Learning with a linear loss function : excess risk and estimation bounds for ERM and minimax MOM estimators, with applications.

Lucie Neirac

INSTITUT POLYTECHNIQUE DE PARIS

ENSAE

IP PARIS

NNT : 2023IPPAG012

Thèse de doctorat

Learning with a linear loss function: excess risk and estimation bounds for ERM and minimax MOM estimators, with applications.

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à École nationale de la statistique et de l'administration économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques fondamentales

Thèse présentée et soutenue à Palaiseau, le 11 décembre 2023, par

LUCIE NEIRAC

Composition du Jury :

Christophe Giraud Professeur, Institut de Mathématiques d'Orsay	Président du jury
Nicolas Verzelen Chargé de recherches, INRAE	Rapporteur
Yohann De Castro Professeur, Institut Camille Jordan	Rapporteur
Alexandre d'Aspremont Directeur de recherche, Ecole Normale Supérieure	Examinateur
Guillaume Lecué Professeur, ENSAE	Directeur de thèse
Matthieu Lerasle Professeur, ENSAE	Co-directeur de thèse

Papa,

Imaginer ta joie si tu avais tenu ce manuscrit entre tes mains est une douleur. La scène tourne en boucle dans ma tête. Ce qui me terrasse, c'est de penser qu'elle a peut-être tourné en boucle dans la tienne. Je vois ton sourire, je sens ta main secouer doucement mon épaule, j'entends ta voix, surtout :

"Ah, ma star !"

Rien de plus. L'amour qui pétille dans tes yeux dit tout le reste.

Je n'ai pas été assez forte pour t'offrir ce bonheur. Je ne me le pardonnerai jamais.

Je t'aime et te dédie ce manuscrit.

Mon premier est ce que qu'une personne née dans les années 90 (disons plutôt 85) dit à une autre personne lorsqu'elle la croise pour la deuxième fois de la journée.
Avec l'amour de ma deuxième, la vie vous fait à l'aube une promesse qu'elle ne tient jamais.
A l'origine, mon dernier est le résultat de la réaction endothermique entre du calcaire et de l'argile qui, mélangés à de l'eau, font prise et permettent d'agglomérer entre eux des sables et des granulats.

Avec mon tout, on peut réer sciemment.

First but not least, je te remercie Guillaume. Si j'ai la chance d'être en train d'écrire ces remerciements aujourd'hui (flûte, voilà que la solution est dévoilée), c'est en premier lieu grâce à toi. Merci de m'avoir proposé ce sujet, alors même que je n'avais pas été une élève de compressed sensing très assidue, et d'avoir accepté d'embarquer avec moi dans cette aventure. Les choses semblent toujours limpides lorsque tu les expliques, c'est un réel moteur. Merci pour ton calme et ta patience que la houle dans laquelle j'ai parfois égaré notre barque n'a pas su ébranler. Je suis très heureuse de la rive sur laquelle on accoste aujourd'hui.

Merci Matthieu d'avoir co-encadré ma thèse. Je garderai le souvenir de tes discours d'introduction toujours chargés d'humour, qui rendent le travail plus doux et agréable.

Un immense merci à vous, Nicolas et Yohann, qui avez accepté de rapporter ma thèse. Merci d'avoir pris le temps de la lire avec tant de minutie, et merci pour vos remarques constructives et avisées. Grâce à vous, je crois que je me dirige vers ma soutenance un peu plus sereinement. Je tiens également à vous remercier grandement, Christophe et Alexandre, pour avoir accepté de faire partie de mon jury.

Stéphane et Mihai, je garde de très bons souvenirs de notre collaboration, notamment des moments passés à Londres à tirer des plans sur le signed clustering. Vous êtes les co-auteurs de mon tout premier papier de recherche, cela nous lie ! Merci Stéphane pour ta gentillesse, et merci de m'avoir accueillie à Lyon, où tu as pris le temps de m'expliquer tant de choses.

Merci Sacha pour ta bienveillance. J'ai beaucoup apprécié d'avoir participé à plusieurs reprises aux rencontres de Statistique Mathématique. Merci aux professeurs de statistique de l'Ensae, Pierre, Arnak, Nicolas.

Merci au pamplemousse, fruit à pépins qui, sous son masque où l'amertume se dispute à l'acidité, cache une splendide douceur, et sans lequel je n'aurais pas eu l'opportunité de commencer cette thèse[1].

Je remercie toutes celles et ceux qui ont fait partie de mon quotidien pendant les trois années que j'ai passées à l'Ensae. L'équipe de la direction des études, Corentin, Laurent, Claude,

[1] Des années plus tard, je regrettai son existence lorsque, assise dans la cantine sans fenêtres de la depp, je devai contempler, quotidiennement et de très longues minutes durant, Laure en décortiquant une moitié avec grand soin avant de la savourer lent-goureusement. Mais ça, c'était après.

Rosalinda, et puis les assistants d'enseignement, Arthur (merci pour le clic droit et même le ctrl+k, on n'arrête pas le progrès), Jérôme, Jérémy, Christophe, Morgane, Jules, Fabien et Nicolas. Merci Anne pour tes petits poissons. Merci aux doctorants[2] de l'Ensae, PEC, Lionel, Gabriel, Avo, Geoffrey, Nicolas, Suzanne, Solène, Boris, Flore. Merci Badr, pour ta gentillesse et tous ces bons moments passés à discuter. Et encore désolée pour ce pois chiche que ma maladresse a fait choir sur tes Timberland flambant neuves, au Cirm. Merci Gautier de m'avoir incitée à aborder mon début de thèse avec tranquilité. Je pense que ça a joué. Merci Amir d'avoir pris de mes nouvelles.

Merci aux Jarrets, vous éclairez ma vie. Quand je pense à ce jour de septembre 2015 où j'ai poussé pour la première fois la porte du café La rencontre - qui porte si bien son nom - je me dis que c'est fou à quel point dans la vie, on peut parfois tout perdre, mais parfois tout braquer, tout rafler, tout gagner. Christian, Hélène, Aurélia, Emmanuelle, Solène, Sandra, Ophélie, fatche, je vous dois tant. Tant de rires, de délires, de plaisir d'écrire, et plein d'autres rimes encore (en ire, mais pas que), sans lesquels la traversée de ces années aurait été très compliquée. Merci Aurélien, Stéphanie, Léa d'être venu.e.s pimenter notre jolie bande.
Vous êtes les co-auteurs que je porte dans mon coeur.

Merci à Yeboutcheva, Edith, Fanny, Jeanne, Anna et Gisèle. Me glisser dans votre peau le temps de quelques mois a été une indispensable soupape. Merci à Domitille aussi. Toutes les deux on ne fait que commencer, mais je nous sens déjà bien parties.

Laure, la vie passe son temps à semer des choses sur nos chemins, et il s'avère que ce sont parfois de très belles surprises. Merci pour la joie de vivre que tu transportes partout avec toi comme si c'était quelque-chose de normal. Ça m'a beaucoup aidée.
Merci Hugo d'avoir fait de la 307 LA pièce où il fait bon vivre, celle où les confidences se cachent sous les éclats de rire. N'en déplaise aux défenseurs de l'ordre, nos entropies se sont trouvées, bel, et surtout bien.
Merci à tous les d'jeun's de B2-2, et à Aïcha aussi. Terminer ma thèse en tant qu'activité extra-scolaire n'aurait sans doute pas été possible en dehors du climat de franche camaraderie qui règne dans notre couloir. A B2-2, le lino au sol n'est pas seul à être rose.
Merci Olivier, d'avoir fait du Scrabble bien plus qu'un jeu de mots, et d'avoir accepté que le silence ne soit pas un oubli.

Merci aux P'tits Charbos de faire comme si j'avais toujours été là. Il faut dire que je vous ai pas mal simplifié la tâche : il y a moins de lettres dans Lucie que dans Kunami, c'est donc plus facile à broder sur une serviette. Albane et François, merci pour votre bienveillance et votre générosité débordante. Liselotte et Régis, merci d'être aussi inspirants - notamment parce-qu'entre deux quintes de rire, il faut savoir inspirer franchement. Aldée et Colas, merci de m'avoir permis de découvrir la vie grenobloise depuis l'intérieur, bien que je n'aie pas forcément la bonne dégaine. Violaine et Théophane, merci de n'avoir jamais oublié un seul de mes anniversaires. Ghislain, merci d'avoir mis mon pied dans l'étrier de la hacke. Ma vie ne serait pas la même sans Ableton.

Thomas, merci de m'avoir pendue par les pieds ce jour de juillet 2011. Peut-être qu'un petit revival serait de circonstance[3] ? Merci de continuer à nous trouver attachantes malgré notre bizarrerie sans doute grandissante. Ludo, merci pour tes idées jamais à court et ton farouche optimisme en l'avenir. Grâce à toi, je peux raconter que j'ai chauffé un poids lourd à travers la France avec mon beauf et ça, ça en jette.

[2] ou devrais-je plutôt dire aujourd'hui maîtres de conférence, chargés de recherche, professeurs émérites, j'en passe, et des meilleures

[3] Cela dit, avec l'âge mon estomac est devenu considérablement plus sensible à tout ce qui est inclinaison de plus de 45 degrés, donc pas sûre que ce soit là une bonne idée.

Au milieu des années
que cette thèse a duré
mon chemin cabossé
soudain s'est éclairé.
Changée en zia Lu,
tout est devenu fou :
Grâce à toi mon Léon
V'là mon coeur en fusion.
—
Un 6 janvier, tu m'as choisie,
Tu m'as choisie et j'ai dit oui,
Nous deux c'est parti pour la vie.

Gaspard, ta bouille a été un véritable rempart contre le blizzard lors de la dernière ligne droite qui a mené à ce manuscrit. Merci pour tes grands sourires qui me rappellent que la vie est belle dans les moment où j'ai tendance à l'oublier.

Merci ma petite Mam, toi en qui s'est logé tout le courage du monde, et que tu distribues à ceux que tu aimes : c'est notre force. Je suis tellement heureuse que tu puisses assister à ma soutenance. Te savoir à l'autre bout du fil me donnera un pep' du tonnerre ! Merci Jean-Marc de m'avoir donné la fibre violonistique, sans quoi j'aurais peut-être opté pour la trompette (voire même la flûte à bec) et serais aujourd'hui une toute autre personne.

Maman, Papa, Camille, Zoé, je vous aime. Vous m'avez tout donné, et égoïstement, j'ai tout pris. Votre amour, votre humour, votre force, qui m'ont portée jusqu'ici. Maman, merci d'avoir toujours semé un pincée de magie, un zeste de surprise, une larme de folie sur chacun de tes gestes. Ça (ça), c'est vraiment toi. Tu m'as appris que l'amour n'a pas de limite, et qu'en cela il permet de les franchir toutes. Papa, tu m'as appris que sans la teindre d'originalité, il manque à la vie sa saveur. Qu'il est malavisé de se robustifier face aux outliers : ils sont la perle rare. Merci d'être parti sans éteindre la lumière. Camille, on dit qu'on ne choisit pas sa famille, pourtant j'ai toujours vécu dans l'impression que tu m'avais choisie : c'est qu'il m'est impossible de concevoir que le hasard puisse si bien faire les choses. Tu m'as appris que quand on veut, on peut, et que quand on peut, on doit. Tu as tracé ma route, je l'ai suivie sans doute. Zoé, être céleste tombé d'une perséide. Difficile encore une fois de penser au hasard. Laisse moi te dire que c'est un comble, pour quelqu'un qui abhorre l'eau gazeuse, d'être si pétillante. Tu m'as appris que l'humour désarme l'obscurité. Et que la force n'a nul besoin de montagne pour se cacher. Les soeurs, io sono la Lucie, voi siete le luci.

Séverin. Peut-on voir les étoiles jusqu'à la magie ? Je balaie le temps, ce truc. Je sais bien qu'on s'est toujours connus, mais c'est quand même vachement mieux depuis qu'on s'est rencontrés.

"J'vous jure, faites des MOM."
Ma mère
(mais elle n'a jamais précisé l'orthographe)

Résumé

La détection de communautés sur des graphes, la récupération de phase, le clustering signé, la synchronisation angulaire, le problème de la coupe maximale, la sparse PCA, ou encore le single index model, sont des problèmes classiques dans le domaine de l'apprentissage statistique. Au premier abord, ces problèmes semblent très dissemblables, impliquant différents types de données et poursuivant des objectifs distincts. Cependant, la littérature récente révèle qu'ils partagent un point commun : ils peuvent tous être formulés sous la forme de problèmes d'optimisation semi-définie positive (SDP). En utilisant cette modélisation, il devient possible de les aborder du point de vue classique du machine learning, en se basant sur la minimisation du risque empirique (ERM) et en utilisant la fonction de perte la plus élémentaire: la fonction de perte linéaire. Cela ouvre la voie à l'exploitation de la vaste littérature liée à la minimisation du risque, permettant ainsi d'obtenir des bornes d'estimation et de développer des algorithmes pour résoudre ces problèmes. L'objectif de cette thèse est de présenter une méthodologie unifiée pour obtenir les propriétés statistiques de procédures classiques en machine learning basées sur la fonction de perte linéaire. Cela s'applique notamment aux procédures SDP, que nous considérons comme des procédures ERM. L'adoption d'un "point de vue machine learning" nous permet d'aller plus loin en introduisant d'autres estimateurs performants pour relever deux défis majeurs en apprentissage statistique : la parcimonie et la robustesse face à la contamination adversaire et aux données à distribution à queue lourde. Nous abordons le problème des données parcimonieuses en proposant une version régularisée de l'estimateur ERM. Ensuite, nous nous attaquons au problème de la robustesse en introduisant un estimateur basé sur le principe de la "Médiane des Moyennes" (MOM), que nous nommons l'estimateur minmax MOM. Cet estimateur permet de faire face au problème de la robustesse et peut être utilisé avec n'importe quelle fonction de perte, y compris la fonction de perte linéaire. Nous présentons également une version régularisée de l'estimateur minmax MOM. Pour chacun de ces estimateurs, nous sommes en mesure de fournir un "excès de risque" ainsi que des bornes d'estimation, en utilisant deux outils clés : les points fixes de complexité locale et les équations de courbure de la fonction d'excès de risque. Afin d'illustrer la pertinence de notre approche, nous appliquons notre méthodologie à cinq problèmes classiques en machine learning, pour lesquels nous améliorons l'état de l'art.

keywords: Programmation Semi-Définie, Minimisation du Risque Empirique, Sparsité, Statistique robuste.

Abstract

Community detection, phase recovery, signed clustering, angular group synchronization, Max-cut, sparse PCA, the single index model, and the list goes on, are all classical topics within the field of machine learning and statistics. At first glance, they are pretty different problems with different types of data and different goals. However, the literature of recent years shows that they do have one thing in common: they all are amenable to Semi-Definite Programming (SDP). And because they are amenable to SDP, we can go further and recast them in the classical machine learning framework of risk minimization, and this with the simplest possible loss function: the *linear loss function*. This, in turn, opens up the opportunity to leverage the vast literature related to risk minimization to derive excess risk and estimation bounds as well as algorithms to unravel these problems. The aim of this work is to propose a unified methodology to obtain statistical properties of classical machine learning procedures based on the linear loss function, which corresponds, for example, to the case of SDP procedures that we look as ERM procedures. Embracing a machine learning view point allows us to go into greater depth and introduce other estimators which are effective in handling two key challenges within statistical learning: sparsity, and robustness to adversarial contamination and heavy-tailed data. We attack the structural learning problem by proposing a regularized version of the ERM estimator. We then turn to the robustness problem and introduce an estimator based on the median of means (MOM) principle, which we call the minmax MOM estimator. This latter estimator addresses the problem of robustness and can be constructed whatever the loss function, including the linear loss function. We also present a regularized version of the minmax MOM estimator. For each of those estimators we are able to provide excess risk and estimation bounds, which are derived from two key tools: local complexity fixed points and curvature equations of the excess risk function. To illustrate the relevance of our approach, we apply our methodology to five classical problems within the frame of statistical learning, for which we improve the state-of-the-art results.

keywords: Semi-Definite Programming, Empirical Risk Minimization, Sparsity, Robust statistics.

Contents

Chapter 1

Introduction

La détection de communautés sur des graphes, la récupération de phase, le clestering signé, la synchronisation angulaire, le problème de la coupe maximale, la sparse PCA parcimonieuse ou encore le modèle à indice simple, sont tous des sujets classiques en machine learning et en statistiques. À première vue, ce sont des problèmes sensiblement différents, avec différents types de données et des objectifs distincts. Cependant, tout comme de nombreux autres problèmes, il a récemment été montré qu'ils pouvaient s'exprimer sous la forme de problèmes d'optimisation semi-définie positive (SDP). La SDP est une classe de problèmes d'optimisation convexe généralisant la programmation linéaire aux problèmes linéaires sur des matrices semi-définies positives (Todd, 2001), (Wolkowicz, Saigal, & Vandenberghe, 2012), (Boyd & Vandenberghe, 2004), et qui s'est avérée être un outil performant dans l'approche computationnelle de problèmes difficiles en machine learning, en optimisation combinatoire, en optimisation polynomiale, en data mining, en statistiques en grande dimension ou encore dans la formalisation de solutions numériques d'équations aux dérivées partielles. Afin s'en savoir un peu plus, nous nous plongeons, dans la suite de cette section, dans les fondements de l'optimisation SDP.

1.1 Contexte historique des estimateurs SDP

L'optimisation SDP est une classe de problèmes d'optimisation dont fait partie la programmation linéaire, et peut être définie comme l'ensemble des problèmes d'optimisation sur des ensembles de matrices symétriques (resp. hermitiennes) semi-définies positives, dont la fonction de perte est linéaire et les contraintes sont affines, c'est-à-dire l'ensemble des problèmes d'optimisation de la forme

$$\max_{Z \succeq 0} \Big(\langle A, Z \rangle : \langle B_j, Z \rangle = b_j \text{ pour } j = 1, \dots, m \Big), \tag{1.1}$$

où les matrices $A, B_1, \dots, B_m$ sont fixées. Les SDP sont des problèmes de programmation convexe qui peuvent être résolus en temps polynomial lorsque l'ensemble de contraintes est compact, et elles jouent un rôle primordial dans la résolution d'un grand nombre de problèmes convexes et non convexes, où elles apparaissent souvent comme une relaxation convexe du problème initial (Anjos & Lasserre, 2011).

1.1.1 Les débuts historiques

La première utilisation de la programmation semi-définie positive en statistiques remonte à Scobey and Kabe (1978) et Fletcher (1981). La même année, Shapiro fait usage d'optimisation SDP pour résoudre un problème d'analyse factorielle (Shapiro, 1982). L'étude des propriétés mathématiques des estimateurs SDP a ensuite pris de l'ampleur avec l'introduction des inégalités matricielles linéaires (LMI) et de leurs nombreuses applications en théorie du contrôle, en identification de systèmes et en traitement du signal. Le livre de Boyd, El Ghaoui, Feron, and Balakrishnan (1994) est la référence standard pour ce type de résultats, principalement obtenus dans les années 90.

1.1.2 La relaxation SDP du Max-Cut par Goemans-Williamson et son héritage

Un tournant notable est franchi avec la publication de Goemans and Williamson (1995), où l'optimisation SDP permet d'obtenir un ration d'approximation de 0.87 au problème du Max-Cut, qui est réputé comme étant NP-difficile. Le problème du Max-Cut est un problème de clustering qui, partant d'un graphe G, consiste à trouver une partition $S \cup S^c$ de l'ensemble V de ses nœuds telle que la somme des poids des arêtes entre S et S^c soit maximale. Dans Goemans and Williamson (1995), les auteurs abordent ce problème combinatoire difficile en utilisant le méthode désormais connue sous le nom de *relaxation SDP de Goemans-Williamson*, et utilisent la factorisation de Choleski de la solution optimale de cette SDP pour produire un schéma aléatoire atteignant la borne de 0.87 en espérance. De plus, ce problème peut être considéré comme une première utilisation du Laplacien d'un graphe pour obtenir un bi-clustering optimal, et constitue ainsi le premier chapitre d'une longue et fructueuse relation entre le clustering, les plongements les Laplaciens de graphes. Les estimateurs SDP ont permis d'approcher la solution d'autres problèmes combinatoires difficiles, notamment le problème de coloration de graphes (Karger, Motwani, & Sudan, 1998), et le problème de satisfaction de contraintes (Goemans & Williamson, 1994, 1995). Ces résultats ont ensuite été synthétisés dans Goemans (1997), Lemaréchal, Nemirovskii, and Nesterov (1995) et Wolkowicz (1999). Le schéma aléatoire introduit par Goemans et Williamson a ensuite été amélioré pour étudier des problèmes quadratiques quadratiquement contraints (QCQP) plus généraux, notamment dans Nesterov (1997) et Zhang (2000), et développé davantage dans He, Luo, Nie, and Zhang (2008). De nombreuses applications au traitement du signal sont abordées dans Olsson, Eriksson, and Kahl (2007) et Ma (2010); une implémentation spécifique à complexité réduite sous la forme d'un problème de minimisation des valeurs propres et son application à la récupération et au débruitage des moindres carrés binaires est présentée dans Chrétien and Corset (2009).

1.1.3 Relaxation des problèmes d'apprentissage automatique et d'estimation statistique en haute dimension

Les applications de la SDP aux problèmes liés à l'apprentissage automatique sont plus récentes et ont probablement commencé avec la relaxation SDP de K-means dans Peng and Xia (2005) et Peng and Wei (2007a), puis dans Ames (2014). Cette approche a ensuite été améliorée en utilisant une analyse statistique plus fine par Royer (2017) et Giraud and Verzelen (2018). Des méthodes similaires ont également été appliquées à la détection de communautés dans Hajek, Wu, and Xu (2016) ou Abbe, Bandeira, and Hall (2015), et pour la reconstitution partielle, dans Guédon and Vershynin (2016). Cette dernière approche a également été réutilisée via la technique du noyau pour le clustering de nuages de points dans Chrétien, Dombry, and Faivre (2016). Une autre utilisation de la SDP en machine learning est l'utilisation extensive des estimateurs des moindres carrés pénalisés par la norme nucléaire comme substitut à la pénalisation par le rang dans les problèmes de recouvrement de matrices de rang faible, tels que la complétion de matrice dans les systèmes de recommandation, le compressed sensing de matrices, le traitement du langage naturel et la tomographie d'état quantique; ces sujets sont étudiés dans Davenport and Romberg (2016). Des liens avec le design de chaînes de Markov à convergence rapide ont également été montrés dans Sun, Boyd, Xiao, and Diaconis (2006).

