\zeta_1(\tilde{f}) = 0$ then $\mathfrak{b} = 0$, and as seen in Remark 3.5.6 the deterministic equivalents will not depend on X. In this case τ and $\mathbf{H}(z)$ will not appear in the results and the assertions (6) and (7) are in essence empty.

3.6.2 Deterministic equivalent and consequences

We let $\nu^{\Sigma_X} = \text{MP}(\gamma_n) \boxtimes \mu_{\Sigma_X}$, and we refer to the beginning of Section 3.4 for the definition of the deterministic equivalent matrix function $\mathbf{G}_{\boxtimes}^{\Sigma_X}(z)$. Similarly to the process used to express $\mathbf{G}_{\boxtimes}^{\Sigma_X}(z)$ as a function of μ_{K_X} and $\mathcal{G}_{K_X}$

(see Remark 3.5.6), we define the objects :

$$\begin{aligned}
\chi &= \mathrm{MP}(\gamma_n) \boxtimes (\mathfrak{a} + \mathfrak{b}\tau), \\
\check{\chi} &= (1-\gamma_n)\cdot\delta_0 + \gamma_n \cdot \chi, \\
l_{\check{\chi}}(z) &= -1/g_{\check{\chi}}(z), \\
\mathbf{K}(z) &= \frac{z^{-1} l_{\check{\chi}}(z)}{\mathfrak{b}} \mathbf{H}\left(\frac{l_{\check{\chi}}(z) - \mathfrak{a}}{\mathfrak{b}}\right) && \text{if } \mathfrak{b} \neq 0, \\
\mathbf{K}(z) &= g_{\mathfrak{a}\mathrm{MP}(\gamma_n)}(z) I_n && \text{if } \mathfrak{b} = 0.
\end{aligned}$$

Lemma 3.6.3. *Under Assumptions 3.6.1,* $|\mathbb{E}[g_{\nu^{\Sigma_X}}(z)] - g_\chi(z)| \leq \zeta_1(\tilde{f}) O_z(\hat{\epsilon}_n + \sqrt{\log n}/n)$, *and* $\left\| \mathbb{E}\left[\mathbf{G}_{\boxtimes}^{\Sigma_X}(z)\right] - \mathbf{K}(z) \right\| \leq \zeta_1(\tilde{f}) O_z(\hat{\epsilon}'_n + 1/\sqrt{n})$.

Proof. If $\zeta_1(\tilde{f}) = 0$, then Σ_X does not depend at all on X, $\mathbf{G}_{\boxtimes}^{\Sigma_X}(z) = \mathbf{K}(z)$ and $\nu^{\Sigma_X} = \chi$. If $\zeta_1(\tilde{f}) \neq 0$, from Proposition 3.2.3 applied to $(X|\mathcal{B})$, there is a constant $c > 0$, and an event $\mathcal{B}' \subset \mathcal{B}$ with $\mathbb{P}(\mathcal{B}'^c) \leq \mathbb{P}(\mathcal{B}^c) + ce^{-n/c} \leq O(\sqrt{\log n}/n)$, on which $(\Sigma_X|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(1/\sqrt{n})$. As explained in the Remark 3.6.2, given the *a priori* bounds on Stieltjes transforms and on the spectral norm of resolvent matrices, we may pass from expectations on $\mathcal{B}'$ to expectations on the full probability set, at the cost of a $O_z(\sqrt{\log n}/n)$ error term. We thus have :

$$\begin{aligned}
|\mathbb{E}_{\mathcal{B}'}[g_{\Sigma_X}(z)] - g_{\mathfrak{a}+\mathfrak{b}\tau}(z)| &= |\mathbb{E}[\mathbf{1}_{\mathcal{B}'}(g_{\Sigma_X}(z) - g_{\mathfrak{a}+\mathfrak{b}\tau}(z))]| \,/\, \mathbb{P}(\mathcal{B}') \\
&\leq \left| \frac{1}{\mathfrak{b}} \mathbb{E}\left[g_{K_X}\left(\frac{z-\mathfrak{a}}{\mathfrak{b}}\right)\right] - \frac{1}{\mathfrak{b}} g_\tau\left(\frac{z-\mathfrak{a}}{\mathfrak{b}}\right)\right| + O_z(\sqrt{\log n}/n) \\
&\leq O_z(\hat{\epsilon}_n + \sqrt{\log n}/n).
\end{aligned}$$

Proposition 3.4.13 applied to $(\Sigma_X|\mathcal{B}')$ implies that :

$$|\mathbb{E}_{\mathcal{B}'}[g_{\nu^{\Sigma_X}}(z)] - g_\chi(z)| \leq O_z(\hat{\epsilon}_n + \sqrt{\log n}/n),$$

hence $|\mathbb{E}[g_{\nu^{\Sigma_X}}(z)] - g_\chi(z)| \leq O_z(\hat{\epsilon}_n + \sqrt{\log n}/n)$. Similarly for the resolvents :

$$\begin{aligned}
\left\| \mathbb{E}_{\mathcal{B}'}[\mathcal{G}_{\Sigma_X}(z)] - \frac{1}{\mathfrak{b}} \mathbf{H}\left(\frac{z-\mathfrak{a}}{\mathfrak{b}}\right)\right\| &= \left\| \mathbb{E}\left[\frac{1}{\mathfrak{b}} \mathcal{G}_{K_X}\left(\frac{z-\mathfrak{a}}{\mathfrak{b}}\right)\right] - \frac{1}{\mathfrak{b}} \mathbf{H}\left(\frac{z-\mathfrak{a}}{\mathfrak{b}}\right)\right\| \\
&\quad + O_z(\sqrt{\log n}/n) \\
&\leq O_z(\hat{\epsilon}'_n + \sqrt{\log n}/n),
\end{aligned}$$

which implies by Proposition 3.4.13 that $\left\| \mathbb{E}_{\mathcal{B}'}\left[\mathbf{G}_{\boxtimes}^{\Sigma_X}(z)\right] - \mathbf{K}(z)\right\| \leq O_z(\hat{\epsilon}_n + \hat{\epsilon}'_n + 1/\sqrt{n}) \leq O_z(\hat{\epsilon}'_n + 1/\sqrt{n})$, and $\left\| \mathbb{E}\left[\mathbf{G}_{\boxtimes}^{\Sigma_X}(z)\right] - \mathbf{K}(z)\right\| \leq O_z(\hat{\epsilon}'_n + 1/\sqrt{n})$. □