Dans une autre direction, A. Singer et ses collaborateurs ont récemment promu l'utilisation de la relaxation SDP pour l'estimation sous invariance de groupe, un domaine florissant trouvant de nombreuses applications (A. S. Bandeira, Charikar, Singer, & Zhu, 2014; Singer, 2011). Des relaxations SDP ont également été mises en œuvre dans Cucuringu (2015) dans le contexte de la synchronisation sur $\mathbb{Z}_2$ dans des réseaux multiplex signés avec contraintes, et dans Cucuringu (2016) dans le cadre du ranking à partir de comparaisons par paires inconsistentes et incomplètes, où une relaxation basée sur la SDP de la synchronisation angulaire sur SO(2)

permet l'état de l'art de la littérature du ranking. La récupération de phase à l'aide de la SDP a été étudiée par exemple dans Waldspurger, d'Aspremont, and Mallat (2015) et Demanet and Hand (2014). Une extension au regroupement multipartite basé sur la SDP a ensuite été proposée dans Karger et al. (1998). D'autres applications importantes de la SDP comprennent la théorie de l'information (Lovász, 1979), l'estimation dans les réseaux électriques (Lavaei & Low, 2011), la tomographie quantique (Gross, Liu, Flammia, Becker, & Eisert, 2010; Mazziotti, 2011) et l'optimisation polynomiale via des relaxations d'estimateurs des moindres carrés (Blekherman, Parrilo, & Thomas, 2012; Lasserre, 2015). Cette dernière méthode a été récemment appliquée à des problèmes statistiques dans De Castro, Gamboa, Henrion, Hess, Lasserre, et al. (2019), Hopkins (2018) et De Castro, Gamboa, Henrion, and Lasserre (2017). L'extension au domaine des nombres complexes, avec $\langle \cdot, \cdot \rangle$ désignant le produit scalaire hermitien, a été à ce jour moins étudiée mais présente de nombreuses applications intéressantes et produit des algorithmes efficaces (Gilbert & Josz, 2017; Goemans & Williamson, 1995).

1.2 Les SDPs en tant que minimiseurs du risque empirique

Le point d'intérêt ici réside dans le fait que bon nombre des problèmes mentionnés ci-dessus peuvent être reformulés dans le cadre, classique en machine learning, de la minimisation du risque (Vapnik, 2000). Il est donc possible de tirer parti de la vaste littérature liée à la minimisation du risque pour en obtenir l'excès de risque, des bornes d'estimation ainsi que des algorithmes efficaces pour traîter ces problèmes. De plus, il apparaît que ces différents problèmes entrent tous dans le cadre défini par une fonction de perte très simple, sans doute la plus simple : la fonction de perte linéaire. Cette observation est est la clef de voûte de ce travail : plusieurs estimateurs introduits récemment dans certains des problèmes cités précédemment sont en fait des minimiseurs du risque empirique (ERM) pour des fonctions de perte linéaires. Ils peuvent donc être analysés en utilisant toute la machinerie (Boucheron, Lugosi, & Massart, 2013a; Koltchinskii, 2011a; Vapnik, 2000) développée au cours des quarante dernières années pour l'ERM dans ce cadre très spécifique de la fonction de perte linéaire.

1.2.1 Rappels sur les ERM

Soient H un espace de Hilbert et $F \subset H$ un ensemble de paramètres. On considère une fonction $\ell : H \times H \to \mathbb{R}$ telle que $\ell(Z, X) := \ell_Z(X)$ quantifie l'erreur commise lors de l'estimation de Z par X. Soit P une distribution de probabilité sur H. Le risque d'un paramètre $Z \in F$ est défini par $P\ell_Z := \mathbb{E}_{X\sim P}[\ell_Z(X)]$. Lorsqu'il existe, nous nous intéressons à la valeur de Z qui minimise ce risque, que nous appelons un oracle :

$$Z^* \in \operatorname*{argmin}_{Z\in F} \ell_Z(X).$$

Pour estimer Z^*, supposons que nous disposions d'un certain nombre de points $X_1, \ldots, X_N \in H$ distribués selon la distribution P. Une idée naturelle est alors d'estimer chaque $P\ell_Z$ par sa valeur empirique $P_N\ell_Z := (1/N)\sum_{i=1}^N \ell_Z(X_i)$. Cela nous donne l'estimateur suivant pour l'oracle Z^* :

$$\hat{Z}_{ERM} \in \operatorname*{argmin}_{Z\in F} P_N\ell_Z \tag{1.2}$$

qui n'est autre que l'illustre estimateur du minimiseur du risque empirique (ERM).

1.2.2 Cadre mathématique

Nous exposons ici le formalisme mathématique qui sous-tend notre approche. Le cadre général dans lequel nous nous plaçons est le suivant. Soient A une matrice aléatoire dans $\mathbb{R}^{d\times d}$ et

$\mathcal{C} \subset \mathbb{R}^{d\times d}$ un ensemble de contraintes (plus tard, nous examinerons également le cas des nombres complexes). L'objet que nous souhaitons estimer, par exemple le vecteur d'appartenance à une communauté dans le cas de la détection de communautés, est lié à un *oracle* défini par

$$Z^* \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle \mathbb{E}A, Z\rangle, \tag{1.3}$$

où $\langle A, B\rangle = \mathrm{Tr}(AB^\top) = \sum A_{ij}B_{ij}$ lorsque $A, B \in \mathbb{R}^{d\times d}$, et où la contrainte $\mathcal{C}$ est de la forme $\mathcal{C} = \{Z \in \mathbb{R}^{n\times n} : Z \succeq 0, \langle Z, B_j\rangle = b_j, j = 1, \dots, m\}$, où $B_1, \dots, B_m \in \mathbb{R}^{n\times n}$. Nous nous intéressons à Z^* car son estimation nous permettra ensuite de récupérer l'objet qui nous importe vraiment (par exemple, en considérant un vecteur singulier associé à la plus grande valeur singulière de Z^*). À cette fin, nous considérons l'estimateur naturel suivant de Z^* donné par

$$\hat{Z} \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle A, Z\rangle, \tag{1.4}$$

qui est simplement obtenu en remplaçant la quantité non observée $\mathbb{E}A$ par l'observation A. Il s'agit donc d'un problème d'optimisation semi-définie positive (SDP), selon la définition que nous en avons donnée précédemment.

Notez que dans de nombreuses situations, Z^* n'est pas l'objet que nous souhaitons estimer, mais il existe une relation directe entre Z^* et cet objet. Par exemple, considérons le problème de détection de communautés sur un graphe, où l'objectif est de récupérer le vecteur d'appartenance aux communautés $\beta^* \in \{-1, 1\}^d$ de d nœuds. Ici, lorsque la conatrainte $\mathcal{C}$ est bien choisie, il existe une relation étroite entre Z^* et β^*, donnée par $Z^* = \beta^*(\beta^*)^\top$. Nous avons donc besoin d'une étape finale pour estimer β^* à partir de $\hat{Z}$, par exemple en considérant un vecteur propre dominant $\hat{\beta}$ de $\hat{Z}$, puis en utilisant le théorème 'sin(Θ)' de Davis-Kahan (C. Davis & Kahan, 1970; Yu, Wang, & Samworth, 2015) afin de contrôler l'estimation de β^* par $\hat{\beta}$ à partir de celle de Z^* par $\hat{Z}$.

Le point de vue que nous adoptons consiste à voir $\hat{Z}$ comme une procédure de minimisation du risque empirique (ERM) construite sur une seule observation A, où la fonction de perte est la fonction linéaire $Z \in \mathcal{C} \to \ell_Z(A) = -\langle A, Z\rangle$, et l'oracle Z^* est celui qui minimise effectivement la fonction de risque $Z \in \mathcal{C} \to \mathbb{E}\ell_Z(A)$ sur $\mathcal{C}$. Notre méthodologie met au jour une approche générale caractérisée par deux éléments cruciaux : la courbure locale de la fonction de risque et le calcul d'un point fixe de complexité local.

1.3 Exemples issus de la littérature

La raison d'être notre intérêt pour l'étude des propriétés statistiques des estimateurs ERM avec fonction de perte linéaire réside dans le fait que la littérature du machine learning regorge de problèmes qui peuvent être modélisés sous cette forme. Nous présentons ici une liste de ces problèmes, qui n'est en aucun cas exhaustive. Pour chacun des problèmes présentés, nous explicitons la valeur de la matrice A et de la variable aléatoire X qui nous permettent de nous conformer au cadre défini dans la Section 2.2.2.

1.3.1 Détection de communautés

Les estimateurs ERM avec une fonction de perte linéaire peuvent être utilisés pour traiter le problème de la détection de communautés sur des graphes. Afin d'illustrer cela, nous considérons ici le cadre du Stochastic Block Model (SBM), tel qu'il est présenté dans Guédon and Vershynin (2016) ou Fei and Chen (2019b), et que nous rappelons ci-dessous. Considérons un ensemble de sommets $V = \{1, \cdots, d\}$, et supposons qu'il est partitionné en K communautés $\mathcal{C}_1, \cdots, \mathcal{C}_K$ de tailles arbitraires $|\mathcal{C}_1|, \cdots, |\mathcal{C}_K|$. Pour toute paire de nœuds $i, j \in V$, nous notons

$i \sim j$ lorsque i et j appartiennent à une même communauté, et $i \not\sim j$ si i et j dans le cas contraire. Pour chaque paire (i,j) de nœuds de V, nous traçons une arête entre i et j avec une probabilité fixe p_{ij} indépendamment des autres arêtes. Nous supposons qu'existent des nombres p et q tels que $0 < q < p < 1$, de sorte que $p_{ij} > p$ si $i \sim j$ et $i \neq j$, $p_{ij} = 1$ si $i = j$ et $p_{ij} < q$ sinon. Nous notons $A = (A_{i,j})_{1 \leq i,j, \leq d}$ la matrice (symétrique) d'adjacence observée telle que, pour tout $1 \leq i \leq j \leq d$, A_{ij} est distribué selon une loi de Bernoulli de paramètre p_{ij}. La structure de communauté d'un tel graphe est capturée par la matrice d'affiliation $\bar{Z} \in \mathbb{R}^{d \times d}$, définie par $\bar{Z}_{ij} = 1$ si $i \sim j$, et $\bar{Z}_{ij} = 0$ sinon. L'objectif est de reconstruire $\bar{Z}$ à partir de l'observation de A. Le Lemme 7.1 de Guédon and Vershynin (2016) montre que la matrice d'affiliation $\bar{Z}$ est donnée par l'oracle suivant :

$$Z^* \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle \mathbb{E}[A], Z \rangle, \qquad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d}, Z \succeq 0, Z \geq 0, \mathrm{d}iag(Z) \preceq I_d, \sum_{i,j=1}^{d} Z_{ij} \leq \lambda \right\}$$

où $\lambda = \sum_{i,j=1}^{d} \bar{Z}_{ij} = \sum_{k=1}^{K} |\mathcal{C}_k|^2$ représente le nombre d'éléments non nuls dans la matrice d'affiliation $\bar{Z}$. Seule la matrice A étant observée, les auteurs considèrent l'estimateur suivant pour Z^* :

$$\hat{Z} \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle A, Z \rangle.$$

D'un 'point de vue machine learning', cet estimateur est un ERM pour la fonction de perte linéaire $Z \to \ell_Z(A) := -\langle A, Z \rangle$, construit à partir d'une seule observation de la matrice aléatoire A. Cette observation constitue notre point de départ dans l'analyse du problème de détection de communautés avec tous les outils développés dans la section 3.

1.3.2 Clustering de variables

Bunea, Giraud, Luo, Royer, and Verzelen (2018) utilisent des estimateurs SDP pour résoudre le problème du clustering de variables. Ce problème consiste à regrouper en clusters les composantes similaires d'un vecteur $X \in \mathbb{R}^d$, c'est-à-dire à trouver une partition $G = \{G_1, \ldots, G_K\}$ de $\{1, \ldots, d\}$ qui sépare les composants de X. Pour ce faire, les auteurs observent N copies indépendantes $X_1, \ldots, X_N$ de X et se placent dans le cas où la matrice de covariance Σ de X suit un modèle par bloc. Afin de décrire ce modèle, nous définissons la matrice d'affiliation $Q \in \mathbb{R}^{p \times K}$ associée à une partition G telle que $Q_{ak} = \mathbb{1}_{\{a \in G_k\}}$. On dit alors que Σ suit un G-modèle exact de covariance par bloc lorsqu'elle se décompose sous la forme $\Sigma = QCQ^\top + \Gamma$, où C est une matrice symétrique $K \times K$ et Γ est une matrice diagonale $d \times d$. Pour une partition donnée G, nous introduisons également la matrice d'affiliation correspondante $Z^* \in \mathbb{R}^{d \times d}$ définie par $Z^*_{ij} = |G_k|^{-1} \mathbb{1}_{\{i \text{ et } j \text{ appartiennent au même groupe } G_k\}}$. Il existe une correspondance bijective entre les partitions G et leurs matrices d'affiliation, de sorte que rechercher G est équivalent à rechercher Z^*. En utilisant l'algorithme des K-means et une relaxation de celui-ci donnée dans Peng and Wei (2007b), les auteurs montrent que la meilleure partition pour les X_i peut être estimée par celle correspondant à la matrice d'appartenance suivante :

$$\hat{Z} \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle A, Z \rangle, \qquad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d} : Z \succeq 0, Z \geq 0, \sum_{j} Z_{ij} = 1 \forall i, \mathrm{Tr}\,(Z) = K \right\}$$

où $A := (1/N) \sum_{i=1}^{N} X_i X_i^\top$ est la matrice de covariance empirique des X_i. Dans le cas non-bruité, nous aurions $Z^* \in \operatorname{argmax}_{Z \in \mathcal{C}} \langle \mathbb{E}[A], Z \rangle$. Encore une fois, de notre point de vue, l'estimateur $\hat{Z}$ peut être considéré comme un ERM pour la fonction de perte linéaire $Z \to \ell_Z(A) := -\langle A, Z \rangle$, construit à partir de l'observation de A, et peut donc analysé selon notre méthodologie.

1.3.3 Synchronisation angulaire

Le problème de la synchronisation angulaire consiste à estimer d angles inconnus $\theta_1, \cdots, \theta_d$ (à un delta de décalage global près), étant donné un sous-ensemble bruité de leurs différences mutuelles $\delta_{ij} = \theta_i - \theta_j$. Ce problème est étudié dans A. Bandeira, Boumal, and Singer (2016). Les auteurs se placent dans la cas où ils observent $d(d-1)/2$ mesures de la forme suivante :

$$a_{ij} = e^{\iota\delta_{ij}} + \epsilon_{ij}, \qquad \text{pour } 1 \leq i < j \leq d.$$

Ils supposent que les $(\epsilon_{ij})_{i<j}$ sont des variables complexes gaussiennes $i.i.d.$ Le problème peut être reformulé sous la forme suivante :

$$A = X\bar{X}^\top + \sigma W$$

où $X \in \mathbb{C}^d$ défini par $X_i = e^{\iota\theta_i}$, W étant une matrice de Wigner complexe et $\sigma > 0$ étant la variance du bruit. L'objectif est alors de reconstruire le vecteur $x^* = (e^{\iota\theta_i})_{i=1}^d$, dont l'estimateur du maximum de vraisemblance est, à une rotation globale de ses coordonnées près, la solution unique du problème de maximisation suivant :

$$\operatorname*{argmax}_{x\in\mathcal{E}} \left\{ \bar{x}^\top \mathbb{E}A\, x \right\}, \quad \text{où} \quad \mathcal{E} := \left\{ x \in \mathbb{C}^d : |x_i| = 1 \text{ pour tous les } i = 1 \dots d \right\}$$

En remarquant que $\mathcal{E} = \{Z \in \mathbb{H}_n : Z \succeq 0, \operatorname{diag}(Z) = \mathbb{1}_d, \operatorname{rank}(Z) = 1\}$, les auteurs construisent la formulation SDP suivante du problème, après avoir supprimé la contrainte de rang :

$$Z^* \in \operatorname*{argmin}_{Z\in\mathcal{C}} -\langle \mathbb{E}[A], Z\rangle, \quad \text{où} \quad \mathcal{C} := \{Z \in \mathbb{H}_n : Z \succeq 0, \operatorname{diag}(Z) = \mathbb{1}_d\} \tag{1.5}$$

Ils montrent que dans ce cadre, x^* peut être obtenu à partir de Z^* en tant que son vecteur propre dominant unitaire. Cependant, $\mathbb{E}[A]$ n'est pas connu et seulement observé à travers A. Z^* est alors approché par l'estimateur naturel suivant :

$$\hat{Z} \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle A, Z\rangle, \quad \text{où} \quad \mathcal{C} := \{Z \in \mathbb{H}_n : Z \succeq 0, \operatorname{diag}(Z) = \mathbb{1}_d\}$$

Cela constitue donc un autre exemple d'estimateur ERM basé sur l'observation de la matrice A et la fonction de perte linéaire $Z \to \ell_Z(A) = -\langle A, Z\rangle$.

1.3.4 Max-Cut

Dans Hong, Lee, and Wei (2021), les auteurs proposent un estimateur SDP pour traiter le problème du Max-Cut. Ce problème est classique en théorie des graphes, et consiste à considérer un graphe consitué des sommets $V := \{1, \dots, d\}$ et des arêtes $E \subset V \times V$, et à déterminer une partition $S \cup \bar{S} = V$ de ses sommets telle que le nombre d'arêtes connectant un sommet de S à un sommet de $\bar{S}$ soit maximal parmi toutes les partitions possibles. La plupart du temps, seule une version partielle ou bruitée $A \in \{0,1\}^{d\times d}$ de la matrice d'adjacence du graphe est effectivement observée. On suppose alors, en général, que la valeur réelle de la matrice d'adjacence est égale à l'espérance $\mathbb{E}A$ de l'observation A. La matrice A est donc considérée comme notre seule donnée, à partir de laquelle nous souhaitons établir une partition optimale S^* du graphe original. Le choix d'une partition S est équivalent au choix de $x \in \{-1,1\}^N$, et il est démontré dans Goemans and Williamson (1995), par un argument de 'lifting', qu'une partition optimale peut être obtenue en considérant le vecteur propre dominant d'une solution du problème d'optimisation suivant :

$$Z^* \in \operatorname*{argmin}_{Z\in\mathbb{R}^{d\times d}} \left(\langle \mathbb{E}[A], Z\rangle : Z \succeq 0, Z_{ii} = 1\ \forall i, \operatorname{rank}(Z) = 1 \right).$$

Ensuite, en utilisant une relaxation SDP en supprimant la contrainte de rang, nous retrouvons les procédures classiques de relaxation SDP pour le MAX-CUT introduites par Goemans et Williamson :

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle A, Z \rangle, \quad \text{où} \quad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d}, Z \succeq 0, Z_{ii} = 1 \ \forall i \right\}$$

Il s'agit effectivement d'un estimateur ERM basé sur l'observation de A et de la fonction de perte linéaire $Z \to \ell_Z(A) := \langle A, Z \rangle$ sur un ensemble convexe.

1.3.5 RÉCUPÉRATION DE PHASE

Le problème précédent est proche de celui de la récupération de phase, qui vise à reconstituer un vecteur $x \in \mathbb{C}^d$ à partir de l'observation bruitée de l'amplitude de N mesures linéaires aléatoires : $X = |Bx| \in \mathbb{R}^N$, où $B \in \mathbb{C}^{N \times d}$ est une matrice aléatoire. Dans WALDSPURGER, D'ASPREMONT, and MALLAT (2013), les auteurs utilisent une stratégie qui consiste à séparer la phase de l'amplitude et à optimiser uniquement les valeurs des variables de phase. Dans le cas non-bruité, iels écrivent $x = B^+ \text{diag(X)u}$, où $u \in \mathbb{C}^N$ est un vecteur de phase et $B^+ \in \mathbb{C}^{N \times N}$ est le pseudo-inverse de B. Dans ce formalisme, iels montrent que trouver $x \in \mathbb{C}^d$ tel que $|Bx| = X$ revient à résoudre le problème suivant :

$$z^* \in \underset{z \in \mathcal{E}}{\operatorname{argmin}} \langle \mathbb{E}[A], z\bar{z}^\top \rangle, \qquad \mathcal{E} := \left\{ z \in \mathbb{C}^N : |z_i| = 1, \forall i \in [N] \right\}$$

où $A := \left(XX^\top\right) \circ \left(I_N - BB^+\right)$ (et $\circ$ est le multiplicateur matriciel composante-par-composante). En écrivant $Z = z\bar{z}^\top$, ce problème est équivalent au suivant :

$$\min_{Z \in \mathcal{C}} \left(\langle \mathbb{E}[A], Z \rangle, Z \succeq 0, Z_{ii} = 1 \forall i, \operatorname{rank}(Z) = 1 \right)$$

qui est classiquement relaxé en abandonnant la contrainte de rang pour aboutir à l'oracle suivant :

$$Z^* \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle \mathbb{E}[A], Z \rangle, \text{ for } \mathcal{C} := \left\{ Z \in \mathbb{R}^{N \times N} : Z \succeq 0, Z_{ii} = 1 \ \forall i \in [N] \right\}$$

La valeur optimale de z^* est ensuite obtenue comme le vecteur propre dominant de l'oracle Z^*. Un estimateur de Z^* à partir de l'observation de A est alors :

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle A, Z \rangle,$$

ce qui est un problème d'optimisation SDP que nous considérons comme un ERM pour la fonction de perte linéaire $Z \to \ell_Z(A) := \langle A, Z \rangle$.

1.3.6 PCA PARCIMONIEUSE

L'analyse en composantes principales (ACP) est l'un des algorithmes de réduction de dimension les plus fondamentaux ainsi que l'un des outils de visualisation de données les plus utilisés. Étant donné un ensemble de données $X_1, \ldots, X_N \in \mathbb{R}^d$, l'objectif de l'ACP est de trouver les composantes principales $(e_1, \ldots, e_d)$ dans $\mathbb{R}^d$ (ou simplement les k premières $(e_1, \ldots, e_k)$) telles que la plus grande variance par une certaine projection scalaire des X_i sur les e_i soit atteinte sur la première coordonnée e_1, la deuxième plus grande variance sur la deuxième coordonnée e_2, et ainsi de suite. L'ACP a été largement étudiée et peut être efficacement exécutée par des algorithmes de SVD tronqués. Cependant, les composantes sont un mélange de caractéristiques qui peuvent être sans signification lorsque ces caractéristiques sont de natures différentes, ou lorsque l'on se trouve dans la cadre de la grande dimension - c'est-à-dire lorsque $d >> N$. C'est là que l'ACP parcimonieuse intervient : l'objectif est de rechercher des composantes principales qui ne sont un mélange que d'un petit nombre k de caractéristiques.

Soient donc $X_1, \dots, X_N \in \mathbb{R}^d$ des points de données aléatoires supposés indépendants et distribués selon une distribution P. Le problème consistant à trouver la première composante principale éparse a été résolu dans (Johnstone & Lu, 2009a; Johnstone & Lu, 2009b), où il est montré qu'elle peut être exprimée comme l'oracle suivant :

$$v^* \in \underset{\|v\|_2=1,\, \|v\|_0 \leq k}{\operatorname{argmax}} \|\mathbb{E}[A]v\|_2,$$

où $\mathbb{E}[A] := \mathbb{E}_{X \sim P}[XX^\top]$ est la matrice de covariance des X_i. Il peut être démontré, et nous le ferons plus en détail ultérieurement, que v^* tel que défini peut être obtenu comme le vecteur propre principal de la solution du problème d'optimisation

$$Z^\star \in \max \left(\langle \mathbb{E}[A], Z \rangle, \quad Z \succeq 0, \operatorname{Tr}(Z) = 1, \operatorname{card}(Z) \leq k^2 \right),$$

qui est classiquement relaxé en supprimant la contrainte de cardinalité pour mener à l'oracle suivant :

$$Z^\star \in \underset{Z \in \mathcal{C}}{\operatorname{argmax}} \langle \mathbb{E}[A], Z \rangle, \quad \text{où} \quad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d} \, Z \succeq 0, \operatorname{Tr}(Z) = 1 \right\}.$$

Puisque nous n'observons pas $\mathbb{E}[A]$ mais seulement les X_i, nous considérons l'estimateur suivant pour l'oracle :

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmax}} \langle A, Z \rangle,$$

où $A := (1/N) \sum_{i=1}^N X_i X_i^\top$, qui est alors un estimateur ERM basé sur la fonction de perte linéaire $Z \to \ell_Z(A) = -\langle A, Z \rangle$.