We remind our reader that the sequences ϵ appearing in the Assumptions 3.6.1 measure the lack of orthogonality of X for ϵ_n, and the convergence speeds

in the deterministic equivalents $g_{K_X}(z) \approx g_\tau(z)$ for $\hat{\epsilon}_n$, and $\mathcal{G}_{K_X}(z) \approx \mathbf{H}(z)$ for $\hat{\epsilon}'_n$ respectively. For the next results, let us denote :

$$\tilde{\epsilon}_n = \sqrt{\log n}/n + \zeta_1(\tilde{f})\hat{\epsilon}_n + \epsilon_n + \sqrt{n}\zeta_2(\tilde{f})^2\epsilon_n^2 + \sqrt{n}\zeta_3(\tilde{f})^2\epsilon_n^3,$$
$$\tilde{\epsilon}'_n = 1/\sqrt{n} + \zeta_1(\tilde{f})\hat{\epsilon}'_n + \epsilon_n + n\zeta_2(\tilde{f})^2\epsilon_n^2 + n\zeta_3(\tilde{f})^2\epsilon_n^3.$$

$\tilde{\epsilon}_n$ and $\tilde{\epsilon}'_n$ will correspond to the new convergence speeds in the deterministic equivalents $g_K(z) \approx g_\chi(z)$ and $\mathcal{G}_K(z) \approx \mathbf{K}(z)$.

Theorem 3.6.4. *Uniformly under Assumptions 3.6.1, there is an event $\mathcal{B}' \subset \mathcal{B}$ with $\mathbb{P}(\mathcal{B}'^c) \leq O(\sqrt{\log n}/n)$, such that the following conditional concentration properties hold true :*

1. $(Y|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(1)$, $(g_K(z)|\mathcal{B}') \propto \mathcal{E}(O_z(1/n))$, *and* $(\mathcal{G}_K(z)|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(O_z(1/\sqrt{n}))$.
2. $(g_K(z)|\mathcal{B}') \in g_\chi(z) \pm \mathcal{E}(O_z(\tilde{\epsilon}_n))$ *and* $(\mathcal{G}_K(z)|\mathcal{B}') \in_{\|\cdot\|} \mathbf{K}(z) \pm \mathcal{E}(O_z(\tilde{\epsilon}'_n))$.

Proof. For a better readability in the upcoming arguments, we choose to omit the spectral parameters z in most of our notations. From Proposition 3.2.3 applied to $(W, (X|\mathcal{B})) \propto_{\|\cdot\|_F} \mathcal{E}(1)$, there is a constant $c > 0$ and an event $\mathcal{B}' \subset \mathcal{B}$, with $\mathbb{P}(\mathcal{B}'^c) \leq \mathbb{P}(\mathcal{B}^c) + ce^{-n/c} \leq O(\sqrt{\log n}/n)$, such that $(WX|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(\sqrt{n})$. The map $(U, B, D) \mapsto f(U + B) + D$ is Lipschitz with respect to the Frobenius norm, and by independence $\Big((WX/\sqrt{d_0}\,|\mathcal{B}'), B, D\Big) \propto_{\|\cdot\|_F} \mathcal{E}(1)$, thus $(Y|\mathcal{B}') = f\Big((WX/\sqrt{d_0}\,|\mathcal{B}') + B\Big) + D \propto_{\|\cdot\|_F} \mathcal{E}(1)$. The general concentration properties recalled in Proposition 3.2.2 imply that $(g_K|\mathcal{B}') \propto \mathcal{E}(O_z(1/n))$ and $(\mathcal{G}_K|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(O_z(1/\sqrt{n}))$.

For the second assertion, given the linear concentration properties $(\mathcal{G}_K|\mathcal{B}') \in_{\|\cdot\|} \mathbb{E}[(\mathcal{G}_K|\mathcal{B}')] \pm \mathcal{E}(O_z(1/\sqrt{n}))$ and $(g_K|\mathcal{B}') \in \mathbb{E}[(g_K|\mathcal{B}')] \pm \mathcal{E}(O_z(1/n))$, we only need to prove that $|\mathbb{E}[(g_K|\mathcal{B}')] - g_\chi| \leq O_z(\tilde{\epsilon}_n)$, and that $\|\mathbb{E}[(\mathcal{G}_K|\mathcal{B}')] - \mathbf{K}\| \leq O_z(\tilde{\epsilon}'_n)$. To compare these expectations, as explained in the Remark 3.6.2, we may assume without loss of generality that $\mathcal{B}' = \Omega$, because the additional $O_z(\sqrt{\log n}/n)$ error terms will not change the final estimates.

We apply Theorem 3.5.7 to the models with deterministic data matrices $X(\omega)$, uniformly in the outcomes $\omega \in \Omega_X$. Since $1/\sqrt{n} \leq \tilde{\epsilon}'_n$ and $\epsilon_n + n\zeta_2(\tilde{f})^2\epsilon_n^2 + n\zeta_3(\tilde{f})^2\epsilon_n^3 \leq \tilde{\epsilon}'_n$, we obtain that $\mathcal{G}_{K(\omega)} \in_{\|\cdot\|} \mathbf{G}_\boxtimes^{\Sigma_X(\omega)} \pm \mathcal{E}(O_z(\tilde{\epsilon}'_n))$. As a consequence, we get uniformly in $\omega \in \Omega_X$:

$$\Big\|\mathbb{E}_{W,B,D}[\mathcal{G}_K(\omega)] - \mathbf{G}_\boxtimes^{\Sigma_X(\omega)}\Big\| \leq O_z(\tilde{\epsilon}'_n).$$

Since X is independent from the other sources of randomness, for any measurable function Φ we have $\mathbb{E}[\Phi(W, X, B, D)|X] = \mathbb{E}_{W,B,D}[\Phi(W, X, B, D)]$.

We integrate the above inequality with respect to X, and using the tower property of conditional expectation :

$$\begin{aligned}\left|\left|\left|\mathbb{E}[\mathcal{G}_K] - \mathbb{E}\left[\mathbf{G}_{\boxtimes}^{\Sigma_X}\right]\right|\right| &= \left|\left|\left|\mathbb{E}\left[\mathbb{E}\left[\mathcal{G}_K - \mathbf{G}_{\boxtimes}^{\Sigma_X}|X\right]\right]\right|\right| \\ &\leq \mathbb{E}_X\left[\left|\left|\left|\mathbb{E}_{W,B,D}[\mathcal{G}_K] - \mathbf{G}_{\boxtimes}^{\Sigma_X}\right|\right|\right|\right] \leq O_z(\tilde{\epsilon}'_n).\end{aligned}$$

From Lemma 3.6.3 we also have $\left|\left|\left|\mathbb{E}\left[\mathbf{G}_{\boxtimes}^{\Sigma_X}\right] - \mathbf{K}\right|\right|\right| \leq \zeta_1(\tilde{f})O_z(\tilde{\epsilon}'_n + 1/\sqrt{n}) \leq O_z(\tilde{\epsilon}'_n)$, hence $|||\mathbb{E}[\mathcal{G}_K] - \mathbf{K}||| \leq O_z(\tilde{\epsilon}'_n)$. The proof for the Stieltjes transforms is similar. □

Corollary 3.6.5. *Uniformly under Assumptions 3.6.1 :*

1. $|g_K(z) - g_\chi(z)| \leq \sqrt{\log n}\, O_z(\tilde{\epsilon}_n)$ *almost surely eventually (a.s.e.), and* $\|\mathcal{G}_K(z) - \mathbf{K}(z)\|_{\max} \leq \sqrt{\log n}\, O_z(\tilde{\epsilon}'_n)$ *a.s.e.*
2. *If f is not linear, or if is f linear and the measures τ are supported on the same compact of $(0, \infty)$, there exists $\theta > 0$ such that $D(\mu_K, \chi) \leq O(\tilde{\epsilon}_n^\theta)$ a.s.e.*
3. *If moreover τ converges weakly to a measure τ_∞, and if $\gamma_n \to \gamma_\infty$, then μ_K converges weakly a.s. to $\chi_\infty = \mathrm{MP}(\gamma_\infty) \boxtimes (\mathfrak{a} + \mathfrak{b}\tau_\infty)$, and more precisely :*

$$D(\mu_K, \chi_\infty) \leq O\Big(D(\tau, \tau_\infty) + |\gamma_n - \gamma_\infty| + \tilde{\epsilon}_n^\theta\Big) \quad \text{a.s.e.}$$

As we did for Corollary 3.4.6, we will not prove this result here, but rather prompt our reader to consult the proof of Corollary 3.5.9 which is extremely similar.

3.6.3 Application to data matrices with i.i.d. columns

In this paragraph we focus on a fairly general setting, where the data matrix X is made of independent samples, and we explore the consequences given by our deterministic equivalents. Let us first mention a general framework on which the Assumption 3.6.1(4) holds true.

Proposition 3.6.6 ([FW20], Proposition 3.3)**.** *Let $X \in \mathbb{R}^{d_0 \times n}$ be a random matrix whose columns are i.i.d. sampled from the distribution of a random vector $x \in \mathbb{R}^{d_0}$, such that $x \propto_{\|\cdot\|} \mathcal{E}(1)$, $\mathbb{E}[x] = 0$, and $\mathbb{E}\left[\|x\|^2\right] = \sigma_x^2 d_0$. We also assume the ratio $\frac{n}{d_0}$ to be bounded from above and away from 0.*

Then there is an event $\mathcal{B}$ with $\mathbb{P}(\mathcal{B}^c) \leq O(1/n)$, such that uniformly in $\omega \in \mathcal{B}$, $|||K_X|||$ and $\left\|\vec{\mathrm{diag}}(\Delta_X)\right\|$ are bounded, and $\|\Delta_X\|_{\max} \leq O(\epsilon_n)$ with $\epsilon_n = \sqrt{\log n / n}$.

Remark 3.6.7. The exact statement in [FW20] is actually more general, because it requires for x to be concentrated in a weaker sense, called convex concentration. We will not digress on this here but rather refer our reader to [LC20, Section 1.7], which presents in all the details this variant of concentration. Also note that in the above result, we only need concentration for the columns of X, while in our deterministic equivalent, we need concentration for the whole matrix X.

To obtain deterministic equivalents for the input data matrix, it is of course possible to use again Theorem 3.4.5 :

Proposition 3.6.8. *If $X \in \mathbb{R}^{d_0 \times n}$ is a random matrix, $\propto_{\|\cdot\|_F} \mathcal{E}(1)$ concentrated, whose columns are i.i.d. sampled from the distribution of a random vector $x \in \mathbb{R}^{d_0}$, with $\|\mathbb{E}[x]\|$ and $\|\|\mathbb{E}[K_X]\|\|$ bounded, and if the ratio n/d_0 is bounded from above and away from 0, then with $\tau = \mathrm{MP}(n/d_0) \boxtimes \mu_{\mathbb{E}[K_X]}$:*

$$|\mathbb{E}[g_{K_X}(z)] - g_\tau(z)| \leq O_z(1/n)$$
$$\left\|\left|\mathbb{E}[\mathcal{G}_{K_X}(z)] - \mathbf{G}_{\boxtimes}^{\mathbb{E}[K_X]}(z)\right|\right\| \leq O_z(1/\sqrt{n}).$$

If we combine the previous propositions, we obtain deterministic equivalents for the conjugate kernel model in a fairly general setting where the data matrix is made of i.i.d. training samples. This encompasses in particular the case where X is a matrix with i.i.d. $\mathcal{N}$ entries, which was the original model studied in [PW17].

Remark 3.6.9. Note that the typical order of magnitude $\epsilon_n = \sqrt{\log n/n}$ given by Proposition 3.6.6 is good enough for a meaningful equivalent of the Stieltjes transform, with a $O_z(\log n/\sqrt{n})$ convergence speed. However it is not small enough for the resolvent, where we would have an $O_z(\zeta_2(\tilde{f})^2 \log n)$ error term, unless of course $\zeta_2(\tilde{f}) = 0$ (see Remark 3.5.8). If $\zeta_2(\tilde{f}) = 0$, we obtain a $O_z\Big((\log n)^{3/2}/\sqrt{n}\Big)$ error term for the deterministic equivalent of the resolvent.