1.3.7 Le modèle monovarié parcimonieux

La situation est la suivante : nous considérons un modèle semi-paramétrique où une sortie $Y \in \mathbb{R}$ est générée à partir d'une entrée $X \in \mathbb{R}^d$ via une fonction de 'liaison' de la manière suivante :

$$Y = f\left(\langle X, \beta^* \rangle \right) + \epsilon.$$

Ici, $\beta^* \in \mathbb{R}^d$ est supposé être un vecteur unitaire k-parcimonieux, $f : \mathbb{R} \to \mathbb{R}$ est une fonction mesurable univariée inconnue, et ϵ est un bruit généralement supposé indépendant de l'entrée. Les composantes de X sont supposées être *i.i.d.* avec une fonction de densité donnée p_0. La fonction de densité jointe de X est alors $p = \otimes_{t=1}^d p_0$ par rapport à la mesure de Lebesgue. Nous introduisons la fonction de score univariée $s : x \in \mathbb{R} \to \mathbb{R}$ par $s(x) = -p_0'(x)/p_0(x)$, définie pour p_0-presque tous les $x \in \mathbb{R}$. À partir de cela, la première et la deuxième fonction de score associées à p sont définies comme suit :

$$S(X) = \left(s(X_t) \right)_{1 \leq t \leq d} \in \mathbb{R}^d, \text{et}$$
$$T(X) := S(X)S(X)^\top - \operatorname{diag}\left(\left(s'(X_t) \right)_{1 \leq t \leq d} \right)$$

Le travail de Yang, Balasubramanian, and Liu (2018) se concentre sur ce problème. Les auteurs montrent que le paramètre β^* peut être obtenu comme le vecteur propre principal de l'oracle suivant :

$$Z^* \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} -\langle \mathbb{E}[A], Z \rangle, \quad \mathcal{C} := \left\{ 0 \preceq Z \preceq I_d \, \text{et} \, \operatorname{Tr}(Z) = 1 \right\}$$

où $A := YT(X)$. Cet oracle peut alors être estimé comme suit :

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \left(-\langle A, Z \rangle + \lambda \|Z\|_1 \right)$$

ce qui est donc un estimateur ERM régularisé basé sur l'observation A et la fonction de perte linéaire $Z \to \ell_Z(A) = -\langle A, Z \rangle$.

1.3.8 Apprentissage d'une métrique de distance

Les estimateurs SDP peuvent également être utilisés dans l'apprentissage de métriques de distance, comme cela est fait dans Xing, Ng, Jordan, and Russell (2002). L'apprentissage des distances est particulièrement important, car le choix d'une métrique correctement adaptée à l'espace des données est crucial pour permettre l'acuité de nombreux algorithmes d'apprentissage, en particulier en clustering, où il est essentiel de prendre pleinement en compte les relations existant entre les données. Considérons un ensemble de points $(X_i)_{i=1\dots N} \in \mathbb{R}^d$ que nous observons partiellement ou de manière bruitée. Maintenant, considérons le problème consistant à apprendre des X_i une métrique de distance de la forme

$$d_Z(X,Y) = \sqrt{\mathrm{Tr}((\mathrm{X}-\mathrm{Y})(\mathrm{X}-\mathrm{Y})^\top \mathrm{Z})},$$

où $Z \succeq 0$ est une matrice semi-définie positive. Notons que, puisque l'on a

$$\mathrm{Tr}((\mathrm{X}-\mathrm{Y})(\mathrm{X}-\mathrm{Y})^\top \mathrm{Z}) = \|Z^{1/2}(X-Y)\|_2^2,$$

apprendre une telle métrique revient à effectuer une mise à l'échelle des données en remplaçant chaque point X par $Z^{1/2}X$, puis à appliquer la métrique euclidienne standard aux données ainsi mises à l'échelle. Supposons maintenant que nous souhaitons que les X_i soient aussi proches que possible les uns des autres pour cette métrique. Cela nous amène à résoudre le problème $\min_{Z\succeq 0} \sum_{i,j=1}^N d_Z(X_i,X_j)^2$. Ce dernier problème étant trivialement résolu par 0, supposons que nous ayons connaissance de M points $(Y_i)_{i=1\dots M}$, distincts des X_i, pour lesquels nous souhaitons que la contrainte suivante soit vérifiée : $\sum_{i,j=1}^M d_Z(Y_i,Y_j) \geq 1$. Cela empêche que la fonction d ne réduise l'ensemble de données à un seul point. Définissons alors $A := \sum_{i,j=1}^N \left(X_i - X_j\right)\left(X_i - X_j\right)^\top$. Dans le cas non-bruité, la matrice Z^* que nous recherchons peut alors être obtenue comme une solution du problème suivant :

$$Z^* \in \operatorname*{argmin}_{Z\in\mathcal{C}} \langle \mathbb{E}[A], Z\rangle, \qquad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \sum_{i,j=1}^M \left\langle \left(Y_i - Y_j\right)\left(Y_i - Y_j\right)^\top, Z\right\rangle^{1/2} \geq 1 \right\},$$

où l'on peut montrer que l'ensemble $\mathcal{C}$ est convexe (voir l'Annexe 7.1). En pratique, l'observation A est une version bruitée de $\mathbb{E}[A]$, et nous remplaçons donc $\mathbb{E}[A]$ par A pour obtenir un estimateur de Z^* :

$$\hat{Z} \in \operatorname*{argmin}_{Z\in\mathcal{C}} \langle A, Z\rangle$$

qui est à nouveau un estimateur ERM pour la fonction de perte linéaire $Z \to \ell_Z(A) = \langle A, Z\rangle$, construit à partir d'une observation de la matrice aléatoire A.

1.3.9 Transport optimal bruité

Soient $\mathcal{X} = (x_1,\dots,x_N)$ et $\mathcal{Y} = (y_1,\dots,y_N)$ deux ensembles de points dans $\mathbb{R}^d$. Le problème du transport optimal quadratique (ou problème d'assignation quadratique) est défini par la distance de Wasserstein W_2 comme suit :

$$W_2^2(\mathcal{X},\mathcal{Y}) = \min_{\tau\in\mathfrak{S}_N} \sum_{i=1}^N \|x_i - y_{\tau(i)}\|^2, \tag{1.6}$$

où $\mathfrak{S}_N$ est l'ensemble de toutes les permutations de $[N]$. Trouver une solution à (2.6) est un problème classique en transport optimal qui peut être reformulé comme un problème matriciel :

$$Z^* \in \operatorname*{argmin}_{Z\in\mathcal{C}} \sum_{i,j} \|x_i - y_j\|_2^2 P_{ij},$$

où $\mathcal{C}$ est l'ensemble de toutes les matrices bi-stochastiques $N \times N$ (c'est-à-dire l'ensemble des matrices à composantes positives dont la somme est égale à 1 le long de chaque ligne et de chaque colonne). En effet, si τ^* est une solution optimale de (2.6), alors pour tout $i \in [N]$, $Z^*_{i\tau^*(i)} = 1$ et $Z^*_{ij} = 0$ lorsque $j \neq \tau^*(i)$.

Supposons maintenant que nous n'observions pas exactement les points de $\mathcal{X}$ et de $\mathcal{Y}$, mais que nous n'ayons accès qu'à une version bruitée de ces points : pour tout $i \in [N]$, $X_i = x_i + \sigma G_i$ et $Y_i = y_i + \sigma G'_i$ où $\sigma \geq 0$ et $(G_i, G'_i)_{i=1}^N$ sont des vecteurs gaussiens standard *i.i.d.* de $\mathbb{R}^d$. Le problème d'assignation quadratique pour ces deux ensembles bruités de points est une solution du problème suivant :

$$\hat{Z} \in \operatorname*{argmin}_{Z \in \mathcal{C}} \langle A, Z \rangle \text{ où } A = \left(\|X_i - Y_j\|_2^2 \right)_{1 \leq i,j \leq N},$$

et il peut être démontré que dans le cas non-bruité, $Z^* \in \operatorname{argmin}_{Z \in \mathcal{C}} \langle \mathbb{E}A, Z \rangle$. Le problème du transport optimal quadratique bruité vise à identifier une transition de phase nette, c'est-à-dire une valeur σ^* telle que 1) si $\sigma < \sigma^*$, alors $\hat{Z} = Z^*$ avec grande probabilité et 2) pour tous $\sigma > \sigma^*$, $\hat{Z} \neq Z^*$ avec une probabilité supérieure à $1/2$.

1.4 Contributions

Comme le montrent les exemples exposés ci-dessus, qui confortent nos déclarations précédentes, l'étude générale des fonctions de perte linéaires en machine learning présente un réel intérêt. L'objectif de cette thèse est de proposer une méthodologie unifiée pour obtenir les propriétés statistiques des procédures classiques en machine learning basées sur la fonction de perte linéaire, telles que les procédures SDP introduites ci-dessus que nous considérons désormais comme des procédures ERM. Le cadre général est celui défini précédemment dans la section 2.2.2. Cependant, certains problèmes reposent sur une structure particulière, telle que la parcimonie, et d'autres sont confrontés à l'écueil de la robustesse. Pour ces problèmes, un ERM n'est pas la bonne réponse. Adopter le point de vue du machine learning nous permet d'aller plus loin en introduisant d'autres estimateurs qui se trouvent être efficaces pour relever ces deux défis cruciaux en apprentissage statistique. Nous abordons le problème de l'apprentissage structurel en proposant une version régularisée de cet estimateur ERM, en ajoutant une fonction de régularisation à la fonction objectif dans (2.4). Ensuite, nous nous tournons vers le problème de la robustesse et introduisons un estimateur basé sur le principe de la médiane des moyennes (MOM), qui a été introduit dans Lecué and Lerasle (2020) et que nous appelons le minimax MOM. Cet estimateur permet de surmonter le problème de la robustesse et peut être construit quelle que soit la fonction de perte. En particulier, il s'adapte à notre configuration où l'on considère la fonction de perte linéaire. Nous montrons que les estimateurs ainsi obtenus sont robustes à la contamination des données ainsi qu'aux données à queue lourde. Tout comme pour l'estimateur ERM, nous présentons une version classique et une version régularisée de l'estimateur minimax MOM dans cette configuration.

Pour chacun des estimateurs précédemment évoqués, nous sommes en mesure de proposer des garanties statistiques lorsque $\mathbb{E}[A]$ n'est observée que partiellement à travers A. En particulier, notre approche conduit à de nouveaux taux de convergence non-asymptotiques ou à des propriétés de reconstruction exacte pour une large gamme d'estimateurs se conformant à notre cadre. Ensuite, afin de montrer la pertinence de notre approche, nous appliquons ces bornes générales à cinq problèmes : le clustering signé, la synchronisation angulaire, le problème du Max-Cut, l'ACP parcimonieuse et le modèle à indice unique. En utilisant notre approche, nous fournissons un excès de risque et des bornes d'estimation en guise de garanties statistiques et améliorons l'état de l'art pour ces cinq problèmes.

Community detection, phase recovery, signed clustering, angular group synchronization, Max-cut, sparse PCA, and the single index model are all classical topics in machine learning and statistics. At first glance, they are pretty different problems with different types of data and different goals. However, as well as many other problems, they have recently been shown to be amenable to Semi-Definite Programming (SDP). SDP is a class of convex optimization problems generalising linear programming to linear problems over semi-definite matrices (Todd, 2001), (Wolkowicz et al., 2012), (Boyd & Vandenberghe, 2004), and which was proved to be an important tool in the computational approach to difficult challenges in automatic control, combinatorial optimization, polynomial optimization, data mining, high-dimensional statistics and the numerical solution to partial differential equations. Let us make a dig into the foundations of SDP optimization.

2.1 Historical background of SDP estimators

SDP is a class of optimization problems which includes linear programming as a particular case, and can be written as the set of problems over symmetric (resp. Hermitian) positive semi-definite matrix variables, with linear cost function and affine constraints, i.e. optimization problems of the form

$$\max_{Z \succeq 0} \Big(\langle A, Z \rangle : \langle B_j, Z \rangle = b_j \text{ for } j = 1, \dots, m \Big), \tag{2.1}$$

where $A, B_1, \dots, B_m$ are given matrices. SDPs are convex programming problems which can be solved in polynomial time when the constraint set is compact and it plays a paramount role in a large number of convex and non-convex problems, for which they often appear as a convex relaxation (Anjos & Lasserre, 2011).

2.1.1 Early history

Early use of Semi-Definite programming in statistics can be traced back to Scobey and Kabe (1978) and Fletcher (1981). In the same year, Shapiro used SDP in factor analysis (Shapiro, 1982). The study of the mathematical properties of SDP then gained momentum with the introduction of Linear Matrix Inequalities (LMI) and their numerous applications in control theory, system identification and signal processing. The book of Boyd et al. (1994) is the standard reference of these type of results, mostly obtained in the 90's.

2.1.2 The Goemans-Williamson SDP relaxation of Max-Cut and its legacy

A notable turning point is the publication of Goemans and Williamson (1995), where SDP was shown to provide a 0.87 approximation to the NP-Hard problem known as Max-Cut. The Max-Cut problem is a clustering problem on graphs which consists in finding two complementary subsets S and S^c of nodes such that the sum of the weights of the edges between S and S^c is maximal. In Goemans and Williamson (1995), the authors approach this difficult combinatorial problem by using what is now known as the Goemans-Williamson *SDP*

relaxation, and use the Choleski factorization of the optimal solution to this SDP in order to produce a randomized scheme achieving the 0.87 bound in expectation. Moreover, this problem can be seen as a first instance where the Laplacian of a graph is employed in order to provide an optimal bi-clustering in a graph, and certainly represents the first chapter of a long and fruitful relationship between clustering, embedding and graph Laplacians. Other SDP schemes for approximating hard combinatorial problems are, to name a few, for the graph coloring problem (Karger et al., 1998), and the satisfiability problem (Goemans & Williamson, 1994, 1995). These results were later surveyed in Goemans (1997), Lemaréchal et al. (1995) and Wolkowicz (1999). The randomized scheme introduced by Goemans and Williamson was then further improved in order to study more general Quadratically Constrained Quadratic Programmes (QCQP) in various references, most notably Nesterov (1997), Zhang (2000) and further extended in He et al. (2008). Many applications to signal processing are discussed in Olsson et al. (2007) and Ma (2010); one specific reduced complexity implementation in the form of an eigenvalue minimization problem and its application to binary least-squares recovery and denoising is presented in Chrétien and Corset (2009).

2.1.3 Relaxation of machine learning and high-dimensional statistical estimation problems

Applications of SDP to problems related to machine learning is more recent and probably started with the SDP relaxation of K-means in Peng and Xia (2005) and Peng and Wei (2007a) and later in Ames (2014). This approach was then further improved using a refined statistical analysis by Royer (2017) and Giraud and Verzelen (2018). Similar methods have also been applied to community detection in Hajek et al. (2016) or Abbe et al. (2015), and for the weak recovery viewpoint, in Guédon and Vershynin (2016). This last approach was also re-used via the kernel trick for the point cloud clustering in Chrétien et al. (2016). Another incarnation of SDP in machine learning is the extensive use of nuclear norm-penalized least-square costs as a surrogate for rank-penalization in low-rank recovery problems such as matrix completion in recommender systems, matrix compressed sensing, natural language processing and quantum state tomography; these topics are surveyed in Davenport and Romberg (2016). Connections with the design of fast converging Markov-Chains were also exhibited in Sun et al. (2006).

In a different direction, A. Singer and collaborators have recently promoted the use of SDP relaxation for estimation under group invariance, an active area with many applications (A. S. Bandeira et al., 2014; Singer, 2011). SDP-based relaxations have also been considered in Cucuringu (2015) in the context of synchronization over $\mathbb{Z}_2$ in signed multiplex networks with constraints, and in Cucuringu (2016) in the setting of ranking from inconsistent and incomplete pairwise comparisons where an SDP-based relaxation of angular synchronization over SO(2) outperformed a suite of state-of-the-art algorithms from the ranking literature. Phase recovery using SDP was studied for example in Waldspurger et al. (2015) and Demanet and Hand (2014). An extension to multi-partite clustering based on SDP was then proposed in Karger et al. (1998). Other important applications of SDP are information theory (Lovász, 1979), estimation in power networks (Lavaei & Low, 2011), quantum tomography (Gross et al., 2010; Mazziotti, 2011) and polynomial optimization via Sums-of-squares relaxations (Blekherman et al., 2012; Lasserre, 2015). Sums of squares relaxations were recently applied to statistical problems in De Castro et al. (2019), Hopkins (2018) and De Castro et al. (2017). Extension to the field of complex numbers, with $\langle \cdot, \cdot \rangle$ denoting the Hermitian inner product, has been less extensively studied but has many interesting applications and comes with efficient algorithms (Gilbert & Josz, 2017; Goemans & Williamson, 1995).

2.2 SDPs as Empirical risk minimizers

The point of interest here lies in the fact that many of the problems mentioned above can be recast in the classical machine learning framework of risk minimization (Vapnik, 2000). It is therefore possible to leverage the vast literature related to risk minimization to derive excess risk and estimation bounds as well as algorithms to unravel them. It appears that the general framework that can encapsulate all these problems relies in fact on a simple loss function, maybe the simplest one: the linear loss function. This observation is the baseline of this work: several estimators introduced recently in some of the problems cited at the beginning are in fact empirical risk minimizers (ERM) for linear loss functions. They can therefore be analyzed using all the machinery (Boucheron et al., 2013a; Koltchinskii, 2011a; Vapnik, 2000) developed during the last forty years for ERM in this very specific framework of linear loss function.

2.2.1 Reminders on ERMs

Let H be a Hilbert space and $F \subset H$ be a set of parameters. Consider a function $\ell : H \times H \to \mathbb{R}$ such that $\ell(Z, X) := \ell_Z(X)$ quantifies the error made when estimating Z with X. Let P be a probability distribution on H. The risk of a parameter $Z \in F$ is quantified by $P\ell_Z := \mathbb{E}_{X\sim P}[\ell_Z(X)]$. When it exists, we are interested on the value of Z that minimizes this risk, which we call an oracle:

$$Z^* \in \operatorname*{argmin}_{Z\in F} P\ell_Z(X).$$

To estimate Z^*, assume we are given some data points $X_1, \dots, X_N \in H$ which are distributed according to the distribution P. A natural idea is then to estimate each $P\ell_Z$ by its empirical value $P_N\ell_Z := (1/N)\sum_{i=1}^N \ell_Z(X_i)$. This gives us the following estimator for the oracle Z^*:

$$\hat{Z}_{ERM} \in \operatorname*{argmin}_{Z\in F} P_N\ell_Z \tag{2.2}$$

which is the celebrated empirical risk minimizer (ERM) estimator.

2.2.2 Mathematical framework

Here, we outline the mathematical formalism that underpins our approach. The general framework in which we stand is as follows. Let A be a random matrix in $\mathbb{R}^{d\times d}$ and $\mathcal{C} \subset \mathbb{R}^{d\times d}$ be a constraint (we will also consider the complex numbers case later on). The object that we want to recover, for instance the community membership vector in community detection, is related to an *oracle* defined as

$$Z^* \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle \mathbb{E}A, Z\rangle, \tag{2.3}$$

where $\langle A, B\rangle = \mathrm{Tr}(AB^\top) = \sum A_{ij}B_{ij}$ when $A, B \in \mathbb{R}^{d\times d}$ and the constraint $\mathcal{C}$ is of the form $\mathcal{C} = \{Z \in \mathbb{R}^{n\times n} : Z \succeq 0, \langle Z, B_j\rangle = b_j, j = 1, \dots, m\}$, where $B_1, \dots, B_m \in \mathbb{R}^{n\times n}$. We would like to estimate Z^*, from which we can ultimately retrieve the object that really matters to us (for instance, by considering a singular vector associated to the largest singular value of Z^*). To that end, we consider the following natural estimator of Z^* given by

$$\hat{Z} \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle A, Z\rangle, \tag{2.4}$$

which is simply obtained by replacing the unobserved quantity $\mathbb{E}A$ by the observation A. This is then a semidefinite program (SDP), according to the definition we gave earlier.

Note that in many situations, Z^* is not the object we want to estimate, but there is a straightforward relation between Z^* and this object. For instance, consider the community detection

problem, where the goal is to recover the class community vector $\beta^* \in \{-1,1\}^d$ of d nodes. Here, when $\mathcal{C}$ is well chosen, there is a close relation between Z^* and β^*, given by $Z^* = \beta^*(\beta^*)^\top$. We therefore need a final step to estimate β^* from $\hat{Z}$, for instance by letting $\hat{\beta}$ denote a top eigenvector of $\hat{Z}$, and then using the Davis-Kahan 'sin(Θ)' Theorem (C. Davis & Kahan, 1970; Yu et al., 2015) to control the estimation of β^* by $\hat{\beta}$ from the one of Z^* by $\hat{Z}$.

The point of view we adopt is to see $\hat{Z}$ as an empirical risk minimization (ERM) procedure built on a single observation A, where the loss function is the linear one $Z \in \mathcal{C} \to \ell_Z(A) = -\langle A, Z\rangle$, and the oracle Z^* is indeed the one minimizing the risk function $Z \in \mathcal{C} \to \mathbb{E}\ell_Z(A)$ over $\mathcal{C}$. Our methodology uncovers a general approach characterized by two crucial points: the local curvature of the excess risk and the computation of a local complexity fixed point.

2.3 Examples from litterature

The rationale behind our interest in studying the statistical properties of ERM estimators with a linear loss functions lies in the fact that the litterature on statistical learning is teem with problems that can be modeled under this form. We provide here a list of those problems, which is by no means exhaustive. For each of the problems presented, we explain the value of the matrix A and the random variable X that enable us to enter the framework defined in Section 2.2.2.