This condition $\zeta_2(\tilde{f}) = 0$, and more generally the Hermite coefficients of the activation function f, appear in other articles that study the conjugate kernel model. Let us consider indeed $\tilde{Y} = \sqrt{\mathfrak{a}}Z + \sqrt{\mathfrak{b}}WX/\sqrt{d_0}$, and $\tilde{K} = \tilde{Y}^\top \tilde{Y}/d$, where Z is a third random matrix, independent from the others, and filled with i.i.d. $\mathcal{N}$ entries. Using Theorem 3.5.7 conditionally on X and similar arguments to those of this paper, we see that $\mathcal{G}_{\tilde{K}}(z)$ also admits $\mathbf{G}_{\boxtimes}^{\mathbb{E}[K_X]}(z)$ as a deterministic equivalent, with a $O_z\Big((\log n)^{3/2}/\sqrt{n}\Big)$ error term in the case where $\zeta_2(\tilde{f}) = 0$.

In [BP21, Theorem 2.3], using combinatorics it is shown that the biggest eigenvalues of both models behave similarly if $\zeta_2(\tilde{f}) = 0$. Although we could not manage to retrieve this property with our deterministic equivalent solely, there is without a doubt a connection between these statements. [BP21] also provides other equivalent models in the case where $\zeta_2(\tilde{f}) \neq 0$, which we could not relate to our results.

3.7 Multi-layer neural network model

In this section we consider the conjugate kernel matrix associated to an artificial neural network with L hidden layers and a random input matrix :

$$\begin{aligned}
X_0 \to X_1 &= f_1\Big(W_1 X_0/\sqrt{d_0} + B_1\Big) + D_1 \\
X_1 \to X_2 &= f_2\Big(W_2 X_1/\sqrt{d_1} + B_2\Big) + D_2 \\
&\vdots \\
X_l \to X_{l+1} &= f_{l+1}\Big(W_{l+1} X_l/\sqrt{d_l} + B_{l+1}\Big) + D_{l+1} \\
&\vdots \\
X_{L-1} \to X_L &= f_L\Big(W_L X_{L-1}/\sqrt{d_{L-1}} + B_L\Big) + D_L
\end{aligned}$$

The initial data $X_0 \in \mathbb{R}^{d_0 \times n}$ is a random matrix, associated to a real parameter $\sigma_{X_0}^2 > 0$ that will appear later in Assumptions 3.7.1. Each layer $l \in [\![1, L]\!]$ is made of :

— a random weight matrix $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$,

— two random biases matrices B_l and $D_l \in \mathbb{R}^{d_l \times n}$,

— an activation function $f_l : \mathbb{R} \to \mathbb{R}$,

— and real parameters $\sigma_{W_l}^2 > 0$ and $\sigma_{B_l}^2, \sigma_{D_l}^2 \geq 0$ that will appear later in Assumptions 3.7.1.

At each layer, we define the conjugate kernel matrix $K_l = X_l^\top X_l/d_l$, and for $z \in \mathbb{C}^+$ its resolvent $\mathcal{G}_{K_l}(z) = (K_l - zI_n)^{-1}$, and Stieltjes transform $g_{K_l}(z) = (1/n)\mathrm{Tr}\mathcal{G}_{K_l}(z)$. We define by induction the following objects :

$$\begin{aligned}
\tilde{\sigma}_l^2 &= \sigma_{W_l}^2 \sigma_{X_{l-1}}^2 + \sigma_{B_l}^2, \\
\tilde{f}_l(t) &= f_l(\tilde{\sigma}_l t), \\
\sigma_{X_l} &= \left\|\tilde{f}_l\right\|_{\mathcal{H}}^2 + \sigma_{D_l}^2, \\
\mathfrak{a}_l &= \left\|\tilde{f}_l\right\|_{\mathcal{H}}^2 - \frac{\sigma_{W_l}^2 \sigma_{X_l}^2}{\tilde{\sigma}_l^2} \zeta_1(\tilde{f}_l)^2 + \sigma_{D_l}^2, \\
\mathfrak{b}_l &= \zeta_1(\tilde{f}_l)^2 \frac{\sigma_{W_l}^2}{\tilde{\sigma}_l^2}, \\
\Delta_{X_l} &= K_l - \sigma_{X_l}^2 I_n, \\
\Sigma_{X_l} &= \mathfrak{a}_l I_n + \mathfrak{b}_l K_l.
\end{aligned}$$

Assumptions 3.7.1. 1. W_l, B_l and D_l are random, independent as a family for $l \in [\![1, L]\!]$, with i.i.d. $\mathcal{N}(\sigma_{W_l}^2)$, $\mathcal{N}(\sigma_{B_l}^2)$ and $\mathcal{N}(\sigma_{D_l}^2)$ entries respectively.

2. $\tilde{f}_l$ are Lipschitz continuous and Gaussian centered, that is $\mathbb{E}\Big[\tilde{f}_l(\mathcal{N})\Big] = \mathbb{E}[f(\tilde{\sigma}_l\mathcal{N})] = 0$.

3. X_0 is random, independent from all the other matrices. There is an event $\mathcal{B}$ with $\mathbb{P}(\mathcal{B}^c) \leq O(\sqrt{\log n}/n)$, such that $(X_0|\mathcal{B}) \propto_{\|\cdot\|_F} \mathcal{E}(1)$.

4. There is a sequence ϵ_n converging to 0, with $\sqrt{\log n/n} \leq O(\epsilon_n)$, such that uniformly in $\omega \in \mathcal{B}$, $|||K_0|||$ and $\left\|\vec{\text{diag}}(\Delta_{X_0})\right\|$ are bounded, and $\|\Delta_{X_0}\|_{\max} \leq O(\epsilon_n)$.

5. The ratios $\gamma_n^{(l)} = \frac{n}{d_l}$ are bounded from above and away from 0.

6. There is a sequence $\hat{\epsilon}_n^{(0)} \geq 0$ such that $\left|\mathbb{E}[g_{K_0}(z)] - g_{\chi_n^{(0)}}(z)\right| \leq O_z(\hat{\epsilon}_n^{(0)})$ for some sequence of measures $\chi_n^{(0)}$ supported on $\mathbb{R}^+$.

7. There is a sequence $\hat{\epsilon'}_n^{(0)} \geq \hat{\epsilon}_n^{(0)}$ such that $|||\mathbb{E}[\mathcal{G}_{K_0}(z)] - \mathbf{G}_0(z)||| \leq O_z(\hat{\epsilon'}_n^{(0)})$ for some sequence of matrix functions $\mathbf{G}_0 : \mathbb{C}^+ \times \mathbb{C}^{n\times n}$, satisfying $|||\mathbf{G}_0(z)||| \leq 1/\Im(z)$.

Starting from the deterministic equivalents $\mu_{K_0} \approx \chi_n^{(0)}$ and $\mathcal{G}_{K_0}(z) \approx \mathbf{G}_0(z)$, we define by induction for $l \in [\![1, L]\!]$:

$$\begin{aligned}
\chi_n^{(l)} &= \text{MP}\Big(\gamma_n^{(l)}\Big) \boxtimes \Big(\mathfrak{a}_l + \mathfrak{b}_l \chi_n^{(l-1)}\Big),\\
\check{\chi}_n^{(l)} &= (1 - \gamma_n^{(l)}) \cdot \delta_0 + \gamma_n^{(l)} \cdot \chi_n^{(l)},\\
l_{\check{\chi}_n^{(l)}}(z) &= -1/g_{\check{\chi}_n^{(l)}}(z),\\
\mathbf{G}_l(z) &= \frac{z^{-1} l_{\check{\chi}_n^{(l)}}(z)}{\mathfrak{b}_l} \mathbf{G}_{l-1}\left(\frac{l_{\check{\chi}_n^{(l)}}(z) - \mathfrak{a}_l}{\mathfrak{b}_l}\right) \qquad \text{if } \zeta_1(\tilde{f}_l) \neq 0,\\
\mathbf{G}_l(z) &= g_{\mathfrak{a}_l \text{MP}(\gamma_n^{(l)})}(z) I_n \qquad \text{if } \zeta_1(\tilde{f}_l) = 0.
\end{aligned}$$

We remind our reader that the sequence ϵ_n measures the lack of orthogonality of the input data matrix X_0. We define by induction the following sequences, corresponding respectively to the error terms in the approximation of the Stieltjes transforms and the resolvent at layer l :

$$\begin{aligned}
\hat{\epsilon}_n^{(l+1)} &= \sqrt{\log n}/n + \zeta_1(\tilde{f}_l)\hat{\epsilon}_n^{(l)} + \epsilon_n + \sqrt{n}\zeta_2(\tilde{f}_l)^2\epsilon_n^2 + \sqrt{n}\zeta_3(\tilde{f}_l)^2\epsilon_n^3\\
\hat{\epsilon}_n'^{(l+1)} &= 1/\sqrt{n} + \zeta_1(\tilde{f}_l)\hat{\epsilon}_n'^{(l)} + \epsilon_n + n\zeta_2(\tilde{f}_l)^2\epsilon_n^2 + n\zeta_3(\tilde{f}_l)^2\epsilon_n^3.
\end{aligned}$$

Theorem 3.7.2. *Uniformly under Assumptions 3.7.1, there is a constant $c > 0$ and an event $\mathcal{B}' \subset \mathcal{B}$ with $\mathbb{P}(\mathcal{B}'^c) \leq O(\sqrt{\log n}/n)$, such that for any $l \in [\![1, L]\!]$, the following conditional concentration properties hold true :*

1. $(X_l|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(1)$, $(g_{K_l}(z)|\mathcal{B}') \propto \mathcal{E}(O_z(1/n))$, *and* $(\mathcal{G}_{K_l}(z)|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(O_z(1/\sqrt{n}))$.
2. $(g_{K_l}(z)|\mathcal{B}') \in g_{\chi_n^{(l)}}(z) \pm \mathcal{E}\Big(O_z(\hat{\epsilon}_n^{(l)})\Big)$ *and* $(\mathcal{G}_{K_l}(z)|\mathcal{B}') \in_{\|\cdot\|} \mathbf{G}_l(z) \pm \mathcal{E}\Big(O_z(\hat{\epsilon}_n'^{(l)})\Big)$.

Proof. By induction on the layers using Proposition 3.2.3, we can prove that there is an event $\mathcal{B}' \subset \mathcal{B}$ with $\mathbb{P}(\mathcal{B}'^c) \leq \mathbb{P}(\mathcal{B}^c) + ce^{-n/c} \leq O(\sqrt{\log n}/n)$, such that $(X_l|\mathcal{B}') \propto_{\|\cdot\|_F} \mathcal{E}(1)$ for all layers $l \in [\![1, L]\!]$. The concentration properties of $\mathcal{G}_{K_l}(z)$ and $g_{K_l}(z)$ follow. Again by induction, using Lemma 3.5.5 on $\mathbb{E}[X_{l+1}|X_l]$ proves that on $\mathcal{B}'$, $\left\|\vec{\text{diag}}(\Delta_{X_l})\right\|$ and $|||K_l|||$ remain bounded, and that $\|\Delta_{X_l}\|_{\max} \leq O(\epsilon_n + \sqrt{\log n/n}) \leq O_z(\epsilon_n)$ for all $l \in [\![1, L]\!]$. The random matrices $(X_l|\mathcal{B}')$ satisfy the Assumptions 3.6.1 uniformly. By repeatedly using Theorem 3.6.4, we get the deterministic equivalents for all layers. The error terms $\hat{\epsilon}_n^{(l)}$ and $\hat{\epsilon}_n'^{(l)}$ are given by the formulas above Theorem 3.6.4. □

Remark 3.7.3. The deterministic equivalents are only meaningful if the error terms $\hat{\epsilon}_n'^{(l)}$ and $\hat{\epsilon}_n^{(l)}$ vanish when $n \to \infty$. Similarly to Remark 3.5.8, let us mention a few cases where their expressions may be greatly simplified :

— If $\zeta_1(\tilde{f}_l) = 0$ for some layer l, then for all subsequent layers $k \geq l$, the error terms $\hat{\epsilon}_n'^{(k)}$ and $\hat{\epsilon}_n^{(k)}$ do not depend on $\hat{\epsilon}_n^{(0)}$ and $\hat{\epsilon}_n'^{(0)}$ anymore.

— If $\epsilon_n = o(n^{-1/4})$, then $\hat{\epsilon}_n^{(l)} = O\Big(\hat{\epsilon}_n^{(0)} + \sqrt{n}\epsilon_n^2\Big)$.

— If $\zeta_2(\tilde{f}_l) = 0$ and $\epsilon_n = o(n^{-1/6})$, then $\hat{\epsilon}_n^{(l)} = O\Big(\hat{\epsilon}_n^{(0)} + \sqrt{n}\epsilon_n^3\Big)$.