2.3.1 Community detection

ERM estimators with a linear loss function can be used to handle the problem of community detection on graphs. Indeed, to present this point of view, we consider the setup of the Stochastic Block Model (SBM), as in Guédon and Vershynin (2016) or Fei and Chen (2019b), that we recall below. We consider a set of vertices $V = \{1, \cdots, d\}$, and assume it is partitioned into K communities $\mathcal{C}_1, \cdots, \mathcal{C}_K$ of arbitrary sizes $|\mathcal{C}_1|, \cdots, |\mathcal{C}_K|$. For any pair of nodes $i, j \in V$, we denote by $i \sim j$ when i and j belong to the same community, and by $i \not\sim j$ if i and j do not belong to the same community. For each pair (i,j) of nodes from V, we draw an edge between i and j with a fixed probability p_{ij} independently from the other edges. We assume that there exist numbers p and q satisfying $0 < q < p < 1$, such that $p_{ij} > p$ if $i \sim j$ and $i \neq j$, $p_{ij} = 1$ if $i = j$ and $p_{ij} < q$ otherwise. We denote by $A = (A_{i,j})_{1 \leq i,j, \leq d}$ the observed symmetric adjacency matrix, such that, for all $1 \leq i \leq j \leq d$, A_{ij} is distributed according to a Bernoulli of parameter p_{ij}. The community structure of such a graph is captured by the membership matrix $\bar{Z} \in \mathbb{R}^{d \times d}$, defined by $\bar{Z}_{ij} = 1$ if $i \sim j$, and $\bar{Z}_{ij} = 0$ otherwise. The objective is to reconstruct $\bar{Z}$ from the observation A. Lemma 7.1 of Guédon and Vershynin (2016) shows that the membership matrix $\bar{Z}$ is given by the following oracle:

$$Z^* \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle \mathbb{E}[A], Z\rangle, \qquad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d}, Z \succeq 0, Z \geq 0, \mathrm{d}iag(Z) \preceq I_d, \sum_{i,j=1}^{d} Z_{ij} \leq \lambda \right\}$$

where $\lambda = \sum_{i,j=1}^{d} \bar{Z}_{ij} = \sum_{k=1}^{K} |\mathcal{C}_k|^2$ denotes the number of nonzero elements in the membership matrix $\bar{Z}$. Since only the A matrix is observed, the authors consider the following estimator for Z^*:

$$\hat{Z} \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle A, Z\rangle.$$

From a machine learning perspective, this estimator is an ERM with the linear loss function $Z \to \ell_Z(A) := -\langle A, Z\rangle$, constructed from a single observation of the random matrix A. This observation is our starting point for the analysis of the community detection problem with all the tools developed in section 3.

2.3.2 Variable clustering

SDP estimators have been considered in Bunea et al. (2018) to solve the variable clustering problem. The problem is that of grouping into clusters similar components of a vector $X \in \mathbb{R}^d$, that is to find a partition $G = \{G_1, \dots, G_K\}$ of $\{1, \dots, d\}$ that separates the components of X. To that end, the authors observe N independent copies $X_1, \dots, X_N$ of X and place themselves in the case where the covariance matrix Σ of X follows a block model. To describe this model, we define the membership matrix $Q \in \mathbb{R}^{d \times K}$ associated with a partition G as $Q_{ak} = \mathbb{1}_{\{a \in G_k\}}$. Then, Σ is said to follow an exact G-block covariance model when it decomposes as $\Sigma = QCQ^\top + \Gamma$, where C is a symmetric $K \times K$ matrix and Γ is a diagonal $d \times d$ matrix. For a given partition G, we also introduce its corresponding partnership matrix $Z^* \in \mathbb{R}^{d \times d}$ defined by $Z^*_{ij} = |G_k|^{-1} \mathbb{1}_{\left\{ i \text{ and } j \text{ belong to the same group } G_k \right\}}$. There is a one-to-one correspondence between partitions G and their corresponding partnership matrices, so that looking for G is equivalent to looking for Z^*. Using the K-means algorithm and a relaxation of it given in Peng and Wei (2007b), the authors show that the best partition for the X_i's can be estimated with the one corresponding to the following partnership matrix:

$$\hat{Z} \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle A, Z \rangle, \qquad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d} : Z \succeq 0, Z \geq 0, \sum_j Z_{ij} = 1 \forall i, \operatorname{Tr}(Z) = K \right\}$$

where $A := (1/N) \sum_{i=1}^N X_i X_i^\top$ is the empirical covariance of the X_i's. In the noiseless case, we would have $Z^* \in \operatorname{argmax}_{Z \in \mathcal{C}} \langle \mathbb{E}[A], Z \rangle$. Again, from our perspective the estimator $\hat{Z}$ can be seen as an ERM with the linear loss function $Z \to \ell_Z(A) := -\langle A, Z \rangle$, constructed from the observation of A, and therefore analysed using our methodology.

2.3.3 Angular synchronization

The angular synchronization problem consists of estimating d unknown angles $\theta_1, \cdots, \theta_d$ (up to a global shift angle) given a noisy subset of their pairwise offsets $\delta_{ij} = \theta_i - \theta_j$. This problem is investigated in A. Bandeira et al. (2016). The authors consider that they observe $d(d-1)/2$ measurements of the following form:

$$a_{ij} = e^{\iota \delta_{ij}} + \epsilon_{ij}, \qquad \text{for } 1 \leq i < j \leq d.$$

They assume the $(\epsilon_{ij})_{i<j}$'s to be *i.i.d.* complex Gaussian variables. The problem can be rewritten under the following form:

$$A = X \bar{X}^\top + \sigma W$$

with $X \in \mathbb{C}^d$ defined by $X_i = e^{\iota \theta_i}$, W being a complex Wigner matrix and $\sigma > 0$ being the variance of the noise. The aim is then to reconstruct the vector $x^* = (e^{\iota \theta_i})_{i=1}^d$, whose maximum likelihood estimator is, up to a global rotation of its coordinates, the unique solution of the following maximization problem:

$$\operatorname*{argmax}_{x \in \mathcal{E}} \left\{ \bar{x}^\top \mathbb{E}A \, x \right\}, \quad \text{where} \quad \mathcal{E} := \left\{ x \in \mathbb{C}^d : |x_i| = 1 \text{ for all } i = 1 \dots d \right\}$$

By noticing that $\mathcal{E} = \{Z \in \mathbb{H}_n : Z \succeq 0, \operatorname{diag}(Z) = \mathbb{1}_d, \operatorname{rank}(Z) = 1\}$, they construct the following SDP formulation of the problem, after removing the rank constraint:

$$Z^* \in \operatorname*{argmin}_{Z \in \mathcal{C}} -\langle \mathbb{E}[A], Z \rangle, \quad \text{where} \quad \mathcal{C} := \{Z \in \mathbb{H}_n : Z \succeq 0, \operatorname{diag}(Z) = \mathbb{1}_d\} \tag{2.5}$$

They show that in this setting, x^* can be obtain from Z^* as its leading unit-length eigen vector. However, $\mathbb{E}[A]$ is not known and only observed through A, so the following constitutes a natural estimator for Z^*:

$$\hat{Z} \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle A, Z \rangle, \quad \text{where} \quad \mathcal{C} := \{Z \in \mathbb{H}_n : Z \succeq 0, \operatorname{diag}(Z) = \mathbb{1}_d\}$$

This is therefore another example of an ERM estimator based on the observation of the matrix A and the linear loss function $Z \to \ell_Z(A) = -\langle A, Z \rangle$.

2.3.4 Max-Cut

In Hong et al. (2021), the authors propose an SDP estimator to handle the Max-Cut problem. The Max-Cut problem is a classical graph theory problem, which consists of taking a graph with vertices $V := \{1, \dots, d\}$ and edges $E \subset V \times V$ and finding a partition $S \cup \bar{S} = V$ of vertices such that the number of edges connecting a vertex in S to a vertex in $\bar{S}$ is maximal among all possible partitions. Most of the time, we observe only $A \in \{0,1\}^{d \times d}$ a noisy or partial version of the adjacency matrix of the graph. Hence, the true adjacency matrix of the graph is not observed but it is usually assumed to be equal to the expectation $\mathbb{E}A$ of the observed one A. Hence, A is considered as our data and from this data, we wish to find an optimal partition S^* of the original graph. Choosing a partition S being equivalent to choosing $x \in \{-1,1\}^N$, it is shown in Goemans and Williamson (1995) via a lifting argument that an optimal partition is a first eigenvector of a solution to the following optimization problem:

$$Z^* \in \underset{Z \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left(\langle \mathbb{E}[A], Z \rangle : Z \succeq 0, Z_{ii} = 1 \ \forall i, \operatorname{rank}(Z) = 1 \right).$$

Then, using an SDP relaxation by removing the rank constraint, we recover the classical Max-Cut SDP relaxation procedures introduced by Goemans and Williamson:

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle A, Z \rangle, \quad \text{where} \quad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d}, Z \succeq 0, Z_{ii} = 1 \ \forall i \right\}$$

It is indeed an ERM procedure based on the observation of A and the linear loss function $Z \to \ell_Z(A) := \langle A, Z \rangle$ over a convex set.

2.3.5 Phase recovery

The former problem is close to the one of phase recovery, which aims at recovering a vector $x \in \mathbb{C}^d$ from the noisy observation of the amplitude of N random linear measurements: $X = |Bx| \in \mathbb{R}^N$, with $B \in \mathbb{C}^{N \times d}$ a random matrix. In Waldspurger et al. (2013), the authors use a strategy that involves separating phase from amplitude and optimizing only the values of the phase variables. In the noiseless case, they write $x = B^+ \text{diag(X)u}$, where $u \in \mathbb{C}^N$ is a phase vector and $B^+ \in \mathbb{C}^{N \times N}$ is the pseudo-inverse of B. In this format, they show that finding $x \in \mathbb{C}^d$ such that $|Bx| = X$ is equivalent to solving the following problem:

$$z^* \in \underset{z \in \mathcal{E}}{\operatorname{argmin}} \langle \mathbb{E}[A], z\bar{z}^\top \rangle, \qquad \mathcal{E} := \left\{ z \in \mathbb{C}^N : |z_i| = 1, \forall i \in [N] \right\}$$

where $A := \left(XX^\top\right) \circ \left(I_N - BB^+\right)$ (and $\circ$ is the component-wise matrix multiplicator). Writing $Z = z\bar{z}^\top$, this problem is equivalent to the following one:

$$\min_{Z \in \mathcal{C}} \left(\langle \mathbb{E}[A], Z \rangle, Z \succeq 0, Z_{ii} = 1 \forall i, \operatorname{rank}(Z) = 1 \right)$$

which is classically relaxed by droping the rank constraint to arrive at the following oracle:

$$Z^* \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle \mathbb{E}[A], Z \rangle, \text{ for } \mathcal{C} := \left\{ Z \in \mathbb{R}^{N \times N} : Z \succeq 0, Z_{ii} = 1 \ \forall i \in [N] \right\}$$

The optimal value of z^* is then obtained as the first eigenvector of the oracle Z^*. An estimator of Z^* from the observation of A is then:

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle A, Z \rangle$$

which is a SDP optimization problem that we see as an ERM with the linear loss function $Z \to \ell_Z(A) := \langle A, Z \rangle$.

2.3.6 Sparse PCA

Principal Components analysis (PCA) is one of the most fundamental dimension reduction algorithm as well as one of the most used data visualization tool. Given a data set $X_1, \dots, X_N \in \mathbb{R}^d$, the aim of PCA is to find principal components $(e_1, \dots, e_d)$ in $\mathbb{R}^d$ (or just the first k ones $(e_1, \dots, e_k)$) such that the highest variance by some scalar projection of the data on the e_i's is attained on the first coordinate e_1, the second highest variance on the second coordinate e_2, and so on. PCA has been widely studied and can be efficently performed via truncated SVD algorithms, however components are a mixture of features that can be meaningless when those features are of different nature, or when we are in a high-dimensional setting - that is when $d >> N$. This is where the sparse PCA is called for: the aim is to look for principal components which are a mixture of only a small number k of features.

Let then $X_1, \dots, X_N \in \mathbb{R}^d$ be some random data points which are supposed independent and distributed according to a distribution P. The problem of finding the first sparse principal component has been settled down in (Johnstone & Lu, 2009a; Johnstone & Lu, 2009b) by finding that it can be expressed as the following oracle:

$$v^* \in \underset{\|v\|_2=1,\, \|v\|_0 \leq k}{\operatorname{argmax}} \quad \|\mathbb{E}[A]v\|_2$$

where $\mathbb{E}[A] := \mathbb{E}_{X\sim P}[XX^\top]$ is the covariance matrix of the X_i's. It can be shown, and we will do so in more detail later, that v^* such defined can be obtain as the leading eigen vector of the solution of the optimization problem $Z^\star \in \max\left(\langle \mathbb{E}[A], Z\rangle, \quad Z \succeq 0, \operatorname{Tr}(Z) = 1, \operatorname{card}(Z) \leq k^2\right)$, which is classically relaxed by removing the cardinality constraint to lead to the following oracle:

$$Z^\star \in \underset{Z\in\mathcal{C}}{\operatorname{argmax}} \langle \mathbb{E}[A], Z\rangle, \quad \text{where} \quad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d\times d}\, Z \succeq 0, \operatorname{Tr}(Z) = 1 \right\}.$$

Since we do not observe $\mathbb{E}[A]$ but only the X_i's, we take the following estimator for the oracle:

$$\hat{Z} \in \underset{Z\in\mathcal{C}}{\operatorname{argmax}} \langle A, Z\rangle.$$

with $A := (1/N)\sum_{i=1}^N X_i X_i^\top$, which is then an ERM estimator based on the linear loss function $Z \to \ell_Z(A) = -\langle A, Z\rangle$.

2.3.7 The sparse Single Index Model

The situation is as follows: we consider a semi-parametric model where an output $Y \in \mathbb{R}$ is generated from an input $X \in \mathbb{R}^d$, via a 'link' function in the following way:

$$Y = f\left(\langle X, \beta^*\rangle\right) + \epsilon.$$

Here, $\beta^* \in \mathbb{R}^d$ is assumed to be a k-sparse unit vector, $f : \mathbb{R} \to \mathbb{R}$ is an unknown univariate measurable function and ϵ is a noise that is generally assumed to be independent of the input. The entries of X are assumed to be *i.i.d.* with a given density function p_0. The joint density function of X is then $p = \otimes_{t=1}^d p_0$ with respect to the Lebesgue measure. We define a univariate score function $s : x \in \mathbb{R} \to \mathbb{R}$ by $s(x) = -p_0'(x)/p_0(x)$, defined for p_0-almost all $x \in \mathbb{R}$. From this, the first and second score functions associated with p are defined as follows:

$$S(X) = \left(s(X_t)\right)_{1\leq t\leq d} \in \mathbb{R}^d, \text{and}$$
$$T(X) := S(X)S(X)^\top - \operatorname{diag}\left(\left(s'(X_t)\right)_{1\leq t\leq d}\right)$$

The work of Yang et al. (2018) focuses on this problem. The authors show that the parameter β^* can be obtain as the leading eigen vector of the following oracle:

$$Z^* \in \underset{Z\in\mathcal{C}}{\operatorname{argmin}} -\langle \mathbb{E}[A], Z\rangle, \quad \mathcal{C} := \left\{0 \preceq Z \preceq I_d \,\text{and}\, \operatorname{Tr}(Z) = 1\right\}$$

where $A := YT(X)$. This oracle can then be estimated as follows:

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \left(-\langle A, Z \rangle + \lambda \|Z\|_1 \right)$$

which takes the form of a regularized ERM estimator based on the observation A and the linear loss function $Z \to \ell_Z(A) = -\langle A, Z \rangle$.

2.3.8 Distance metric learning

SDP estimators can also be used in learning distance metrics, as it is done in Xing et al. (2002). Learning distances is particularly important, as the choice of a metric that is correctly adapted to the input space is crucial to the acuity of many learning algorithms, especially in clustering, where it is essential to take deep account of the relationships between the data. Let's consider a set of points $(X_i)_{i=1\dots N} \in \mathbb{R}^d$ that we observe partially or with noise. Now, consider the task of learning a distance metric of the form

$$d_Z(X, Y) = \sqrt{\mathrm{Tr}((\mathrm{X} - \mathrm{Y})(\mathrm{X} - \mathrm{Y})^\top \mathrm{Z})},$$

where $Z \succeq 0$ is positive semidefinite. We note that, since one has $\mathrm{Tr}((\mathrm{X} - \mathrm{Y})(\mathrm{X} - \mathrm{Y})^\top \mathrm{Z}) = \|Z^{1/2}(X - Y)\|_2^2$, learning such a distance metric amounts to finding a rescaling of data that replaces each point X with $Z^{1/2}X$ and applying the standard Euclidean metric to the rescaled data. Now, assume that we want the X_i's to be as close as possible to each other for this metric. This leads us to solve the problem $\min_{Z \succeq 0} \sum_{i,j=1}^N d_Z(X_i, X_j)^2$. Since this last problem is trivially solved by 0, suppose we have knowledge of M points $(Y_i)_{i=1\dots M}$, distinct from the X_i's, for which we want the following constraint to be verified: $\sum_{i,j=1}^M d_Z(Y_i, Y_j) \geq 1$. This prevent the situation where d_Z collapses the dataset into a single point. Let us then define $A := \sum_{i,j=1}^N (X_i - X_j)(X_i - X_j)^\top$. In the noiseless case, the matrix Z^* we are looking for can then be taken as a solution to the following problem:

$$Z^* \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle \mathbb{E}[A], Z \rangle, \qquad \mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d} : Z \succeq 0, \sum_{i,j=1}^M \langle (Y_i - Y_j)(Y_i - Y_j)^\top, Z \rangle^{1/2} \geq 1 \right\}$$

where one can show that the set $\mathcal{C}$ is convex (see Appendix 7.1). In practice, the observation A is a noisy version of $\mathbb{E}[A]$, so we just replace $\mathbb{E}[A]$ with A to get an estimator of Z^*:

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle A, Z \rangle$$

which is again an ERM estimator with the linear loss function $Z \to \ell_Z(A) = \langle A, Z \rangle$, constructed from an observation of the random matrix A.

2.3.9 Noisy optimal transport

Let $\mathcal{X} = (x_1, \dots, x_N)$ and $\mathcal{Y} = (y_1, \dots, y_N)$ be two clouds of points in $\mathbb{R}^d$. The quadratic optimal transport problem (or quadratic assignment problem) is defined by the W_2-Wasserstein distance

$$W_2^2(\mathcal{X}, \mathcal{Y}) = \min_{\tau \in \mathfrak{S}_N} \sum_{i=1}^N \|x_i - y_{\tau(i)}\|^2 \tag{2.6}$$

where $\mathfrak{S}_N$ is the set of all permutations of $[N]$. Finding a solution to (2.6) is a standard problem in optimal transport that can be lifted to the matrix problem

$$Z^* \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \sum_{i,j} \|x_i - y_j\|_2^2 Z_{ij}$$

where $\mathcal{C}$ is the set of all $N \times N$ bi-stochastic matrices (i.e. of matrices with non-negative entries summing to one along each rows and columns). Indeed, if τ^* denotes an optimal solution to (2.6) then for all $i \in [N]$, $Z^*_{i\tau^*(i)} = 1$ and $Z^*_{ij} = 0$ when $j \neq \tau^*(i)$.

Let us now assume that we do not observe exactly the points in $\mathcal{X}$ and $\mathcal{Y}$ but we only have access to a noisy version of these points: for all $i \in [N]$, $X_i = x_i + \sigma G_i$ and $Y_i = y_i + \sigma G'_i$ where $\sigma \geq 0$ and $(G_i, G'_i)_{i=1}^N$ are $2N$ *i.i.d.* standard Gaussian vectors in $\mathbb{R}^d$. The quadratic assignment problem for this two noisy cloud of points is a solution to the problem

$$\hat{Z} \in \operatorname*{argmin}_{Z \in \mathcal{C}} \langle A, Z \rangle \text{ where } A = \left(\|X_i - Y_j\|_2^2 \right)_{1 \leq i,j \leq N}$$

and it can be shown that in the free noise case, we have $Z^* \in \operatorname{argmin}_{Z \in \mathcal{C}} \langle \mathbb{E}A, Z \rangle$. The noisy quadratic optimal transport problem is to identify a sharp phase transition that is a σ^* such that 1) if $\sigma < \sigma^*$ then with high probability $\hat{Z} = Z^*$ and 2) for all $\sigma > \sigma^*$, with probability larger than $1/2$, $\hat{Z} \neq Z^*$.

2.4 Contributions

As substantiated by the examples we enumerated above, in line with our prior statements, there is a real interest in the general study of linear loss functions in machine learning. The aim of this thesis is to propose a unified methodology to obtain statistical properties of classical machine learning procedures based on the linear loss function such as the SDP procedures introduced above that we are now looking as ERM procedures. The general framework is the one previously defined in Section 2.2.2. However, some problems rely on some structure such as sparsity, and other ones are facing the problem of robustness. For these issues, ERM is not the right answer. Embracing a machine learning view point allows us to explore further and introduce other estimators which are effective in handling those two key challenges within statistical learning. We attack the structural learning problem by proposing a regularized version of this ERM estimator, adding a regularization function to the objective function in (2.4). Afterwards, we turn to the robustness problem and introduce an estimator based on the median of means (MOM) principle, which have been introduced in Lecué and Lerasle (2020) and that we call the minmax MOM. This latter estimator addresses the problem of robustness and can be constructed whatever the loss function. In particular, it fits our linear loss function setup. We show that the resulting estimators are robust to data contamination as well as to heavy-tailed data. As for ERMs, we present a classical and a regularized version of the minmax MOM estimator in this setup.

For each of those estimators we are able to propose statistical guarantees when $\mathbb{E}[A]$ is only partially observed through A. In particular, our approach leads to new non-asymptotic rates of convergence or exact reconstruction properties for a wide range of estimators that fall within our framework. Then, in order to show the versatility of our approach, we apply these general bounds to four problems: signed clustering, angular synchronization, Max-Cut and sparse PCA. Using our approach, we provide general excess risk and estimation bounds as optimal statistical guarantees and improve the state-of-the-art results for these five problems.

In this chapter, we provide high probability excess risk and estimation bounds satisfied by four procedures (ERM, minmax MOM and their regularized versions) in the setup introduced above, that is for the linear loss function. The proofs of all the results are postponed to Chapter 6. They use state-of-the art machinery such as localization, homogeneity arguments, local curvature and fixed point complexity parameters. In particular, there are several ways to localize around the oracle depending on the metric used; it can be either the excess risk itself or a natural local curvature metric, denoted later by the G function or the standard L_2 metric with respect to the probability measure of the data. Depending on the metric, this defines different local curvatures and different fixed points. For each type of localization, we state a statistical result. We therefore obtain various bounds for each of the four estimators in this section. Hence, this section provides a complete description of the results one can obtain for these estimators in the setup of linear loss functions and for any regularization norm. We will then apply these results to several of the problems discussed in Chapter 2 to show how these general bounds can be applied in concrete examples.

3.1 General framework

Throughout this section, we place ourselves in the classical context considered in machine learning and provide its relation with the setup from the Introduction section, in particular, we specify for each example the random matrix A appearing in (2.3) and (2.4).

Let H be a Hilbert space and X be a random vector with values in H distributed according to a distribution P. For any function $g : H \to \mathbb{R}$ for which it makes sense, we denote by $Pg := \mathbb{E}_{X\sim P}[g(X)]$ the expectation of the g function under the distribution P. For each $p \geq 1$, we denote by $\|g\|_{L_p} = (P[|g|^p])^{\frac{1}{p}}$ its $L_p(P)$-norm. Let $\mathcal{C}$ be a subset of H. For all Z in H, the loss function of Z is the *linear loss function*, $\ell_Z : X \in H \to -\langle X, Z\rangle$, which can be seen as an alignment measure and quantifies the error made when estimating Z with X. As usual in machine learning, we are interested in the best element in H that minimizes the risk (i.e. the expectation of the loss function) over $\mathcal{C}$, that is we want to estimate/learn/infer/test

$$Z^* \in \operatorname*{argmin}_{Z\in\mathcal{C}} P\ell_Z(X). \tag{3.1}$$

Sometimes Z^* is called the oracle because it is a quantity we would like to know but we usually cannot have a direct access to it because the distribution P of X is not known to the Statistician and so is the risk function $Z \to P\ell_Z(X)$. However, we have access to a sample distributed according to P. This sample (or dataset) is denoted by $\{X_i : i \in [N]\}$ where $N \in \mathbb{N}$ is called the sample size. From a mathematical point of view $(X_i)_{i\in[N]}$ is a family of *i.i.d.* random variables distributed according to P – in the section below concerning median-of-means estimators, we will relax this assumption and consider a situation where a fraction of the dataset may have been corrupted by an adversary, in that case the X_i's are not anymore assumed to be *i.i.d.*.