— If $\zeta_2(\tilde{f}_l) = \zeta_3(\tilde{f}_l) = 0$, then $\hat{\epsilon}_n^{(l)} = O\Big(\hat{\epsilon}_n^{(0)} + \sqrt{\log n}/n\Big)$.

— If $\epsilon_n = o(n^{-1/2})$, then $\hat{\epsilon}_n'^{(l)} = O\Big(\hat{\epsilon}_n^{(0)} + \hat{\epsilon}_n'^{(0)} + n\epsilon_n^2\Big)$.

— If $\zeta_2(\tilde{f}_l) = 0$ and $\epsilon_n = o(n^{-1/3})$, then $\hat{\epsilon}_n'^{(l)} = O\Big(\hat{\epsilon}_n'^{(0)} + n\epsilon_n^3\Big)$.

— If $\zeta_2(\tilde{f}_l) = \zeta_3(\tilde{f}_l) = 0$, then $\hat{\epsilon}_n'^{(l)} = O\Big(\hat{\epsilon}_n'^{(0)} + 1/\sqrt{n}\Big)$.

In the case $\epsilon_n = O(\sqrt{\log n/n})$ corresponding to a data matrix X_0 with i.i.d. columns (see Proposition 3.6.6), and starting from typical $\hat{\epsilon}_n^{(0)} = O_z(1/n)$ and $\hat{\epsilon}_n'^{(0)} = O_z(1/\sqrt{n})$ equivalents for the Stieltjes transform and the resolvent of K_0 respectively, Theorem 3.7.2 gives an $O_z(\log n/\sqrt{n})$ equivalent for the Stieltjes transform. The error for the resolvents does not vanish in general because of the $O_z(\zeta_2(\tilde{f})^2 \log n)$ error term. If $\zeta_2(\tilde{f}) = 0$ however, we obtain a $O_z((\log n)^{3/2}/\sqrt{n})$ approximation.

Corollary 3.7.4. *Uniformly under Assumptions 3.7.1 :*

1. $\left|g_{K_l}(z) - g_{\chi_n^{(l)}}(z)\right| \leq \sqrt{\log n}\, O_z(\hat{\epsilon}_n^{(l)})$ *almost surely eventually (a.s.e.), and* $\|\mathcal{G}_{K_l}(z) - \mathbf{G}_l(z)\|_{\max} \leq \sqrt{\log n}\, O_z(\hat{\epsilon}_n')$ *a.s.e.*
2. *If $\tilde{f}_l$ is not linear, or if the measures $\chi_n^{(0)}$ are supported on the same compact of $(0, \infty)$, there exists $\theta > 0$ such that $D(\mu_{K_l}, \chi_n^{(l)}) \leq O(\hat{\epsilon}_n^{(l)\theta})$ a.s.e.*

3. *If moreover* $\chi_n^{(0)}$ *converges weakly to a measure* $\chi_\infty^{(0)}$*, and if all dimension ratios* $\gamma_n^{(k)} \to \gamma_\infty^{(k)} > 0$ *, then* μ_{K_l} *converges a.s. to the measure* $\chi_\infty^{(l)}$ *defined by induction as :*

$$\chi_\infty^{(l)} = \mathrm{MP}\left(\gamma_\infty^{(l)}\right) \boxtimes \left(\mathfrak{a}_l + \mathfrak{b}_l \chi_\infty^{(l-1)}\right).$$

More precisely :

$$D(\mu_{K_l}, \chi_\infty^{(l)}) \leq O\left(D(\chi_n^{(0)}, \chi_\infty^{(0)}) + \max_{1\leq k\leq l} |\gamma_n^{(k)} - \gamma_\infty^{(k)}| + \hat{\epsilon}_n^{(l)\theta}\right) \quad \text{a.s.e.}$$

Again we will not prove this result here, but rather refer to the proof of Corollary 3.5.9 which is similar.

Remark 3.7.5. Our result generalizes previously known global laws on the conjugate kernel model. Employing our notations, [FW20, Theorem 3.4] states that, without bias in the model, if f is twice differentiable, $\epsilon_n = o(n^{-1/4})$, and $\chi_n^{(0)}$ converges weakly to $\chi_\infty^{(0)}$, then μ_{K_l} converges weakly to $\chi_\infty^{(l)}$ a.s. Taking into account the Remark 3.7.3, in this setting we have $\hat{\epsilon}_n^{(l)} = O(\hat{\epsilon}_n^{(0)} + \sqrt{n}\epsilon_n^2) = O(\hat{\epsilon}_n^{(0)} + o(1))$, and we retrieve the a.s. convergence of μ_{K_l} towards $\chi_\infty^{(l)}$ weakly, supplemented with quantitative estimates for the Stieltjes transforms and the Kolmogorov distances.

Remark 3.7.6. The coefficients $\mathfrak{a}_l$ and $\mathfrak{b}_l$, appearing in the deterministic equivalent, remain unchanged if we swap $\tilde{f}_l$ for the linear function $t \mapsto \zeta_1(\tilde{f}_l)t$, and $\sigma_{D_l}^2$ for ${\sigma'_{D_l}}^2 = \sigma_{D_l}^2 + \left\|\tilde{f}_l\right\|_{\mathcal{H}}^2 - \zeta_1(\tilde{f}_l)^2$. In other words, our original model, with non-linear activations functions f_l, admits the same deterministic equivalents as the following model, involving only linear functions :

$$X_l \to X_{l+1} = \frac{\zeta_1(\tilde{f}_l)}{\tilde{\sigma}_l}\left(W_{l+1}X_l/\sqrt{d_l} + B_{l+1}\right) + D_{l+1} + \left(\left\|\tilde{f}_l\right\|_{\mathcal{H}}^2 - \zeta_1(\tilde{f}_l)^2\right)^{1/2} D'_{l+1},$$

where D'_{l+1} is an additional bias matrix with i.i.d. $\mathcal{N}$ entries.

This phenomenon may be seen as a linearization principle for the conjugate kernel model, similar to the Gaussian equivalence principle highlighted in a different setting in [GLR+22].

3.8 Appendix : Bounds on Kolmogorov distances between empirical measures

Let us remind the notations $\mathcal{F}_\nu$ for the cumulative distribution function of a measure ν, and $D(\nu,\mu) = \sup_{t\in\mathbb{R}} |\mathcal{F}_\nu(t) - \mathcal{F}_\mu(t)|$ for the Kolmogorov distance between two measures ν and μ.

It is a well-known fact that the convergence in Kolmogorov distance implies the weak convergence for probability measures, and there is even an equivalence if the limiting measure admits a Hölder continuous cumulative distribution function ([GH03]). In [BM20] and [Cho22], the authors propose a general method to derive a convergence speed in Kolmogorov distance from estimates on the Stieltjes transforms. This method implies for instance our Proposition 3.2.8. However the techniques employed are not well suited to work with two discrete measures like empirical spectral distributions.

Two matrices close in spectral norm admit the same limiting spectral distribution if it exists. This does not imply any bound on the Kolmogorov distances in general because the measures are discrete. In this section, we show how a quantitative result may still be obtained, provided the limiting empirical measure is regular enough. We strongly incite the reader to first examine [Cho22, Section 8], where the technical tools are explained in full detail.

Proposition 3.8.1. *Let Σ and $\tilde{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric matrices such that :*

1. *$|||\Sigma|||$ and $\left|\left|\left|\tilde{\Sigma}\right|\right|\right|$ are bounded, and $\left|\left|\left|\Sigma - \tilde{\Sigma}\right|\right|\right|$ converges to 0.*
2. *$\mu_{\tilde{\Sigma}}$ converges weakly to some probability measure ν^∞, and $\mathcal{F}_{\nu^\infty}$ is Hölder continuous for some parameter $\beta > 0$.*

Then μ_Σ converges weakly to ν^∞, and more precisely in Kolmogorov distance :

$$D(\mu_\Sigma, \nu^\infty) \leq O\left(\left|\left|\left|\Sigma - \tilde{\Sigma}\right|\right|\right|^{\frac{\beta}{4+2\beta}} + D(\mu_{\tilde{\Sigma}}, \nu^\infty) \right).$$

Lemma 3.8.2. *For any $y \in (0,1)$ and $A > 0$:*

$$D(\mu_{\tilde{\Sigma}}, \mu_\Sigma) \leq O\left(A\,\frac{\left|\left|\left|\Sigma - \tilde{\Sigma}\right|\right|\right|}{y^2} + \frac{1}{y^2 A} + y^\beta + D(\mu_{\tilde{\Sigma}}, \nu^\infty) \right).$$

Proof. We closely follow [Cho22, Section 3.1], with the only key difference that we plug in Bai's Inequality the following bound :

$$\left|\mathcal{F}_{\mu_{\tilde{\Sigma}}}(x+t) - \mathcal{F}_{\mu_{\tilde{\Sigma}}}(x)\right| \leq |\mathcal{F}_{\nu^\infty}(x+t) - \mathcal{F}_{\nu^\infty}(x)| + 2\,D(\mu_{\tilde{\Sigma}}, \nu^\infty).$$

We thus obtain :

$$D(\mu_{\tilde{\Sigma}}, \mu_\Sigma) \leq \frac{2}{\pi}\left(\int_{\mathbb{R}} \left|g_{\mu_{\tilde{\Sigma}}} - g_{\mu_\Sigma}\right|(t+iy)dt \right. \\ \left. + \frac{1}{y}\sup_{x\in\mathbb{R}} \int_{\left[\pm 2y\tan\left(\frac{3\pi}{8}\right)\right]} |\mathcal{F}_{\nu^\infty}(x+t) - \mathcal{F}_{\nu^\infty}(x)|dt\right) \\ + O(D(\mu_{\tilde{\Sigma}}, \nu^\infty)).$$

For $z \in \mathbb{C}^+$ a classical application of the resolvent identity gives $|g_\Sigma(z) - g_{\tilde{\Sigma}}(z)| \leq \frac{|||\Sigma - \tilde{\Sigma}|||}{\Im(z)^2}$. The rest of the proof is exactly the same as in [Cho22, Section 3.1]. □

Proof of Theorem 3.8.1. We optimize y and A in the above lemma by choosing $y_n = |||\Sigma - \tilde{\Sigma}|||^{\frac{1}{4+2\beta}}$ and $A_n = |||\Sigma - \tilde{\Sigma}|||^{-1/2}$, which leads to the bound $D(\mu_{\tilde{\Sigma}}, \mu_\Sigma) \leq O\left(|||\Sigma - \tilde{\Sigma}|||^{\frac{\beta}{4+2\beta}} + D(\mu_{\tilde{\Sigma}}, \nu^\infty)\right)$. A final triangular inequality proves the proposition. □

Bibliographie

[Ada15] Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20 :1–13, 2015.

[ASS20] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132 :428–446, 2020.

[Bai93] Zhidong D Bai. Convergence rate of expected spectral distributions of large random matrices. part ii. sample covariance matrices. *The Annals of Probability*, pages 649–672, 1993.

[Bai08] Zhi Dong Bai. Convergence rate of expected spectral distributions of large random matrices part i : Wigner matrices. In *Advances In Statistics*, pages 60–83. World Scientific, 2008.

[BEK+14] Alex Bloemendal, László Erdős, Antti Knowles, Horng-Tzer Yau, Jun Yin, et al. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electronic Journal of Probability*, 19, 2014.

[Bel03] Serban Teodor Belinschi. The atoms of the free multiplicative convolution of two probability distributions. *Integral Equations and Operator Theory*, 46(4) :377–386, 2003.

[BG10] Florent Benaych-Georges. On a surprising relation between the Marčenko-Pastur law, rectangular and square free convolutions. In *Annales de l'IHP Probabilités et statistiques*, volume 46, pages 644–652, 2010.

[BHZ12] Zhidong Bai, Jiang Hu, and Wang Zhou. Convergence rates to the Marčenko-Pastur type distribution. *Stochastic Processes and their Applications*, 122(1) :68–92, 2012.

[BM20] Marwa Banna and Tobias Mai. Hölder continuity of cumulative distribution functions for noncommutative polynomials un-

der finite free fisher information. *Journal of Functional Analysis*, 279(8) :108710, 2020.

[BP21] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26(none), 2021.

[BP22] Lucas Benigni and Sandrine Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks, 2022.

[BS10] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[BV93] Hari Bercovici and Dan Voiculescu. Free convolution of measures with unbounded support. *Indiana University Mathematics Journal*, 42(3) :733–773, 1993.

[BZ08] Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442, 2008.

[Cap17] Mireille Capitaine. Deformed ensembles, polynomials in random matrices and free probability theory. 2017.

[Cho22] Clément Chouard. Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. *arXiv preprint arXiv :2211.13044*, 2022.