The setup we just introduced is pretty much the same as in the Introductory section. We

just have to identify the random matrix A for each particular examples. Since, the 'linear loss function' setup is not standard in machine learning, we provide the connection between A and the X_i's for each example:

- in community detection, $N = 1$ and $A = X_1$ is the adjacency matrix of the observed graph
- in variable clustering, $A := (1/N)\sum_{i=1}^N X_i X_i^\top$ is the empirical covariance of the observed variables X_i's
- in angular synchronization, $A = \left(e^{\iota\delta_{ij}} + \epsilon_{ij}\right)_{1\leq i<j\leq d}$ is made of the noisy measurements of the pairwise offsets.
- in the Max-Cut problem, A is the adjacency matrix of the oriented graph we observe. We want to know what is the maximum cut of the graph associated with $\mathbb{E}[A]$.
- in phase recovery, $A := \left(XX^\top\right) \circ \left(I_N - BB^+\right)$, where X is the vector of the N observed measurements $|Bx|$ and B is the measurement matrix, that is the linear operator through which the target vector x is observed.
- in sparse PCA, $A = (1/N)\sum_{i=1}^N X_i X_i^\top$ is the empirical covariance matrix of the dataset from which we want to extract the principal components.
- in the single index model, $A = (1/N)\sum_{i=1}^N Y_i T(X_i)$, where for any $i \in [N]$, $Y_i = f(\langle X_i, \beta^*\rangle) + \epsilon_i$ is the noisy output associated to the input X_i via the link function f, and $T(X_i) \in \mathbb{R}^{d\times d}$ is the second order score matrix of X_i.
- in distance metric learning, $A := \sum_{i,j=1}^N \left(X_i - X_j\right)\left(X_i - X_j\right)^\top$ where the X_i's are the observed data from which we want to learn the metric.
- in noisy optimal transport, $A = (\|X_i - Y_j\|_2^2)_{1\leq i,j\leq N}$, where $\{X_1, \ldots, X_N\}$ and $\{Y_1, \ldots, Y_N\}$ are the two sets of observed data that we wish to transport one over the other.

Remark 1. Most of the problems introduced in section 2 are presented as maximization problems, whereas ERM is a minimization problem. One can just take the opposite of the linear loss function, or replace A with $-A$, or $\mathcal{C}$ with $-\mathcal{C}$. Here, we consider the loss function $\ell_Z : A \to -\langle A, Z\rangle$.

Moving back to the 'learning with a linear loss function' introduced at the beginning of this section, we want to estimate/learn the oracle Z^* from the data $(X_i)_{i\in[N]}$. Let $\hat{Z}$ be an estimator constructed with these data. The quality of prediction of $\hat{Z}$ is measured via the excess risk $P\mathcal{L}_{\hat{Z}}$ where $Z \in \mathcal{C} \to \mathcal{L}_Z := \ell_Z - \ell_{Z^*}$ is called the excess loss. The quality of estimation of $\hat{Z}$ is measured by the error rate $\|\hat{Z} - Z^*\|_{L_2}^2$, where L_2 is taken with respect to the P distribution.

There are many ways to construct estimators in the machine learning context considered here. We will see four of them below. The most classical one is the empirical risk minimization procedure (Vapnik, 2000) introduced in the next section. Before moving to the construction of estimators, we say a word about the set $\mathcal{C}$. In all examples introduced in Section 2, $\mathcal{C}$ is a convex set because of algorithmic considerations. For our theoretical purpose, we will however need a weaker assumption given now: the star-shaped property.

Definition 1. We say that a set $\mathcal{C}$ is star-shaped in Z^* when for all $Z \in \mathcal{C}$, the segment $[Z, Z^*]$ is in $\mathcal{C}$.

In all our results we will assume $\mathcal{C}$ to be star-shaped in Z^*. This property is satisfied in all examples introduced in the Introduction chapter because a convex set is star-shaped in any of its elements.

3.2 The ERM estimator and its regularized version: definitions and general bounds

In this section, we consider the '*i.i.d.* setup' introduced in the previous section and consider the standard ERM estimator and its regularized version for which we provide high probability excess risk and estimation bounds.

3.2.1 The ERM estimator for the linear loss function

For any loss function, and in particular for the linear one considered here, $\ell_Z : X \in H \to -\langle X, Z\rangle$, which is defined for any $Z \in \mathcal{C}$, the ERM is

$$\hat{Z} \in \operatorname*{argmin}_{Z\in\mathcal{C}} P_N \ell_Z \tag{3.2}$$

where $P_N\ell_Z = (1/N)\sum_{i=1}^N \ell_Z(X_i) = (1/N)\sum_{i=1}^N -\langle X_i, Z\rangle$. The ERM is the natural empirical version of the oracle Z^* since $P\ell_Z$ appearing in the definition of Z^* in (3.1) has been replaced by its empirical counterpart $P_N\ell_Z$. When there is only one observation, that is $N = 1$, for instance in the community detection problem, we simply have $P_N\ell_Z = P_1\ell_Z = -\langle X_1, Z\rangle = -\langle A, Z\rangle$.

The study of the statistical properties of ERM estimators goes back to Vapnik and Chervonenkis (1974) and has been at the heart of many researches since then (Koltchinskii, 2011a). The results presented below are for the special case of the linear loss function. They are however based on nowadays classical concepts in machine learning.

One key quantity driving the rate of convergence of the ERM with a linear loss function is a local complexity fixed point parameter. This kind of parameter carries all the statistical complexity of the problem. It can however be hard to compute (see for instance Section 4.1 below), since it requires to control with large probability the supremum of empirical processes indexed by 'localized classes', that is the set $\mathcal{C}$ intersected with a neighborhood of the oracle. We now define such a complexity fixed point related to the problem we are considering here.

Definition 2. [Complexity fixed point parameter] Consider $0 < \delta < 1$. The fixed point complexity parameter at deviation $1-\delta$ is

$$r^*(\delta) = \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z\in\mathcal{C}: P\mathcal{L}_Z \le r} (P_N - P)\mathcal{L}_Z \le \frac{r}{2}\right] \ge 1-\delta\right). \tag{3.3}$$

Fixed point complexity parameters have been extensively used in Learning Theory since the introduction of the localization argument (Birgé & Massart, 1993; Koltchinskii, 2011b; Massart, 2007; van de Geer, 2000). When they can be computed, they are preferred to the (global) analysis developed by Chervonenkis and Vapnik (Vapnik, 1998) to study ERM, since the latter analysis always yields slower rates given that the Vapnik-Chervonenkis bound is a high-probability bound on the non-localized empirical process $\sup_{Z\in\mathcal{C}}\langle A - \mathbb{E}A, Z - Z^*\rangle$, which is an upper bound for $r^*(\delta)$ since $\{Z \in \mathcal{C} : P\mathcal{L}_Z \le r\} \subset \mathcal{C}$. The gap between the two global and local analysis can be important since fast rates cannot be obtained using the global approach, whereas the localization argument resulting in fixed points such as the one in Definition 2 may yield fast rates of convergence or even exact recovery results when $r^*(\delta) = 0$.

An example of a Vapnik-Chervonenkis's type of analysis of SDP estimators (that is a global approach) can be found in Guédon and Vershynin (2016) for the community detection problem. An improvement of the latter approach has been obtained in Fei and Chen (2019b) thanks to a localization argument. Somehow, a fixed point such as 3.4 is a sharp way to measure the statistical performances of ERM estimators, and in particular for the SDP estimators that we considered in the previous chapter. They can even be proved to be optimal (in a minimax sense)

when the noise $A - \mathbb{E}A$ is Gaussian (Lecué & Mendelson, 2013), and under mild conditions on the complexity of $\mathcal{C}$.

In what follows, we give some statistical properties of the ERM $\hat{Z}$ build from this complexity parameter (all proofs are postponed to Section 6).

Theorem 3. *We assume that the constraint $\mathcal{C}$ is star-shaped in Z^*. Then, for all $0 < \delta < 1$, with probability at least $1 - \delta$, it holds true that $P\mathcal{L}_{\hat{Z}} \leq r^*(\delta)$.*

This result shows that $\hat{Z}$ is almost a minimizer of the true objective function $Z \to -\langle \mathbb{E}A, Z \rangle$ over $\mathcal{C}$ up to $r^*(\delta)$. In particular, when $r^*(\delta) = 0$, $\hat{Z}$ is exactly a minimizer such as Z^* and, in that case, we can work with $\hat{Z}$ as if we were working with Z^* without any loss. Then, in this 'exact reconstruction case', the information contained about A on $\mathbb{E}[A]$ is enough for inferring Z^* exactly, just as if we had known $\mathbb{E}[A]$.

The importance of Theorem 3 stems from the fact that it puts forward two important concepts originally introduced in Learning Theory, namely that the complexity of the problem comes from the one of the local subset $\mathcal{C} \cap \{Z : P\mathcal{L}_Z \leq r^*(\delta)\}$, and that the 'radius' $r^*(\delta)$ of the localization is the solution of a fixed point equation.

The main conclusion of Theorem 3 is that all the information for the problem of estimating Z^* via $\hat{Z}$ is contained in the fixed point $r^*(\delta)$. We therefore have to compute or upper bound such a fixed point. This might be difficult in great generality but some tools exist that can help to find upper bounds on $r^*(\delta)$.

A first task is to understand the shape of the local sets $\mathcal{C} \cap \{Z : \langle \mathbb{E}A, Z^* - Z \rangle \leq r\}$ for $r > 0$, and to that end it is helpful to characterize the *curvature* of the excess risk $Z \to \langle \mathbb{E}A, Z^* - Z \rangle$ around its minimizer Z^*. This type of local characterization of the excess risk is also a tool used in Learning Theory that goes back to classical conditions such as the Margin assumption (Mammen & Tsybakov, 1999; Tsybakov, 2004) or the Bernstein condition (Bartlett & Mendelson, 2006). The latter condition was initially introduced as an upper bound of the variance term by its expectation: for all $Z \in \mathcal{C}$, $\mathbb{E}\left[\mathcal{L}_Z(A)^2\right] \leq c_0 \mathbb{E}\left[\mathcal{L}_Z(A)\right]$ for some absolute constant c_0, but it has now been better understood as a way to discriminate the oracle from the other points in the model $\mathcal{C}$. These assumptions were *global* assumption in the sense that they concern all Z in $\mathcal{C}$. It has been recently shown (Chinot, Lecué, & Lerasle, 2018) that only the *local curvature* of the excess risk needs to be understood. We now introduce this tool in our setup. In what follows, G is some function from H to $\mathbb{R}$. The radius defining the local subset onto which we need to understand the curvature of the excess risk is also solution of a fixed point equation.

Definition 4. [Complexity fixed point parameter with G-localization] Consider $0 < \delta < 1$. The fixed point complexity parameter with respect to a G-localization at deviation $1 - \delta$ is

$$r_G^*(\delta) = \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z \in \mathcal{C} : G(Z - Z^*) \leq r} (P_N - P)\mathcal{L}_Z \leq \frac{r}{2} \right] \geq 1 - \delta \right) \tag{3.4}$$

The difference between r^* and r_G^* lies in the fact that the local subsets are not defined with the same proximity function: r^* used the excess risk function for localization whereas r_G^* uses the G function. The latter G function should play the role of a simple description of the curvature of the excess risk around the oracle as it is granted in the following assumption.

Assumption 3.1. For all $Z \in \mathcal{C}$, if $P\mathcal{L}_Z \leq r_G^*(\delta)$ then $P\mathcal{L}_Z \geq G(Z^* - Z)$.

Typical examples of curvature functions G will have the form $G(Z^* - Z) = \theta \|Z^* - Z\|^\kappa$ for some $\kappa \geq 1$, $\theta > 0$ and some norm $\|\cdot\|$. In that case, the parameter κ was initially called the

margin parameter (Mammen & Tsybakov, 1999; Tsybakov, 2003). Even though the relation given in Assumption 3.1 has been typically referred to as a margin condition or a Bernstein condition in the Learning Theory literature, we will rather call it a *local curvature assumption*, following Guédon and Vershynin (2016) and Chinot et al. (2018), since this type of relation describes the behavior of the risk function locally around its oracle. The main advantage for finding a local curvature function G is that $r^*_G(\delta)$ should be easier to compute than $r^*(\delta)$ and $r^*(\delta) \leq r^*_G(\delta)$ because of the definition of $r^*_G(\delta)$ and $\{Z \in \mathcal{C} : \langle \mathbb{E}A, Z^* - Z\rangle \leq r^*_G(\delta)\} \subset \{Z \in \mathcal{C} : G(Z^* - Z) \leq r^*_G(\delta)\}$ (thanks to Assumption 3.1).

If this assumption holds, then we can get the following bound on the excess risk.

Theorem 5. *We assume that the constraint $\mathcal{C}$ is star-shaped in Z^* and that the 'local curvature' Assumption 3.1 holds for some $0 < \delta < 1$. With probability at least $1 - \delta$, it holds true that*

$$r^*_G(\delta) \geq P\mathcal{L}_{\hat{Z}} \geq G(Z^* - \hat{Z}).$$

When it is possible to describe the local curvature of the excess risk around its oracle by some G function and when some estimate of $r^*_G(\delta)$ can be obtained, Theorem 5 applies and estimation results of Z^* by $\hat{Z}$, with respect to both the 'excess risk' metric $\langle \mathbb{E}A, Z^* - \hat{Z}\rangle$ and the G metric follow. If not, either because understanding the local curvature of the excess risk or the computation of $r^*_G(\delta)$ is difficult, it is still possible to apply Theorem 3 with the *global* VC approach, which boils down to simply upper bound the fixed point $r^*(\delta)$ used in Theorem 3 by a global parameter that is a complexity measure of the entire set $\mathcal{C}$

$$r^*(\delta) \leq \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z \in \mathcal{C}} \langle A - \mathbb{E}A, Z - Z^*\rangle \leq \frac{r}{2}\right] \geq 1 - \delta\right). \tag{3.5}$$

Interestingly, if the latter last resort approach is used then, following the approach from Guédon and Vershynin (2016), Grothendieck's inequality (Grothendieck, 1953; Pisier, 2012) appears to be a powerful tool to upper bound the right-hand side of (3.5) in the case of the community detection problem, such as in Guédon and Vershynin (2016), as well as in the Max-Cut problem. Of course, when it is possible to avoid this likely suboptimal global approach, one should do so because the local approach will always provide better results.

Finally, proving a 'local curvature' property such as in Assumption 3.1 may be difficult because it requires to understand the shape of the local subsets $\mathcal{C} \cap \{Z : \langle \mathbb{E}A, Z^* - Z\rangle \leq r\}, r > 0$. It is however possible to simplify this assumption if getting estimation results of Z^* only with respect to the G function – and not necessarily an upper bound on the excess risk $\langle \mathbb{E}A, Z^* - \hat{Z}\rangle$ – is enough. In that case, Assumption 3.1 may be replaced by the following one.

Assumption 3.2. For all $Z \in \mathcal{C}$, if $G(Z^* - Z) \leq r^*_G(\delta)$, then $P\mathcal{L}_Z \geq G(Z^* - Z)$.

Assumption 3.2 assumes a curvature of the excess risk function in a 'G neighborhood' of Z^*, unlike Assumption 3.1 which grants this curvature in an 'excess risk neighborhood'. The shape of a neighborhood defined by the G function may be easier to understand – for instance when G is a norm, a neighborhood defined by G is the ball of a norm centered at Z^* with radius $r^*_G(\delta)$. In general, the latter assumption and Assumption 3.1 do not compare. The following result establishes that, under Assumption 3.2, $\hat{Z}$ is a good estimate of Z^* with respect to the G function, but no guarantee on the excess risk is obtained.

Theorem 6. *We assume that the constraint $\mathcal{C}$ is star-shaped in Z^* and that the 'local curvature' Assumption 3.2 holds for some $0 < \delta < 1$. We assume that the G function is continuous, $G(0) = 0$ and $G(\lambda(Z^* - Z)) \leq \lambda G(Z^* - Z)$ for any $\lambda \in [0, 1]$ and $Z \in \mathcal{C}$. Then, with probability at least $1 - \delta$, it holds true that $G(Z^* - \hat{Z}) \leq r^*_G(\delta)$.*

Results like Theorem 3, 5 and 6 appeared in many papers on ERM in Learning Theory such as in Bartlett and Mendelson (2006), Koltchinskii (2011b), Massart (2007) and Lecué and Mendelson (2013). In all these results, typical loss functions, such as the quadratic or logistic loss functions, were not linear ones such as the one we are using here. From that point of view, our problem is easier. What is much more complicated here than in other more classical problems in Learning Theory is the computation of the fixed point. First, because the stochastic process $Z \to \langle A - \mathbb{E}A, Z - Z^* \rangle$ may be far from being a Gaussian process if the noise matrix $A - \mathbb{E}A$ is complicated. Secondly, because the local sets $\{Z \in \mathcal{C} : \langle \mathbb{E}A, Z^* - Z \rangle \leq r\}$ or $\{Z \in \mathcal{C} : G(Z^* - Z) \leq r\}$ for $r > 0$ may be very hard to describe in a simple way. However, instrumental results are available in the literature to circumvent this kind of problems, such as in Fei and Chen (2019b), or those we will present in Section 4.1.

In the next chapter, we will see how these results apply in community detection, signed clustering, angular group synchronization and the Max-Cut problem. All these problems share the feature that the oracle Z^* does not have some special structure onto which one can leverage on to improve the rates of convergence. They are however situations such as in sparse PCA where the target has a structure that can be used to improve statistical the performances. In such cases, one may consider some regularization procedures like in the following section.

3.2.2 The Regularized ERM estimator for the linear loss function

We focus here on structural learning in which targets/oracles have a structure (such as sparsity, low rank or regularity) onto which the statistician can leverage to construct more statistically efficient estimators. The typical approach to this problem is to regularize the ERM in order to force the estimator toward the desired structure.

We place ourselves in the framework defined above in Section 3.1 except that we need here a regularization function, i.e. a function that favors some structure. In this work, we consider a general norm defined at least on the span of $\mathcal{C}$ and denoted by $\|\cdot\|$. Typical examples are the ℓ_1 norm and the trace-norm used in high-dimensional statistics to induce sparsity or low-rank. When Z^* has some structure a natural way to force an estimator toward Z^* is by adding a mutliple of this norm. This yields to the regularized ERM, later called RERM:

$$\hat{Z}^{\mathrm{RERM}} \in \operatorname*{argmin}_{Z \in \mathcal{C}} \left(P_N \ell_Z + \lambda \|Z\| \right) \tag{3.6}$$

where $\lambda > 0$ is called the regularization parameter and has the role to make a trade-off between the data adequation term $P_N \ell_Z$ and the regularization term $\|Z\|$.

As for the ERM, convergence rates achieved by the RERM $\hat{Z}^{\mathrm{RERM}}$ are driven by a local complexity fixed point parameter. However, the regularization norm appears in this type of parameter: it is now the set $\mathcal{C}$ intersected with balls with respect to $\|\cdot\|$ centered at Z^* (and for some radius) that are "localized" by some neighborhood of Z^*. Somehow the model in structural learning is of the form $\mathcal{C} \cap \{Z : \|Z - Z^*\| \leq r\}$. As in the ERM case, one may consider two different ways to construct localization: either via the excess risk or via a local curvature G function. However, to avoid a lengthy presentation, we focus only on the latter one, i.e. on a localization via a local curvature G function because it is this result that we will use for the our application later in sparse PCA. In what follows, we consider a function $G : H \to \mathbb{R}$, which characterizes the curvature of the objective function, i.e. the risk, $Z \in H \to P\ell_Z$ around its minimizer Z^*.

Definition 7. For parameter $A > 0$, radius $\rho > 0$ and deviation parameter $\delta \in (0, 1)$, we define

the complexity fixed point for the structural learning with a linear loss function by

$$r^*_{\mathrm{RERM,G}}(A,\rho,\delta) = \inf\left(r>0 : \mathbb{P}\left(\sup_{Z\in\mathcal{C}:\|Z-Z^*\|\leq\rho, G(Z-Z^*)\leq r} |(P-P_N)\mathcal{L}_Z| \leq \frac{r}{3A}\right) \geq 1-\delta\right),$$

where we recall that for all $Z \in \mathcal{C}$, $\mathcal{L}_Z = \ell_Z - \ell_{Z^*}$ is the excess loss function of Z.

After introducing the fixed point $r^*_{\mathrm{RERM,G}}(A,\rho,\delta)$, we are now in a position to introduce the G function. As we already mentioned above, the G function describes the curvature of the excess risk locally around the oracle.

Assumption 3.3. We assume there exist $A>0$, $\rho^*>0$ and $\delta\in(0,1)$ such that, for all $Z\in\mathcal{C}$ satisfying $G(Z-Z^*) = r^*_{\mathrm{RERM,G}}(A,\rho^*,\delta)$ and $\|Z-Z^*\|\leq\rho^*$, then $AP\mathcal{L}_Z \geq G(Z-Z^*)$.

We now leverage on the structure inducing property of the regularization norm and explain what features must the radius ρ^* appearing in Assumption 3.3 have in relation to this property. We will use the assumption below, that is adapted from the one in Lecué and Mendelson (2017), to get the statistical bounds satisfied by the RERM estimator $\hat{Z}^{RERM}$. The idea is that the regularization norm $\|.\|$ is expected to promote some structure by having a large subdifferential at elements in H having this structure. First, let us recall what the subdifferential of $\|.\|$ at a point Z is:

$$(\partial\|.\|)_Z := \left\{\Phi\in H : \|Z+h\|-\|Z\| \geq \langle\Phi,h\rangle \text{ for all } h\in H\right\}.$$

Elements in $(\partial\|.\|)_Z$ are called the *subgradients* of $\|\cdot\|$ in Z. What matters in structural learning to get fast rates is that Z^* is close to an element with a structure induced by the regularization norm. Therefore we consider the set of all subgradients of $\|\cdot\|$ of points close to Z^*:

$$\text{for any } \rho>0 : \quad \Gamma_{Z^*}(\rho) = \bigcup_{Z\in Z^*+\frac{\rho}{20}B} (\partial\|.\|)_Z,$$

where B is the unit ball of $\|.\|$. We expect $\Gamma_{Z^*}(\rho)$ to be a large subset of the unit dual sphere (or dual ball, when $0\in Z^*+(\rho/20)B$) of $\|.\|$ when Z^* is structured or close to a structured element in H, for the notion of structure associated with $\|.\|$. This intuition is formalized in the following definition.

Definition 8. For $A>0$, $\rho>0$ and $\delta\in(0,1)$ we define:

$$H_{\rho,A} := \left\{Z\in\mathcal{C} : \|Z-Z^*\| = \rho \;\; \textit{and} \;\; G(Z-Z^*) \leq r^*_{\mathrm{RERM},G}(A,\rho,\delta)\right\}$$

and

$$\Delta(\rho,A) := \inf_{Z\in H_{\rho,A}} \sup_{\Phi\in\Gamma_{Z^*}(\rho)} \langle\Phi, Z-Z^*\rangle.$$

We say that $\rho>0$ satisfies the **A-sparsity equation** when $\Delta(\rho,A)\geq(4/5)\rho$.

Note that it is always true that $\Delta(\rho,A)\leq\rho$ – because $\|Z-Z^*\|=\rho$ and Φ is a subgradient of $\|\cdot\|$ – hence, a radius ρ satisfying the A-sparsity equation is somehow extremal up to the absolute constant $4/5$ (the analysis works for any other absolute constant, there is nothing special with $4/5$). It means that $\Gamma_{Z^*}(\rho)$ is almost as big as the unit dual sphere (or ball) of $\|\cdot\|$.

All the material introduced above (complexity fixed points, local curvatures and the sparsity equation) are the corner stones of our statistical analysis of RERMs. Once introduced, we are in a position to state our main result on RERM estimators for linear loss functions and a general regularization norm.

Theorem 9. *Let $\delta \in (0,1)$. Assume that the constraint set $\mathcal{C}$ is star-shaped in Z^*. Consider a continuous function $G : H \to \mathbb{R}$ such that $G(0) = 0$. Suppose the existence of $A > 0$ and $\rho^* > 0$ such that Assumption 3.3 holds and $\rho^* > 0$ satisfies the A-sparsity equation from Definition 8. Define the function $r^*(.) := r^*_{\mathrm{RERM,G}}(A, ., \delta)$ and assume that*

$$\frac{10}{21A}\frac{r^*(\rho^*)}{\rho^*} < \lambda < \frac{2}{3A}\frac{r^*(\rho^*)}{\rho^*}. \tag{3.7}$$

Then, with probability at least $1-\delta$, the following bounds hold for the RERM estimator defined in (3.6):

$$\|\hat{Z}^{\mathrm{RERM}} - Z^*\| \leq \rho^* \quad , \quad G(\hat{Z}^{\mathrm{RERM}} - Z^*) \leq r^*(\rho^*) \quad \textit{and} \quad P\mathcal{L}_{\hat{Z}^{\mathrm{RERM}}} \leq \frac{r^*(\rho^*)}{A}.$$

We note that in the case where G is the risk function $Z \to P\ell_Z$ - that is when the excess risk is used for localization, because, by linearity $G(Z - Z^*) = P\ell_{Z-Z^*} = P\mathcal{L}_Z$ - Assumption 3.3 is trivially verified with $A = 1$, and as a consequence Theorem 9 applies.

3.3 The Median of Means estimator and its regularized version: definitions and general bounds

In this section, we move to the construction and the statistical analysis of another family of estimators introduced in Lecué and Lerasle (2020) whose aims are to solve robustness issues related to adversarial contamination of the dataset as well as heavy-tailed data. We are interested here in the case where our data could be contaminated by possible outliers generated by an adversary and the inliers data may be heavy-tailed. Even though the framework seems not in favor of statisticians because the dataset is of poor quality, we still want to achieve the same statistical performance as if there was no outliers and light-tailed (such as sub-gaussian) data. It is known that the classical ERM or RERM approaches from the previous section do not perform well in general on this type of dataset and that is the reason why we move to MOM estimators.

The statistical framework considered in this section cannot be the ideal *i.i.d.* setup considered in the previous section that fits well for ERM and RERM. Indeed, the *i.i.d.* framework do not allow for adversarial corruption. That is why we consider the following setup in this section.

Assumption 3.4. [Adversarial contamination setup] Let N *i.i.d.* random vectors $(\widetilde{X}_i)_{i=1}^N$ in H. These vectors are first given to an adversary who is allowed to modify up to $|\mathcal{O}|$ of them. This modification does not have to follow any rule and is unknown to the statistician. This leads to the modified dataset $\{X_1, \ldots, X_N\}$ that the adversary gives to the statistician. Hence, the dataset at hands $\{X_1, \ldots, X_N\}$ is said to be 'adversarially' contaminated. It can be partitioned into two groups: the modified data $(X_i)_{i\in\mathcal{O}}$, which can be seen as outliers and the 'good data', or inliers, $(X_i)_{i\in\mathcal{I}}$ such that for any $i \in \mathcal{I}$, $X_i = \widetilde{X}_i$. Of course, the statistician does not know which data has been modified or not so that the partition $\mathcal{O} \cup \mathcal{I} = \{1, \ldots, N\}$ is unknown to the statistician.

Remark 2. Since there are two types of data considered in Assumption 3.4 (the 'good' $\widetilde{X}_i$'s and the corrupted ones X_i's), we need to be clear on the objects we will be using later: the risk function and its associated oracle are the one associated with the 'good' data:

$$Z \in \mathcal{C} \to P\ell_Z = \mathbb{E}\langle -\widetilde{X}, Z\rangle \text{ and } Z^* \in \operatorname*{argmin}_{Z\in\mathcal{C}} P\ell_Z,$$

where $\widetilde{X}$ has the same probability distribution as $\widetilde{X}_1, \dots, \widetilde{X}_N$. It is also the same for the L_2-norm: for all $Z \in H$, $\|Z\|_{L_2} = \sqrt{\mathbb{E}\langle \widetilde{X}, Z\rangle^2}$. Note that the L_2-norm is in general different from the original Hilbert norm defining H, which is denoted by $\|\cdot\|_2$.

The adversarial contamination setup addresses several questions in statistics regarding the rates of convergence, the probability deviations and the number of outliers. Many approaches have been introduced to answer these questions (Huber & Ronchetti, 2009). There was an important renewal of this topic during the last ten years (Catoni, 2012; Diakonikolas et al., 2016). The approach we use in this section is based on the median-of-means principle (Jerrum, Valiant, & Vazirani, 1986; Nemirovsky & Yudin, 1983): $[N]$ is partitioned into K equal-size groups $B_1, \dots, B_K$ (without loss of generality, K is assumed to divide N, otherwise we only have to remove some data). For any function $g : H \to \mathbb{R}$ and $k \in [K]$ we define $P_{B_k} g = (K/N) \sum_{i \in B_k} g(X_i)$, the empirical mean of g over B_k. Then, we define $\mathrm{MOM}_k(g)$ as the median of these K empirical means:

$$\mathrm{MOM}_K(g) := \mathrm{Med}(P_{B_1} g, \dots, P_{B_K} g).$$

This data partition scheme is at the heart of our approach to answer the robustness issues. It is used as a building block in the minmax MOM estimator. We recall its construction and provide its statistical properties in the remaining of this section as well as for its regularized version for the robust structural learning problem.

3.3.1 The minmax MOM estimator for the linear loss function.

To solve the robustness to adversarial corruption as well as to heavy-tailed data, one can use a systematic approach called the minimax MOM estimator in Lecué and Lerasle (2020). It works whenever a loss function exists and a robust gradient descent algorithm may also be constructed out of it (see Lecué and Lerasle (2020) for more details). When the dataset has been splited into K equal size blocks, it takes the following form:

$$\hat{Z}_K^{\mathrm{MOM}} \in \operatorname*{argmin}_{Z \in \mathcal{C}} \sup_{Z' \in \mathcal{C}} \mathrm{MOM}_K(\ell_Z - \ell_{Z'}) \tag{3.8}$$

and can therefore be used in the particular case studied here of the linear loss function $x \to \ell_Z(x) = -\langle Z, x\rangle$. From our theoretical perspective, the aim of the minmax MOM estimator $\hat{Z}_K^{\mathrm{MOM}}$ is to achieve the rates of convergence for the same deviation probabilty in the contaminated and heavy-tailed setup as in the ideal *i.i.d.* setup with light-tailed data, as long as the number of outliers is not too large. It is the aim of the next section to prove such statistical bounds. As for the ERM case, rates of convergence are given by local complexity fixed points that depends on the choice of localization. Below, we consider three different ways to localize: either via the $L_2(P)$-norm, or via the excess risk or via some general curvature function G.

MOM estimator with excess-risk localization.

As previously for ERMs, the convergence rate of the minmax MOM estimator is driven by a local complexity fixed point parameters. In this section, we consider the case where the excess risk is simple enough so that it can serve as a localization. In that case, there is no need to identify the curvature of the excess risk locally around Z^* since the excess risk describes it by itself. There is therefore no curvature assumption. In the next two paragraphs the picture will be different.

Definition 10. Let $\sigma_1, \dots, \sigma_N$ be N independent Rademacher variables which are independent of the $\widetilde{X}_i$'s. For $\gamma > 0$, we define:

$$r^*_{\mathrm{MOM,ER}}(\gamma) := \inf\left\{ r > 0 \ : \ \max\left(\frac{\mathrm{E}(r)}{\gamma}, \sqrt{12800} V_K(r) \right) \le r^2 \right\}$$

where, for all $r > 0$,

$$\mathrm{E}(r) := \mathbb{E}\left[\sup_{Z\in\mathcal{C}:P\mathcal{L}_Z\leq r^2}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right|\right] \text{ and } V_K(r) := \sqrt{\frac{K}{N}}\sup_{Z\in\mathcal{C}:P\mathcal{L}_Z\leq r^2}\sqrt{\boldsymbol{V}ar(\mathcal{L}_Z(\widetilde{X}))}.$$

In the case of excess risk localization, there is no need for other tools than the fixed point $r^*_{\mathrm{MOM,ER}}(\gamma)$ to describe the rate of convergence of the minmax MOM. This is what shows the following result.

Theorem 11. *We consider the adversarial contamination setup of Assumption 3.4. We assume that the constraint set $\mathcal{C}$ is star-shaped in Z^*. Let $\gamma = 1/6400$ and consider K, a divisor of N such that $K \geq 100|\mathcal{O}|$. Then, it holds true that with probability at least $1 - \exp(-72K/625)$, $P\mathcal{L}_{\hat{Z}^{\mathrm{MOM}}_K} \leq r^*_{\mathrm{MOM,ER}}(\gamma)^2$.*

Compared to the fixed point from Definition 2 describing the rate of convergence of the ERM, we note that the one from Definition 10 uses a local Rademacher complexity, denoted by $E(r)$, and a variance term, denoted by $V_K(r)$; there is no need to upper bound with high probability the supremum of an empirical process. For minmax MOM estimators, the task of computing fixed point complexity parameters is therefore easier. Moreover, as one can see in Theorem 11, the convergence rate is obtained with an exponentially large probability even though no strong concentration property is assumed; only the existence of a second moment (so that the variance term $V_K(r)$ exists) is required. This shows the robustness to heavy-tail data of minmax MOM estimators for the linear loss function as well as its robustness with respect to adversarial contamination since it is proved in the setup of Assumption 3.4. However, the computation of the complexity term $E(r)$ may require more moments than just 2 in order to recover a Gaussian regime, i.e. a rate achieved when the data have a light subgaussian tail.

MOM estimator with L_2-localization.

In this section, we consider the case where the behaviour/curvature of the excess risk locally around the oracle Z^* is well described by the L_2-norm to the square. This is the situation when a margin assumption $AP\mathcal{L}_Z \geq \|Z - Z^*\|^2_{L_2}, \forall Z \in \mathcal{C}$ holds, i.e. with a margin parameter equal to 2 Mammen and Tsybakov (1999). In that case, one needs to modify the definition of the complexity fixed point parameter by using a L_2-localization.

Definition 12. Let $\sigma_1, \ldots, \sigma_N$ be independent Rademacher variables which are independent of the $\widetilde{X}_i$'s. For $\gamma > 0$, we define

$$r^*_{\mathrm{MOM},L_2}(\gamma) := \inf\left(r > 0 : \mathbb{E}\left[\sup_{Z\in\mathcal{C}:\ \|Z-Z^*\|_{L_2}\leq r}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right|\right] \leq \gamma r^2\right)$$

where we recall that $\|Z\|_{L_2} = \sqrt{\mathbb{E}\langle\widetilde{X}, Z\rangle^2}$ for all $Z \in H$.

As we said above, we use the L_2-norm in the localization to define the fixed point $r^*_{\mathrm{MOM},L_2}(\gamma)$ when it describes the curvature of the excess risk around Z^*. We now formalize this property in the next assumption.

Assumption 3.5. There exists $A > 0$ such that for any $Z \in \mathcal{C}$, if $\|Z - Z^*\|^2_{L_2} \leq C_{K,A}$, then $\|Z - Z^*\|^2_{L_2} \leq AP\mathcal{L}_Z$, where $C_{K,A} := \max\left(r^*_{\mathrm{MOM},L_2}(\gamma)^2, \gamma^{-1}A^2(K/N)\right)$ for $\gamma = 1/3200$.

Looking at Assumption 3.5, this may be surprising to have a quadratic term $\|Z - Z^*\|_{L_2}^2$ describing a linear term $P\mathcal{L}_Z = \langle \mathbb{E}\tilde{X}, Z^* - Z \rangle$. However, one may see that the local curvature of the excess risk from Assumption 3.5 holds only for Z in $\mathcal{C}$ not in H. Thanks to the two tools introduced above (a local complexity fixed point and a curvature assumption), we are now ready to state our main result on the minmax MOM estimator in the adversarial contamination setup for a L_2-localization.

Theorem 13. *We consider the adversarial contamination setup of Assumption 3.4. We assume that the constraint set $\mathcal{C}$ is star-shaped in Z^*. Let $\gamma = 1/3200$. Assume the existence of $0 < A < 1$ such that Assumption 3.5 holds. Let K be a divisor of N such that $K \geq 100|\mathcal{O}|$. Then, it holds true that with probability at least $1 - \exp(-72K/625)$:*

$$P\mathcal{L}_{\hat{Z}_K^{\mathrm{MOM}}} \leq \frac{C_{K,A}}{A} \text{ and } \|\hat{Z}_K^{\mathrm{MOM}} - Z^*\|_{L_2}^2 \leq C_{K,A}.$$

Theorem 13 can be used under a margin assumption with a margin parameter equal to 2. It can be extended to margin parameter other than 2. However, one may be interested in other situations where the local curvature of the excess risk is not described by the square of the L_2 norm but for instance by the square of the native Hilbert norm of H - as it will be the case for the sparse PCA problem. In the next paragraph, we provide a statistical bound for the minmax MOM estimator for a local curvature of the excess risk described by a general G function.

MOM estimator with G localization.

In this final paragraph regarding the minmax MOM estimator, we consider a general G function describing locally the excess risk around Z^* and derive statistical bounds when this function is used for localization. When applied to the particular cases of the excess risk or the L_2 norm to the square, we recover the last two results. However, other G functions may be considered, for instance, if the calculation of $r^*_{\mathrm{MOM,ER}}(\gamma)$ is too hard or if L_2-norm to the square does not describe well enough the excess risk. We need first to define a complexity fixed point for a localization with respecto to a general G function. Unlike in the previous section dealing with the L_2 to the square localization and as in the last but one section dealing with a excess risk localization, there is a variance term in this fixed point equation.

Definition 14. Let $\sigma_1, \ldots, \sigma_N$ be N independent Rademacher variables which are independent of the $\tilde{X}_i$'s. For $G : H \to \mathbb{R}$ and $\gamma > 0$, we define:

$$r^*_{\mathrm{MOM,G}}(\gamma) := \inf\left\{ r > 0 \ : \ \max\left(\frac{\mathrm{E}_G(r)}{\gamma}, \sqrt{12800} V_{K,G}(r) \right) \leq r^2 \right\}$$

where, for all $r > 0$,

$$\mathrm{E}_G(r) := \mathbb{E}\left[\sup_{Z \in \mathcal{C} : G(Z - Z^*) \leq r^2} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathcal{L}_Z(\tilde{X}_i) \right| \right]$$

$$\text{and} \quad V_{K,G}(r) := \sqrt{\frac{K}{N}} \sup_{Z \in \mathcal{C} : G(Z - Z^*) \leq r^2} \sqrt{\boldsymbol{V}ar(\mathcal{L}_Z(\tilde{X}))}.$$

The function G characterizes the curvature of the excess risk $Z \in \mathcal{C} \to P\mathcal{L}_Z = \langle \mathbb{E}X, Z^* - Z \rangle$ locally around its minimizer Z^*. This is formalized in the following assumption.

Assumption 3.6. There exist $A > 0$ and $\gamma > 0$ such that for all $Z \in \mathcal{C}$, if $G(Z - Z^*) \leq (r^*_{\mathrm{MOM,G}}(\gamma))^2$, then $AP\mathcal{L}_Z \geq G(Z - Z^*)$.

The difference between $r^*_{\mathrm{MOM,ER}}$ and $r^*_{\mathrm{MOM,G}}$ is that the local subsets are not defined using the same proximity function to the oracle Z^*. The main advantage in finding a curvature function G satisfying Assumption 3.6 is that $r^*_{\mathrm{MOM,G}}$ may be easier to compute than $r^*_{\mathrm{MOM,ER}}$, since the shape of a neighborhood defined by G may be easier to understand than the one defined by the excess risk. However, one always has $r^*_{\mathrm{MOM,ER}} \leq r^*_{\mathrm{ERM,G}}$ since there is no better way to describe the excess risk than the excess risk itself. We now obtain statistical bounds satisfied by the minmax MOM estimator (3.8) under this local curvature assumption.

Theorem 15. *We consider the adversarial contamination setup of Assumption 3.4. We assume that the constraint set $\mathcal{C}$ is star-shaped in Z^*. We consider a continuous function $G : H \to \mathbb{R}$ be a continuous function. Let $\gamma = 1/6400$. We assume the existence of $0 < A < 2$ such that the local curvature Assumption 3.6 holds for those values of γ and G. Then, with probability at least $1 - \exp(-72K/625)$ it holds true that:*

$$P\mathcal{L}_{\hat{Z}^{\mathrm{MOM}}_K} \leq \frac{1}{2} r^*_{\mathrm{MOM,G}}(\gamma)^2 \qquad \text{and} \qquad G(Z^* - \hat{Z}^{\mathrm{MOM}}_K) \leq r^*_{\mathrm{MOM,G}}(\gamma)^2.$$

Theorem 15 may be applied in the examples introduced from Section 2 if one is willing to handle robustness issues for these (none structured) learning problems. If one wants to handle the robustness issues in structural learning then one may consider regularized versions of the minmax MOM estimator.

3.3.2 The regularized minmax MOM estimator for the linear loss function

We are now considering the setup of structural learning that allows for high-dimensional statistics, i.e. when the dimension of the parameter to estimate Z^* is larger than the number of observations. In that case, some structure is usually assumed to be satisfied by Z^* and should be taken into account for the construction of estimators. On top of that, we consider a setup where the data may have been corrupted by some outliers and the inliers may be heavy-tailed. We therefore have to face several issues related to robustness and high-dimensions that we propose to solve using a regularized version of the minmax MOM estimator introduced in Section 3.3.1:

$$\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} \in \operatorname*{argmin}_{Z \in \mathcal{C}} \sup_{Z' \in \mathcal{C}} \left(\mathrm{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\| - \|Z'\|) \right) \tag{3.9}$$

where $\lambda > 0$ is some regularization parameter and $\|\cdot\|$ is a norm inducing some structure. In the following sections, we provide statistical guarantees for this estimator. As in the previous sections, the convergence rates depend on local complexity fixed points, local curvature properties of the excess risk and of the 'structure inducing power' of the regularization norm $\|\cdot\|$. As previously, the choice of the localization function plays a key role in the definition of all these concepts. We therefore consider three paragraphs depending on the localization function used: it can either be the excess risk, the L_2-norm or some general function G.

RMOM estimator with excess-risk localization.

As in the previous section, we start with the excess risk localization.

Definition 16. Let $\sigma_1, \ldots, \sigma_N$ be independent Rademacher variables which are independent of the $\tilde{X}_i$'s. For $\gamma > 0$ and $\rho > 0$, we define:

$$r^*_{\mathrm{RMOM,ER}}(\gamma, \rho) := \inf \left\{ r > 0 \ : \ \max \left(\frac{\mathrm{E}(r, \rho)}{\gamma}, 400\sqrt{2} V_K(r, \rho) \right) \leq r^2 \right\}$$

where, for all $\rho, r > 0$ and $\mathcal{C}_{\rho,r} = \left\{Z \in \mathcal{C} : \|Z - Z^*\| \leq \rho, P\mathcal{L}_Z \leq r^2\right\}$,

$$\mathrm{E}(r,\rho) := \mathbb{E}\left[\sup_{Z\in\mathcal{C}_{\rho,r}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\tilde{X}_i)\right|\right] \text{ and } V_K(r,\rho) := \sqrt{\frac{K}{N}}\sup_{Z\in\mathcal{C}_{\rho,r}}\sqrt{\boldsymbol{V}ar(\mathcal{L}_Z(\tilde{X}))}.$$

The sparsity equation introduced for the study of the RERM in Definition 8 has to be slightly modified according to this new definition of the complexity parameter.

Definition 17. For $\gamma > 0$ and $\rho > 0$, let

$$\bar{H}_\rho := \left\{Z \in \mathcal{C} : \|Z - Z^*\| = \rho \text{ and } P\mathcal{L}_Z \leq r^*_{\mathrm{RMOM,ER}}(\gamma,\rho)^2\right\}$$

and

$$\bar{\Delta}(\rho) := \inf_{Z\in\bar{H}_\rho}\sup_{\Phi\in\Gamma_{Z^*}(\rho)}\left\langle\Phi, Z - Z^*\right\rangle.$$

We say that ρ satisfies the **sparsity equation** if $\bar{\Delta}(\rho) \geq 4\rho/5$.

We are now ready to state our main statistical result satisfied by the regularized minmax MOM estimator for the linear loss function and for an excess-risk localization.

Theorem 18. *We consider the adversarial contamination setup of Assumption 3.4. Let $K \in [N]$ be such that $K \geq 100|\mathcal{O}|$. Let $\rho^* > 0$ satisfying the sparsity equation from Definition 17. Let $\gamma = 1/3200$ and take $\lambda = (11/(40\rho^*))r^*_{\mathrm{RMOM,ER}}(\gamma, 2\rho^*)$ as regularization parameter. Then, with probability at least $1 - 2\exp(-72K/625)$,*

$$P\mathcal{L}_{\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}} \leq r^*_{\mathrm{RMOM,ER}}(\gamma,2\rho^*)^2 \quad \textit{and} \quad \|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\| \leq 2\rho^*.$$

Note that one may replace $r^*_{\mathrm{RMOM,ER}}(\gamma, 2\rho^*)$ by any real number r^* larger than $r^*_{\mathrm{RMOM,ER}}(\gamma, 2\rho^*)$. This observation is particularly useful since we usually only know how to upper bound local complexity fixed points such as $r^*_{\mathrm{RMOM,ER}}(\gamma, 2\rho^*)$ and that we use it to define λ, the regularization parameter.

RMOM estimator with L_2 localization.

In this section, we look at the case where the L_2-norm to the square is used to describe the local curvature of the excess risk. As we mentioned above, it is the case when the margin assumption with margin parameter equals to 2 holds. We define below the appropriate complexity fixed point parameter, the local curvature assumption and the associated sparsity equation.

Definition 19. Let $(\sigma_i)_{i\leq N}$ be independent Rademacher variables independent of the X_i's. For $\rho > 0$ and $\gamma > 0$, we define:

$$r^*_{\mathrm{RMOM},L_2}(\gamma,\rho) := \inf\left(r > 0 : \mathbb{E}\left[\sup_{Z\in\mathcal{C}:\|Z-Z^*\|\leq\rho,\|Z-Z^*\|_{L_2}\leq r}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\tilde{X}_i)\right|\right] \leq \gamma r^2\right).$$

We turn now to the sparsity equation that is used to construct the radius ρ^* which defines the model $\mathcal{C} \cap (Z^* + \rho^* B)$ where both Z^* and $\hat{Z}^{RMOM}_{K,\lambda}$ lie (with high probability).