[Cho23] Clément Chouard. Deterministic equivalent of the conjugate kernel matrix associated to artificial neural networks. *arXiv preprint arXiv :2306.05850*, 2023.

[Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4) :303–314, 1989.

[ESY09] László Erdős, Benjamin Schlein, and Horng-Tzer Yau. Local semicircle law and complete delocalization for Wigner random matrices. *Communications in Mathematical Physics*, 287(2) :641–655, 2009.

[FOBS06] Reza Rashidi Far, Tamer Oraby, Wlodzimierz Bryc, and Roland Speicher. Spectra of large block matrices, 2006.

[FW20] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020.

[GH03] Jeffrey S Geronimo and Theodore P Hill. Necessary and sufficient condition that the limit of Stieltjes transforms is a Stieltjes transform. *Journal of Approximation Theory*, 121(1) :54–60, 2003.

[GLR+22] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

[GT10] Friedrich Götze and Aleksandr Nikolaevich Tikhomirov. The rate of convergence of spectra of sample covariance matrices. *Theory of Probability & Its Applications*, 54(1) :129–140, 2010.

[HLN07] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3) :875–930, 2007.

[HW71] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3) :1079–1083, 1971.

[Joh90] Charles R Johnson. *Matrix theory and applications*, volume 40. American Mathematical Soc., 1990.

[JSS+20] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR, 2020.

[KY17] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1) :257–352, 2017.

[LC18] Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. 2018.

[LC20] Cosme Louart and Romain Couillet. Concentration of solutions to random equations with concentration of measure hypotheses. *arXiv preprint arXiv :2010.09877*, 2020.

[LC21] Cosme Louart and Romain Couillet. Spectral properties of sample covariance matrices arising from random matrices with independent non identically distributed columns. *arXiv preprint arXiv :2109.02644*, 2021.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[MP67] V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4) :457–483, 1967.

[MS17] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[NM20] Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33 :11961–11972, 2020.

[Noi21] Nathan Noiry. Spectral measures of spiked random matrices. *Journal of Theoretical Probability*, 34(2) :923–952, 2021.

[PS21] Vanessa Piccolo and Dominik Schröder. Analysis of one-hidden-layer neural networks via the resolvent method. *Advances in Neural Information Processing Systems*, 34 :5225–5235, 2021.

[PW17] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems 30*. 2017.

[PY14] Natesh S Pillai and Jun Yin. Universality of covariance matrices. 2014.

[Pé19] S. Péché. A note on the Pennington-Worah distribution. *Electronic Communications in Probability*, 24(none) :1 – 7, 2019.

[Ros58] Frank Rosenblatt. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6) :386, 1958.

[San59] Giovanni Sansone. *Orthogonal functions*, volume 9. Interscience Publishers, 1959.

[SB95] J.W. Silverstein and Z.D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2) :175–192, 1995.

[SCDL23] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning, 2023.

[Sil86] Jack W Silverstein. Eigenvalues and eigenvectors of large dimensional sample covariance matrices. *Contemporary Mathematics*, 50 :153–159, 1986.

[Sil89] Jack W Silverstein. On the eigenvectors of large dimensional sample covariance matrices. *Journal of multivariate analysis*, 30(1) :1–16, 1989.

[Sil95] J.W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2) :331–339, 1995.

[Tal95] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1) :73–205, 1995.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[VW15] Van Vu and Ke Wang. Random weighted projections, random quadratic forms and random eigenvectors. *Random Structures & Algorithms*, 47(4) :792–821, 2015.

[Wig57] Eugene Paul Wigner. *Statistical properties of real symmetric matrices with many dimensions*. Princeton University, 1957.

[Wis28] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2) :32–52, 1928.

[WZ21] Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *arXiv preprint arXiv :2109.09304*, 2021.

[WZ23] Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks, 2023.

[Yan20] Fan Yang. Linear spectral statistics of eigenvectors of anisotropic sample covariance matrices, 2020.

[YS19] Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv :1907.10599*, 2019.

Matrices aléatoires de covariance et réseaux de neurones artificiels

Résumé : Cette thèse est consacrée à l'étude asymptotique de certains modèles de matrices aléatoires de covariance empirique, en utilisant une approche analytique et notamment en étudiant leur matrice résolvante. Elle se décompose en deux parties.

Dans un premier temps nous nous intéressons aux matrices de covariance issues de matrices avec une structure de dépendance partielle en colonnes. Nous fournissons un équivalent déterministe explicite pour la matrice résolvante de tels modèles, quantitatif en la dimension et le paramètre spectral de la résolvante. Ce résultat de type loi locale permet d'étudier les statistiques spectrales fines de ces matrices, et notamment de préciser la vitesse de convergence des mesures spectrales empiriques en distance de Kolmogorov.

Dans un second temps nous étudions le modèle du noyau conjugué, qui reproduit le comportement d'un réseau de neurones artificiels à propagation avant lors de sa phase d'initialisation. Ce modèle se distingue des modèles habituellement considérés en matrices aléatoires par l'application d'une fonction réelle entrée par entrée. Nous utilisons les résultats généraux obtenus précédemment pour obtenir un équivalent déterministe quantitatif pour la résolvante du noyau conjugué. Cet équivalent permet de mieux comprendre les propriétés spectrales de ce modèle, et aussi de donner un sens à un phénomène d'universalité appelé principe d'équivalence gaussienne dans le domaine de l'apprentissage machine.

Sample covariance matrices arising in artificial neural networks

Abstract : This thesis is devoted to the asymptotic study of some sample covariance random matrices models, by means of analytical methods and in particular by studying their resolvent. It subdivides into two parts.

First, we focus on sample covariance matrices derived from matrices with a partial dependence structure in columns. We provide an explicit deterministic equivalent for the resolvent of such models, quantitative in both the dimension and the spectral parameter. This local law-type result enables us to study fine spectral statistics of such matrices, and in particular to specify the speed of convergence of the empirical spectral measures in Kolmogorov distance.

Secondly, we study the conjugate kernel model, which modelizes a feed-forward artificial neural network at initialization. This model differs from classical random matrix models in that a real function is applied entry-wise on the matrices. We use the aforementioned results to obtain a quantitative deterministic equivalent for the conjugate kernel resolvent. This equivalent provides a better understanding of the spectral properties of this model, and it highlights a universality phenomenon also known as Gaussian equivalence principle in machine learning.