Definition 20. For γ, ρ and $A > 0$, let:

$$C_K(\gamma, \rho, A) := \max\left(320000A^2\frac{K}{N}, r^*_{\mathrm{RMOM},L_2}(\gamma,\rho)^2\right),$$
$$\widetilde{H}_{\rho,A} := \left\{Z \in \mathcal{C} : \|Z - Z^*\| = \rho \text{ and } \|Z - Z^*\|_{L_2} \le \sqrt{C_K(\gamma,\rho,A)}\right\}$$
$$\text{and} \quad \widetilde{\Delta}(\rho, A) := \inf_{Z\in\widetilde{H}_{\rho,A}} \sup_{\Phi\in\Gamma_{Z^*}(\rho)} \langle \Phi, Z - Z^*\rangle.$$

A real number $\rho > 0$ is said to satisfy the **A-sparsity equation** if $\widetilde{\Delta}(\rho, A) \ge 4\rho/5$.

The next definition is the formal way to say that the L_2-norm to the square can be used to describe the curvature of the excess risk closed to the oracle.

Assumption 3.7. There exists A, γ and $\rho^* > 0$ such that ρ^* satisfies the A-sparsity equation from Definition 20 and for both $b \in \{1, 2\}$ and all $Z \in \mathcal{C}$, if $\|Z - Z^*\|^2_{L_2} = C_K(\gamma, b\rho^*, A)$ and $\|Z - Z^*\| \le b\rho^*$, then $\|Z - Z^*\|^2_{L_2} \le AP\mathcal{L}_Z$.

After introducing the three key concepts in structural learning: local complexity fixed point, local curvature assumption and the sparsity equation, we can now state our excess risk and estimation (with respect to both L_2 and the regularization norm) bounds.

Theorem 21. *We consider the adversarial contamination setup of Assumption 3.4. Let K be a divisor of N and assume that $K \ge 100|\mathcal{O}|$. Grant Assumption 3.7 for some $A \in (0, 1]$, $\gamma = 1/32000$ and ρ^* that satisfies the A-sparsity equation from Definition 20. Define $\lambda = (11/(40\rho^*))C_K(\gamma, 2\rho^*, A)$. Then it holds true that with probability at least $1-2\exp(-72K/625)$:*

$$\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\| \le 2\rho^*, \quad P\mathcal{L}_{\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}} \le \frac{93}{100} r^*_{\mathrm{RMOM,L_2}}(\gamma, 2\rho^*)^2, \quad \textit{and}$$
$$\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|^2_{L_2} \le r^*_{\mathrm{RMOM,L_2}}(\gamma, 2\rho^*)^2.$$

Again the same result as the one of Theorem 21 holds if one replaces $r^*_{\mathrm{RMOM,L_2}}$ by any upper bound on $r^*_{\mathrm{RMOM,L_2}}$.

RMOM estimator with G localization.

Finally, we consider a function $G : H \to \mathbb{R}$ that is expected to describe well the local curvature of the excess risk and that is used to define all the subsequent localization. An example of such a G function is given in the sparse PCA case studied later. Indeed, in Lemma 39 below, we will use $Z \in \mathbb{R}^{d\times d} \to G(Z) = \|Z\|_2^2$ as a localization function (we recall that $\|\cdot\|_2$ is the canonical norm over H; it is in general different from the L_2 one that was used above for localization). We are now introducing a complexity fixed point that uses the G function for localization.

Definition 22. Let $\sigma_1, \ldots, \sigma_N$ be independent Rademacher variables independent of the $\tilde{X}_i$'s. For $G : H \to \mathbb{R}$ and A, γ and $\rho > 0$, we define:

$$r^*_{\mathrm{RMOM,G}}(\gamma, \rho) := \inf\left\{r > 0 \ : \ \max\left(\frac{\mathrm{E}_G(r,\rho)}{\gamma}, 400\sqrt{2}V_{K,G}(r,\rho)\right) \le r^2\right\}$$

where, for all $r, \rho > 0$,

$$\mathrm{E}_G(r,\rho) := \mathbb{E}\left[\sup_{Z\in\mathcal{C}_{\rho,r}}\left|\frac{1}{N}\sum_{i=1}^N \sigma_i\mathcal{L}_Z(\tilde{X}_i)\right|\right] \text{ and } V_{K,G}(r,\rho) := \sqrt{\frac{K}{N}}\sup_{Z\in\mathcal{C}_{\rho,r}}\sqrt{\boldsymbol{V}ar(\mathcal{L}_Z(\tilde{X}))},$$

with $\mathcal{C}_{\rho,r} = \{Z \in \mathcal{C} : \|Z - Z^*\| \le \rho, G(Z - Z^*) \le r^2\}$

An example of computation of an upper bound of the local complexity fixed point $r^*_{\mathrm{RMOM,G}}(\gamma,\rho)$ is provided in the sparse PCA example in Lemma 49 below. The final ingredient to derive the rate of convergence is the radius ρ that needs to satisfies a sparsity equation.

Definition 23. For all γ and $\rho > 0$, consider

$$\bar{H}_\rho := \left\{ Z \in \mathcal{C} : \|Z - Z^*\| = \rho \text{ and } G(Z - Z^*) \leq r^*_{\mathrm{RMOM,G}}(\gamma,\rho)^2 \right\}$$

and

$$\bar{\Delta}(\rho) := \inf_{Z \in \bar{H}_\rho} \sup_{\Phi \in \Gamma_{Z^*}(\rho)} \langle \Phi, Z - Z^* \rangle.$$

We say that ρ satisfies the **sparsity equation** if $\bar{\Delta}(\rho) \geq 4\rho/5$.

Finally, we write the assumption saying that the G function is indeed appropriate to describe the excess risk locally around Z^*.

Assumption 3.8. There exists $A > 0$, $\gamma > 0$ and $\rho^* > 0$ such that ρ^* satisfies the spartsity equation from Definition 23 and for both $b \in \{1, 2\}$ and all $Z \in \mathcal{C}$, if $G(Z - Z^*) = r^*_{\mathrm{RMOM,G}}(\gamma^*, b\rho^*)^2$ and $\|Z - Z^*\| \leq b\rho^*$, then $AP\mathcal{L}_Z \geq G(Z - Z^*)$.

We are now ready to state the following result on the statistical properties of the regularized minimax MOM in the context of robust structural learning with a linear loss function and for a general G function describing the local curvature of the excess risk.

Theorem 24. *We consider the adversarial contamination setup of Assumption 3.4. Let $G : H \to \mathbb{R}$ be a continuous function such that $G(0) = 0$ and for all $\alpha \geq 1$ and $Z \in \mathcal{C}, G(\alpha(Z - Z^*)) \geq \alpha G(Z - Z^*)$. Let $K \in [N]$ be such that $K \geq 100|\mathcal{O}|$. Grant Assumption 3.8 for some $A \in (0, 1]$, $\gamma = 1/32000$ and ρ^* that satisfies the sparsity equation from Definition 23. Define $\lambda = (11/(40\rho^*))r^*_{\mathrm{RMOM,G}}(\gamma, 2\rho^*)$. Then with probability at least $1 - 2\exp(-72K/625)$, it holds true that:*

$$\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\| \leq 2\rho^*, \quad P\mathcal{L}_{\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}} \leq \frac{93}{100} r^*_{\mathrm{RMOM,G}}(\gamma, 2\rho^*)^2, \quad \text{and}$$
$$G(\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*) \leq r^*_{\mathrm{RMOM,G}}(\gamma, 2\rho^*)^2.$$

In the sparse PCA example, Theorem 24 will be applied for the study of a ℓ_1-regularized minmax MOM estimator. However, applying Theorem 24 requires several intermediate results, such as proving that $Z \to G(Z) = \|Z\|_2^2$ can be used as a local curvature of the excess risk, find a ρ^* satisfying the sparsity equation of Definition 23 and compute an upper bound for the local complexity fixed point $r^*_{\mathrm{RMOM,G}}(\gamma,\rho)$. For the last task, one needs to handle the variance term $V_{K,G}$ as well as the complexity term $E_G(r,\rho)$. For the latter, we need to find an upper bound on the expected supremum of a Rademacher process over the interpolation body $\mathcal{C}_{\rho,r} = \{Z \in \mathcal{C} : \|Z - Z^*\| \leq \rho, G(Z - Z^*) \leq r^2\}$. This step is usually the hardest one and requires some techniques from empirical process theory that we are now developing in the next section.

In the previous chapter, we developed theoretical tools that enable us to obtain high probability excess risk and estimation bounds for procedures that take the form of ERM or minmax MOM estimators, all based on a linear loss function. In this opening chapter, we deploy these theoretical tools to address four classical problems in the realm of statistical learning: the signed clustering problem, the angular synchronization problem, the Max-Cut problem and the sparse PCA problem. Using our methodology to handle those problems, we are able to improve the state-of-the art results as well as provide statistical optimal guarantees for adversarially corrupted and heavy-tailed data.

4.1 Tools for the computation of local complexity fixed points

In this section, we present concentration and in expectation results for two specific interpolation norms of the difference between the covariance matrix and its empirical version. These results are typical results that we used to compute local complexity fixed points like the ones introduced in the previous section. Indeed, in order to use any of the general statistical bounds presented in Chapter 3, we have to compute local complexity fixed points. We provide two such examples in this section that will be useful for the rest of the chapter when we will apply our methodology to several classic problems. Note that the bounds presented here hold under weak moment assumptions.

4.1.1 Statistical setup

In this section, $X_1, \dots, X_N$ are *i.i.d.* centered random vectors in $\mathbb{R}^d$ and we denote by Σ their covariance matrix, that is $\Sigma := \mathbb{E}X_1^\top X_1$. The entries of Σ are denoted by Σ_{pq}: $\Sigma_{pq} = \mathbb{E}X_{1p}X_{1q}$ for all $p, q \in [d]$. We denote the empirical covariance matrix by $\hat{\Sigma}_N = (1/N)\sum_{i=1}^N X_i^\top X_i$ and its entries by $\hat{\Sigma}_{pq}$, $p, q \in [d]$. The aim of this section is to provide large deviation and in expectation upper bounds for the norm of $\Sigma - \Sigma_N$, for two norms defined by interpolation bodies. We define B_1 and B_2 as the unit balls for the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively.

4.1.2 Control of $\|\Sigma - \hat{\Sigma}_N\|$ for a B_2/B_1 interpolation norm.

In order to upper bound the deviation of the empirical covariance matrix $\hat{\Sigma}_N$ around Σ with respect to some norm, we need to assume some concentration properties on the X_i's. We therefore consider such an assumption now.

Assumption 4.1. There exists $w \geq 0$ and $t \geq 1$ such that the following holds. For all $p, q \in [d]$ and all $2 \leq r \leq 2\log(ed/k) + t$ we have $\|X_{1p}X_{1q} - \mathbb{E}(X_{1p}X_{1q})\|_{L_r} \leq w^2 r$.

In other words, Assumption 4.1 is a growth condition on the first $2\log(ed/k) + t$ moments of the products $X_{1p}X_{1q}$ of the coordinates of X_1. This growth condition is the one exhibited by sub-exponential (i.e. ψ_1) variables.

This is, for instance, the case of a product of two sub-gaussian (i.e. ψ_2) variables because $\|UV\|_{\psi_1} \leq \|U\|_{\psi_2}\|V\|_{\psi_2}$ and the r-th moment of a ψ_α variable growths like $r^{1/\alpha}$ (see Chapter 1 in Chafaï, Guedon, Lecue, and Pajor (2012) for more details).

Assumption 4.1 does not require the existence of any moment beyond the $2\log(ed/k)+t$-th moment and is therefore called a weak moment assumption: Assumption 4.1 essentially assumes the existence of $\log(ed/k)$ subgaussian moments on the coordinates of the data. We will see below that this assumption is enough to get estimation result for the estimator of the first sparse principal component in deviation with the improved rate of order

$$\sqrt{\frac{k^2\log(ed/k)}{N}}. \tag{4.1}$$

Consider $k \in [d]$. We denote by $\|\cdot\|$ the following interpolation pseudo-norm onto $\mathbb{R}^{d\times d}$ defined by

$$\|A\| = \sup\left(\langle A, Z\rangle : Z \in kB_1 \cap B_2, Z = Z^\top\right). \tag{4.2}$$

Theorem 25. *There exists an absolute constant c_0 such that the following holds. Grant Assumption 4.1 for some w and $t \geq 1$ and assume that $N \geq 2\log(ed/k)+t$. With probability at least $1-\exp(-t)$, it holds true that*

$$\|\hat{\Sigma}_N - \Sigma\| \leq c_0 w^2 \sqrt{\frac{k^2(\log(ed/k)+t)}{N}}$$

Moreover, if Assumtion 4.1 holds for some $w \geq 0$ and $t = 1$, and provided that $N \geq 2\log(ed/k)+1$, it holds true that $\mathbb{E}\left[\|\hat{\Sigma}_N - \Sigma\|\right] \leq c_0 w^2 \sqrt{\frac{6k^2\log(ed/k)}{N}}$.

Remark 3. Classical estimation result require the number of observations to be larger than $k\log(ed/k)$ where s is the sparsity of signal to be reconstructed. Here, we observe in Theorem 25 that N is only asked to be larger than $\log(ed/k)$ so it is a much weaker assumption. This is due to the fact that we do not have to lower bound a quadratic process, since our loss function is linear. It is usually isomorphic or just lower bounds results on a quadratic process which requires N to be larger than the sparsity up to a log factor, a property that we don't need here in the context of linear loss functions.

4.1.3 Control of $\|\Sigma - \hat{\Sigma}_N\|$ for a B_2/SLOPE interpolation norm.

We consider the following assumption.

Assumption 4.2. There exists $w \geq 0$ and $t \geq 3$ such that the following holds. For all $p, q \in [d]$ and all $2 \leq r \leq \log(ed^2)+t$ we have $\|X_{1p}X_{1q} - \mathbb{E}(X_{1p}X_{1q})\|_{L_r} \leq w^2 r$.

Our aim is to analyze the statistical properties of a SLOPE regularization for various problems and to show that the optimal rate

$$\sqrt{\frac{k^2\log(ed/k)}{N}}$$

can be achieved by a unique regularization method which does not require the a priori knowledge on the sparsity parameter k. To that end we introduce the SLOPE regularization norm of a $d\times d$ matrix A

$$\|A\|_{SLOPE} = \sum_{p,q=1}^{d} b_{pq} A^*_{(p,q)}$$

where $\mathbf{b} := (b_{pq} : p, q \in [d])$ are decreasing weights for some lexicographical order over $[d]^2$ starting at $(1,1)$ such that for all $k \in [d]$, $b_{kk} = \sqrt{\log(ed^2/k^2) + t}$. For instance, one may assume that b is a symetric matrice and set $b_{pq} = \sqrt{\log(ed^2/(pq)) + t}$ when $q \geq p$. We also denote by $(A^*_{(p,q)} : p, q \in [d])$ the non-increasing sequence (for the same lexicographical order over $[d]^2$ used before) of the rearrangement of the absolute values of the entries of A, for instance $A^*_{(d,d)} = \min(|A_{pq}| : p, q \in [d])$ and $A^*_{(1,1)} = \max(|A_{pq}| : p, q \in [d])$. We denote by B_{SLOPE} the unit ball of the $SLOPE$ norm.

Consider $\rho > 0$. We denote by $\|\cdot\|_\rho$ the following interpolation pseudo-norm onto $\mathbb{R}^{d\times d}$ defined by

$$\|A\|_\rho = \sup\left(\langle A, Z\rangle : Z \in \rho B_{SLOPE} \cap B_2\right). \tag{4.3}$$

Theorem 26. *There exists an absolute constant c_0 such that the following holds. Consider $k \in [d]$ and $\gamma \geq 1$. Grant Assumption 4.2 for some w and $t \geq \max\left(2\log(\lceil\log(k^2)\rceil), \gamma\log(ed^2/k^2)\right)$ and assume that $N \geq \log(ed^2) + t$. With probability at least $1 - 2\exp(-t/2)$, it holds true that*

$$\|\hat{\Sigma}_N - \Sigma\|_\rho \leq \frac{c_0 w^2}{\sqrt{N}} \min(\rho, d).$$

4.2 The graph clustering problem (stochastic block model): revisiting results from literature

In order to support the relevance of the methodology developed in the previous chapter, based on the curvature of excess risk, we begin by showing that it is possible - and this is the starting point of this work - by looking at some of the existing evidence in the literature through the new prism it provides, to achieve the same result in a simpler way. To this end, we look at the graph clustering problem, for which the existing literature is abundant.

Indeed, the rapid growth of social networks on the Internet has led many statisticians and computer scientists to focus their research on data coming from graphs, and it turns out that an important topic that has attracted particular interest during the last decades is that of community detection (Fortunato, 2010; Porter, Onnela, & Mucha, 2009), where the goal is to recover mesoscopic structures in a network, the so-called *communities*. A community consists of a group of nodes that are relatively densely connected to each other, but sparsely connected to other dense groups present within the network. The motivation for this line of work stems not only from the fact that finding communities in a network is an interesting and challenging problem of its own, as it leads to understanding structural properties of networks, but community detection is also used as a data pre-processing step for other statistical inference tasks on large graphs, as it facilitates parallelization and allows one to distribute time-consuming processes on several smaller subgraphs (that is, the extracted communities).

One challenging aspect of the community detection problem arises in the setting of sparse graphs. Many of the existing algorithms, which enjoy theoretical guarantees, do so in the relatively dense regime for the edge sampling probability, where the expected average degree is of the order $\Theta(\log d)$. The problem becomes challenging in very sparse graphs with bounded average degree. To this end, Guédon and Vershynin (2016) proposed a semidefinite relaxation for a discrete optimization problem, an instance of which encompasses the community detection problem, and showed that it can recover a solution with any given relative accuracy even in the setting of very sparse graphs with an average degree of order $O(1)$.

A subset of the existing literature for community detection and clustering relies on spectral methods, which consider the adjacency matrix associated with a graph and employ its eigenvalues, and especially eigenvectors, in the analysis process or to propose efficient algorithms to solve the task at hand. Along these lines, Le, Levina, and Vershynin (2016) proposed a

general framework for optimizing a general function of the graph adjacency matrix over discrete label assignments by projecting onto a low-dimensional subspace spanned by vectors that approximate the top eigenvectors of the expected adjacency matrix. The authors consider the problem of community detection with $k = 2$ communities, which they frame as an instance of their proposed framework, combined with a regularization step that shifts each entry in the adjacency matrix by a small constant τ, which renders their methodology applicable in the sparse regime as well.

In the remainder of this section, we focus on the community detection problem on random graphs under the general stochastic block model. We build on the works of Guédon and Vershynin (2016) and Fei and Chen (2019b), who both tackle this problem, and revisit the evidence they provide via the prism of our methodology.

4.2.1 Statistical setup

We place ourselves within the context of the stochastic block model (SBM). We consider a set of vertices $V = \{1, \cdots, d\}$, and assume it is partitioned into K communities $\mathcal{C}_1, \cdots, \mathcal{C}_K$ of arbitrary sizes $|\mathcal{C}_1| = \ell_1, \cdots, |\mathcal{C}_K| = \ell_K$.

Definition 27. For any pair of nodes $i, j \in V$, we denote by $i \sim j$ when i and j belong to the same community (that is, there exists $k \in \{1, \dots, K\}$) such that $i, j \in \mathcal{C}_k$), and we denote by $i \not\sim j$ if i and j do not belong to the same community.

For each pair (i, j) of nodes from V, we draw an edge between i and j with a fixed probability p_{ij} independently from the other edges. We assume that there exist numbers p and q satisfying $0 < q < p < 1$, such that

$$\begin{cases} p_{ij} > p, \text{ if } i \sim j \text{ and } i \neq j, \\ p_{ij} = 1, \text{ if } i = j, \\ p_{ij} < q, \text{ otherwise.} \end{cases} \tag{4.4}$$

We denote by $A = (A_{i,j})_{1 \leq i,j, \leq d}$ the observed symmetric adjacency matrix, such that, for all $1 \leq i \leq j \leq d$, A_{ij} is distributed according to a Bernoulli of parameter p_{ij}.

The community structure of such a graph is captured by the membership matrix $\bar{Z} \in \mathbb{R}^{d \times d}$, defined by $\bar{Z}_{ij} = 1$ if $i \sim j$, and $\bar{Z}_{ij} = 0$ otherwise. The main goal in community detection is to reconstruct $\bar{Z}$ from the observation A.

4.2.2 Revisiting Guédon and Vershynin's results

Spectral methods for community detection are very popular in the literature (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Clauset, Newman, & Moore, 2004; Fei & Chen, 2019b; Guédon & Vershynin, 2016; Vershynin, 2018). There are many ways to introduce such methods, one of which being via convex relaxations of certain graph cut problems aiming to minimize a modularity function such as the RatioCut (Newman, 2006). Such relaxations often lead to SDP estimators, such as the ones introduced in Chapter 3.

Considering a random graph distributed according to the generalized stochastic block model, and its associated adjacency matrix A (i.e. $A = A^\top$ and $A_{ij} \sim Bern(p_{ij})$ for $1 \leq i \leq j \leq d$ and p_{ij} as defined in (4.4)), we will estimate its membership matrix $\bar{Z}$ via the following SDP estimator

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmax}} \langle A, Z \rangle,$$

where $\mathcal{C} = \{Z \in \mathbb{R}^{d\times d}, Z \succeq 0, Z \geq 0, \mathrm{diag}(Z) \preceq I_d, \sum_{i,j=1}^d Z_{ij} \leq \lambda\}$ and $\lambda = \sum_{i,j=1}^d \bar{Z}_{ij} = \sum_{k=1}^K |\ell_k|^2$ denotes the number of nonzero elements in the membership matrix $\bar{Z}$. The motivation for this approach stems from the fact that the membership matrix $\bar{Z}$ is actually the oracle, that is, $Z^* = \bar{Z}$ (see Lemma 7.1 in Guédon and Vershynin, 2016 or Lemma 29 below), where

$$Z^* \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle \mathbb{E}A, Z\rangle.$$

Following the strategy from Theorem 3 and from our point of view, the upper bound on $r^*(\delta)$ from Guédon and Vershynin (2016) is the one that is based on the global approach – that is, without localization. Indeed, they use the observation that, for all $r > 0$, it holds true that

$$\sup_{Z\in\mathcal{C}:\langle \mathbb{E}A, Z^*-Z\rangle \leq r} \langle A - \mathbb{E}A, Z - Z^*\rangle \overset{(a)}{\leq} \sup_{Z\in\mathcal{C}} \langle A - \mathbb{E}A, Z - Z^*\rangle \overset{(b)}{\leq} 2K_G \|A - \mathbb{E}A\|_{\mathrm{cut}}, \tag{4.5}$$

where (b) is Grothendieck's inequality they recall in Lemma 28 below. Therefore, the localization around the oracle Z^* by the excess risk 'band' $B_r^* := \{Z : \langle \mathbb{E}A, Z^* - Z\rangle \leq r\}$ is simply removed in inequality (a). As a consequence, the resulting statistical bound is based on the complexity of the entire class $\mathcal{C}$ whereas, in a localized approach, only the complexity of $\mathcal{C} \cap B_r^*$ matters.

Lemma 28 (Grothendieck's inequality)**.** *For $C \in \mathbb{R}^{d\times d}$, we define its 'cut' norm in the following way:*

$$\|C\|_{cut} := \max_{s,t\in\{-1,1\}^d} \sum_{i,j=1}^d C_{ij} s_i t_j\,, \quad \mathcal{C} = \{Z \succeq 0 : Z_{ii} = 1, i = 1, \ldots, d\}$$

Then, there exists an absolute constant K_G, called the Grothendieck constant, such that for any $C \in \mathbb{R}^{d\times d}$, one has:

$$\sup_{Z\in\mathcal{C}} \langle C, Z\rangle \leq K_G \|C\|_{cut} = K_G \max_{s,t\in\{-1,1\}^d} \sum_{i,j=1}^d C_{ij} s_i t_j$$

The next step in the proof of Guédon and Vershynin (2016) is a high-probability upper bound on $\|A - \mathbb{E}A\|_{\mathrm{cut}}$ which follows from Bernstein's inequality (see (6.88) in Appendix 6.2) and a union bound: since one has $\|A - \mathbb{E}A\|_{\mathrm{cut}} = \max_{x,y\in\{-1,1\}^d} \langle A - \mathbb{E}A, xy^\top\rangle$, then for all $t > 0$, $\|A - \mathbb{E}A\|_{\mathrm{cut}} \leq td(d-1)/2$ with probability at least $1 - \exp\left(2d\log 2 - (d(d-1)t^2)/(16\bar{p} + 8t/3)\right)$ where $\bar{p} := 2/[d(d-1)] \sum_{i<j} p_{ij}(1-p_{ij})$. The resulting upper bound on the fixed point obtained in Guédon and Vershynin (2016) is:

$$r^*(\delta) \leq \frac{8}{3} K_G \left(2d\log(2) + \log(\frac{1}{\delta})\right). \tag{4.6}$$

Then, applying Theorem 3 gives that for any $\delta \in (0, 1)$, one has with probability at least $1 - \delta$, that $\langle \mathbb{E}A, Z^* - \hat{Z}\rangle \leq r^*(\delta)$. Let us then place ourselves under the condition of Theorem 1 in

Guédon and Vershynin (2016): we consider $\epsilon \in (0,1)$, $d \in \mathbb{N}$ such that $d \geq 5.10^4/\epsilon^2$ and we assume that $\max\big(p(1-p), q(1-q)\big) \geq 20/d$, as well as the existence of a and b such that $p = a/d > b/d = q$ and $(a-b)^2 \geq 2.10^4\epsilon^{-2}(a+b)$. Then, taking $\delta = e^3 5^{-d}$, we obtain

$$\big\langle \mathbb{E}A, Z^* - \hat{Z} \big\rangle \leq r^*(\delta) \leq \epsilon d^2 = \epsilon \|Z^*\|_2^2$$

which is the result from Theorem 1 in Guédon and Vershynin (2016). Finally, Guédon and Vershynin use a (global) curvature property of the excess risk in their Lemma 7.2.

Lemma 29 (Lemma 7.2 in Guédon and Vershynin (2016))**.** *For all $Z \in \mathcal{C}$,*

$$\big\langle \mathbb{E}A, Z^* - Z \big\rangle \geq \frac{p-q}{2} \|Z^* - Z\|_1$$

Therefore, a (global – that is for all $Z \in \mathcal{C}$) curvature assumption holds for a G function which is here the $\ell_1^{d\times d}$ norm, a margin parameter $\kappa = 1$ and $\theta = (p-q)/2$ for the community detection problem. However, this curvature property is not used to compute a 'better' fixed-point parameter, but only to obtain a $\ell_1^{d\times d}$ estimation bound since

$$\|\hat{Z} - Z^*\|_1 \leq \left(\frac{2}{p-q}\right) \big\langle \mathbb{E}A, Z^* - \hat{Z} \big\rangle \leq \frac{16 K_G (2d\log(2) + \log(1/\delta))}{3(p-q)}.$$

The latter bound together with the Davis-Kahan sin-Theta theorem (see Corollary 1 in Yu, Wang, and Samworth (2014)) allow the authors to obtain estimation bounds for the community membership vector x^*.

4.2.3 Revisiting Fei and Chen results

The approach from Fei and Chen (2019b) improves upon the one in Guédon and Vershynin (2016) because it uses a localization argument: the curvature property of the excess risk function from Lemma 29 is used to improve the upper bound in (4.6) obtained following a global approach. Indeed, Fei and Chen (2019b) obtain a high-probability upper bound on the quantity

$$\sup_{Z \in \mathcal{C} : \|Z^* - Z\|_1 \leq r} \big\langle A - \mathbb{E}A, Z - Z^* \big\rangle$$

depending on r. This leads to an exact reconstruction result in the 'dense' case and exponentially decaying rates of convergence in the 'sparse' case. This is a typical example where the localization argument shows its advantage upon the global approach. The price to pay is usually a more technical proof for the local approach compared with the global one. However, the argument from Fei and Chen, 2019b also uses an unnecessary peeling argument together with an unnecessary a priori upper bound on $\|\hat{Z} - Z^*\|_1$ (which is actually the one from Guédon and Vershynin (2016)). It appears that this peeling argument and this a priori upper bound on $\|\hat{Z} - Z^*\|_1$ can be avoided thanks to our approach from Theorem 3. This improves the probability estimate and simplifies the proofs (since the result from Guédon and Vershynin, 2016 is not required anymore, and neither is the peeling argument). For the signed clustering problem we consider in the next section as an application of our main results, we will mostly adapt the probabilistic tools from Fei and Chen (2019b) (in the 'dense' case) to the methodology associated with Theorem 3, without needing either of these two arguments.

4.3 The signed clustering problem (signed stochastic block model)

Much of the clustering literature, including both spectral and non-spectral methods, has focused on unsigned graphs, where each edge carries a non-negative scalar weight that encodes a measure of affinity (similarity, trust) between pairs of nodes. However, in numerous instances, the above-mentioned affinity takes negative values, and encodes a measure of dissimilarity or distrust. Such applications arise in social networks where users relationships denote trust-distrust or friendship-enmity, shopping bipartite networks which capture like-dislike relationships between users and products (Banerjee, Sarkar, Gokalp, Sen, & Davulcu, 2012), online news and review websites, such as Epinions and Slashdot, that allow users to approve or denounce others (Leskovec, Huttenlocher, & Kleinberg, 2010a), and clustering financial or economic time series data (Aghabozorgi, Shirkhorshidi, & Wah, 2015). Such applications have spurred interest in the analysis of signed networks, which has recently become an increasingly important research topic (Leskovec, Huttenlocher, & Kleinberg, 2010b), with relevant lines of work in the context of clustering signed networks including, in chronological order, Kunegis et al. (2010), Chiang, Whang, and Dhillon (2012) and Cucuringu, Davies, Glielmo, and Tyagi (2019). The latter work proposed regularized versions of signed clustering methods to handle sparse graphs – a regime where standard spectral methods are known to underperform.

4.3.1 Statistical setup

We focus on the problem of clustering a K-weakly balanced graph: a signed graph is said to be K-weakly balanced if and only if all the edges are positive, or the nodes can be partitioned into $K \in \mathbb{N}$ disjoint sets such that positive edges exist only within clusters, and negative edges are only present across clusters (J. A. Davis, 1967). We consider a signed stochastic block model (SSBM) similar to the one introduced in Cucuringu et al. (2019), where we are given a graph G with d nodes $\{1, \dots, d\}$ which are divided into K communities, $\{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$, such that, in the noiseless setting, edges within each community are positive and edges between communities are negative.

The only information available to the user is given by a $d \times d$ sparse adjacency matrix A constructed as follows: A is symmetric, with $A_{ii} = 1$ for all $i \in [d]$, and for all $1 \leq i < j \leq d$, $A_{ij} = s_{ij}(2B_{ij} - 1)$ where

$$\mathrm{B}_{ij} \sim \begin{cases} \mathrm{Bern}(p) \text{ if } i \sim j \\ \mathrm{Bern}(q) \text{ if } i \not\sim j \end{cases} \quad \text{and} \quad s_{ij} \sim \mathrm{B}ern(\delta),$$

for some $0 \leq q < 1/2 < p \leq 1$ and $\delta \in (0, 1)$. Moreover, all the variables B_{ij}, s_{ij} for $1 \leq i < j \leq n$ are assumed to be independent.

We remark that this SSBM model is similar to the one considered in Cucuringu et al. (2019), which was governed by two parameters, the sampling probability δ as above, and the noise level η, which may flip entries of the adjacency matrix.

Our aim is to recover the community membership matrix or cluster matrix $\bar{Z} = (\bar{Z}_{ij})_{i,j \leq d}$, with $\bar{Z}_{ij} = 1$ when $i \sim j$ and $\bar{Z}_{ij} = 0$ when $i \not\sim j$ using only the observed censored adjacency matrix A.

Our approach is similar in nature to the one used by spectral methods in community detection. We first observe that for $\alpha := \delta(p + q - 1)$ and $J = (1)_{d \times d}$ we have $\bar{Z} = Z^*$ where

$$Z^* \in \underset{Z \in \mathcal{C}}{\operatorname{argmax}} \langle \mathbb{E}A - \alpha J, Z \rangle, \quad \mathcal{C} = \{Z \in \mathbb{R}^{d \times d} : Z \succeq 0, Z_{ij} \in [0, 1], Z_{ii} = 1, \forall i \in [d]\} \tag{4.7}$$

The proof of (4.7) is detailed in Section 6.2.2.

Since we do not know $\mathbb{E}A$ and α, we should estimate both of them. We will estimate $\mathbb{E}A$ with A but, for simplicity, we will assume that α is known. The resulting estimator of the cluster matrix $\bar{Z}$ is

$$\hat{Z} \in \operatorname*{argmax}_{Z \in \mathcal{C}} \langle A - \alpha J, Z \rangle, \tag{4.8}$$

which is indeed an ERM estimator based on the observation of the matrix $A - \alpha J$ and the loss function $Z \to \ell_Z(A) = -\langle A, Z \rangle$. We can therefore use the tools introduced in section 3.2.1 to obtain statistical properties on the estimation of Z^* from (4.7) by $\hat{Z}$.

We will use the following notations: $s := \delta(p-q)^2$, $\theta := \delta(p-q)$, $\rho := \delta \max\{1-\delta(2p-1)^2, 1-\delta(2q-1)^2\}$, $\nu := \max\{2p-1, 1-2q\}$, $[m] := \{1, \cdots, m\}$ for all $m \in \mathbb{N}$, $l_k := |\mathcal{C}_k|$ for all $k \in [K]$, $\lambda^2 := \sum_{k=1}^K l_k^2$, $\mathcal{C}^+ := \bigcup_{k=1}^K (\mathcal{C}_k \times \mathcal{C}_k)$ and $\mathcal{C}^- := \bigcup_{k \neq k'} (\mathcal{C}_k \times \mathcal{C}_{k'})$. We also use the notation $c_0, c_1, \ldots$ to denote absolute constants whose values may change from one line to another.

4.3.2 Main result for the estimation of the cluster matrix in signed clustering

Our main result concerns the reconstitution of the K communities from the observation of the matrix A. In order to avoid solutions with some communities of degenerated size (too small or too large), we consider the following assumption.

Assumption 4.3. Up to constants, the elements of the partition $\mathcal{C}_1 \sqcup \cdots \sqcup \mathcal{C}_K$ of $\{1, \ldots, d\}$ are of the same size: there are absolute constants c_0 and c_1 such that for any $k \in [K]$, $d/(c_1 K) \leq |\mathcal{C}_k| = l_k \leq c_0 d/K$.

We are now ready to state the main result on the estimation of the cluster matrix Z^* from (4.7) by the SDP estimator $\hat{Z}$ from (4.8).

Theorem 30. *There is an absolute positive constant c_0 such that the following holds. Grant Assumption 4.3. Assume that*

$$d\nu\delta \geq \log d \tag{4.9}$$

$$sd \geq c_0 K^2 \nu \tag{4.10}$$

$$\text{and} \quad \frac{K \log(2eKd)}{d} \leq \max\left(\frac{\theta^2}{\rho}, \frac{9\rho}{32}\right) \tag{4.11}$$

Then, with probability at least $1 - \exp(-\delta\nu d) - 3/(2eKd)$, exact recovery holds true, that is $\hat{Z} = Z^*$. We recall the constants defined above : $s := \delta(p-q)^2$, $\theta := \delta(p-q)$, $\rho := \delta \max\{1-\delta(2p-1)^2, 1-\delta(2q-1)^2\}$, $\nu := \max\{2p-1, 1-2q\}$.

Therefore, we have exact reconstruction in the dense case (that is under assumption (4.9)), which stems from condition (4.11).

The latter condition is in the same spirit as the one in Theorem 1 of Fei and Chen (2019b), it measures the SNR (signal-to-noise ratio) of the model which captures the hardness of the SSBM. As mentioned in Fei and Chen (2019b), it is related to the Kesten-Stigum threshold (Mossel, Neeman, & Sly, 2015). If this condition is dropped out, then we do not have anymore exact reconstruction but only a controlled exponential rate of convergence: there exists a universal constant C_2 such that, with probability at least $1 - \exp(-\delta\nu d) - 3/(2eKd)$, it holds true that

$$\|Z^* - \hat{Z}\|_1 \leq \frac{2ed^2}{c_1\theta} \exp\left(-\frac{sd}{C_2 K}\right). \tag{4.12}$$

A proof of (4.12) can be found in 6.2.2.

This shows that in the dense case, exact reconstruction is possible when $(K^2 + K\log(d))/d \lesssim 1$ and, otherwise, we only have a control of the estimation error with an exponential convergence rate.

We then obtain results of the same nature as in Fei and Chen (2019b), or in the more recent paper of Xu, Jog, Sun, and Loh (2020). In those two articles, the authors show the existence of a phase transition, with exact recovery in the regime $(K^2 + K\log(d))/d \lesssim 1$, and exponential rate with exponent $\simeq -sd/K$ otherwise, where s is some measurement of the SNR of the problem. Note that the estimation bound is given with respect to the $\ell_1^{d\times d}$ norm. This is not a surprise since it is the behavior of the excess risk over $\mathcal{C}$ around Z^*.

In some recent works (Fei & Chen, 2019a, 2020), the authors were able to obtain sharp constants in the rate (4.12) for the Synchronization model, the Censored Block Model as well as the Stochastic Block Model. Their proof relies on the construction of a dual certificate and goes through the study of the dual problem. We see the proof technique behind Theorem 30 of different nature as a straight 'primal' approach and it is not clear how to relate the two approaches. The two similar approaches were both developed in the compressed sensing and matrix completion problems (to name a few) where the 'primal' approach was based on the Null Space Property or the Restricted Isometry Property or some neighborliness property (Foucart & Rauhut, 2013) and, at the same time and for the same problems, a 'dual' approach relying on the construction of a dual or approximate dual certificate was performed (Candès & Tao, 2010; Gross, 2011). But, to the best of our knowledge, no clear connection has been made between the two approaches. It would be however interesting to have a clear picture on the two types of approaches, and see if they are actually the same or coming from a more general approach.

4.4 The angular synchronization problem

In this section, we introduce the group synchronization problem as well as a stochastic model for this problem. We consider a SDP relaxation of the original problem (which is exact) and construct the associated SDP estimator such as in (2.4).

4.4.1 Statistical setup

The angular synchronization problem consists of estimating d unknown angles $\theta_1, \cdots, \theta_d$ (up to a global shift angle) given a noisy subset of their pairwise offsets $(\theta_i - \theta_j)[2\pi]$, where $[2\pi]$ is the modulo 2π operation. The pairwise measurements can be realized as the edge set of a graph G, typically modeled as an Erdös-Renyi random graph (Singer, 2011).

The aim of this section is to show that the angular synchronization problem can be analyzed using our methodology. In order to keep the presentation as simple as possible, we assume that all pairwise offsets are observed up to some Gaussian noise: we are given $\delta_{ij} = (\theta_i - \theta_j + \sigma g_{ij})[2\pi]$ for all $1 \le i < j \le d$ where $(g_{ij} : 1 \le i < j \le d)$ are $d(d-1)/2$ *i.i.d.* standard Gaussian variables and $\sigma > 0$ is the noise variance. We may rewrite the problem as follows: we observe a $d \times d$ complex matrix A defined by

$$A = S \circ [x^*(\overline{x^*})^\top] \text{ where } S = (S_{ij})_{d\times d}, S_{ij} = \begin{cases} e^{\iota\sigma g_{ij}} & \text{if } i < j \\ 1 & \text{if } i = j \\ e^{-\iota\sigma g_{ij}} & \text{if } i > j \end{cases} \tag{4.13}$$

ι denotes the imaginary number such that $\iota^2 = -1$, $x^* = (x_i^*)_{i=1}^d \in \mathbb{C}^d$, $x_i^* = e^{\iota\theta_i}, i = 1, \ldots, d$, $\bar{x}$ denotes the conjugate vector of x and $S \circ [x^*(\overline{x^*})^\top]$ is the element-wise product $(S_{ij}x_i\bar{x}_j)_{d\times d}$. In particular, S is a Hermitian matrix (i.e. $\bar{S}^\top = S$), with $\mathbb{E}[S_{ij}] = \exp(-\sigma^2/2)$ for $i \neq j$ and $\mathbb{E}[S_{ii}] = 1$ if $i = j$. We want to estimate $(\theta_1, \ldots, \theta_d)$ (up to a global shift) from the observation of the matrix of data A.

Unlike the statistical model introduced in A. Bandeira et al. (2016), the noise here is multiplicative in A. From a physical point of view, it makes more sense to consider an additive noise on the offsets, that is we observe $(\theta_i - \theta_j + \sigma g_{ij})[2\pi]$. The noise becomes multiplicative by passing to the exponential. However, to compare our methodology with the one from A. Bandeira et al. (2016), we also consider the model therein (that is, an additive noise on the matrix $Z^* = x^*(\overline{x^*})^\top$ instead of the additive noise in A). We recover similar results in this latter model than the one in A. Bandeira et al. (2016). For the moment, we consider the multiplicative noise and the data matrix A as introduced above in (4.13). We will turn to the addtive noise model from A. Bandeira et al. (2016) at the very end of this section in a remark.

The first step is to find an (vectorial) optimization problem which solutions are given by $(\theta_i)_{i=1}^d$ (up to global angle shift) or some bijective function of it. Estimating $(\theta_i)_{i=1}^d$ up to global angle shift is equivalent to estimating the vector $x^* = (e^{\iota\theta_i})_{i=1}^d$. The latter is, up to a global rotation of its coordinates, the unique solution of the following maximization problem

$$\operatorname*{argmax}_{x\in\mathbb{C}^d:|x_i|=1} \left\{ \bar{x}^\top \, \mathbb{E}A \, x \right\} = \{(e^{\iota(\theta_i+\theta_0)})_{i=1}^d : \theta_0 \in [0, 2\pi)\}. \tag{4.14}$$

A proof of (4.14) is given in Section 6.2.3. Let us now rewrite (4.14) as an ERM with a linear loss function. To that end, we classically use a lifting procedure (d'Aspremont, El Ghaoui, Jordan, & Lanckriet, 2007; Lemaréchal & Oustry, 2018) that we will now describe. For all $x \in \mathbb{C}^d$, we have $\bar{x}^\top \mathbb{E}Ax = \operatorname{tr}(\mathbb{E}AX) = \langle \mathbb{E}A, X \rangle$ where $X = x\bar{x}^\top$ and $\{Z \in \mathbb{C}^{d\times d} : Z = x\bar{x}^T, |x_i| = 1\} = \{Z \in \mathbb{H}_d : Z \succeq 0, \operatorname{diag}(Z) = \mathbf{1}_d, \operatorname{rank}(Z) = 1\}$ where $\mathbb{H}_d$ is the set of all $d \times d$ Hermitian matrices and $\mathbf{1}_d \in \mathbb{C}^d$ is the vector with all coordinates equal to 1. It is therefore straightforward to construct a SDP relaxation of (4.14) by dropping the rank constraint. It appears that this relaxation is exact since, for $\mathcal{C} = \{Z \in \mathbb{H}_d : Z \succeq 0, \operatorname{diag}(Z) = \mathbf{1}_d\}$,

$$\operatorname*{argmax}_{Z\in\mathcal{C}} \langle \mathbb{E}A, Z \rangle = \{Z^*\} \tag{4.15}$$

where $Z^* = x^*(\overline{x^*})^\top$. A proof of (4.15) can be found in Section 6.2.3. Finally, as we only observe A, we consider the following SDP estimator of Z^*

$$\hat{Z} \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle A, Z \rangle. \tag{4.16}$$

In the next section, we use the strategy from Theorem 5 to obtain statistical guarantees for the estimation of Z^* by $\hat{Z}$.

Intuitively, the above maximization problem (4.15) attempts to preserve the given angle offsets as best as possible, by aiming to maximize the following objective function

$$\operatorname*{argmax}_{\theta_1,\dots,\theta_d\in[0,2\pi)} \sum_{i,j=1}^{d} e^{-\iota\theta_i} A_{ij} e^{\iota\theta_j}, \tag{4.17}$$

where the objective function value is incremented by $+1$ whenever an assignment of angles θ_i and θ_j perfectly satisfies the given edge constraint $\delta_{ij} = (\theta_i - \theta_j)[2\pi]$ (that is, for a *clean* edge for which $\sigma = 0$), while the contribution of an incorrect assignment (that is, of a *very noisy* edge) will be almost uniformly distributed on the unit circle in the complex plane. Due to non-convexity of optimization in (4.17), it is difficult to solve it computationally (Zhang & Huang, 2006); one way to overcome this problem is to consider the SDP relaxation from (4.15) such as in (4.16) but it is also possible to consider a spectral relaxation such as the one proposed by Singer (2011), which replaces the individual constraints that all z_i's should have unit magnitude by the much weaker single constraint $\sum_{i=1}^d |z_i|^2 = d$, leading to

$$\operatorname*{argmax}_{z_1,\dots,z_d\in\mathbb{C};\ \sum_{i=1}^d |z_i|^2=d} \sum_{i,j=1}^{d} \bar{z}_i A_{ij} z_j. \tag{4.18}$$

The solution to the resulting maximization problem is simply given by a top eigenvector of the Hermitian matrix A, followed by a normalization step. We remark that the main advantage of the SDP relaxation (4.15) is that it explicitly imposes the unit magnitude constraint for $e^{\iota\theta_i}$, which we cannot otherwise enforce in the spectral relaxation solved via the eigenvector method in (4.18) (at the end of the day, our estimator $\hat{x}$ from Corollary 32 below is a top eigenvector which may not satisfied the unit magnitude constraint). The above SDP program (4.15) is very similar to the well-known Goemans-Williamson SDP relaxation for the seminal MAX-CUT problem of finding the maximal cut of a graph (the MAX-CUT problem is considered in Section 4.5 below), with the main difference being that here we optimize over the cone of complex-valued Hermitian positive semidefinite matrices, not just real symmetric matrices.

4.4.2 Main results for phase recovery in the synchronization problem (in the multiplicative noise model)

Our main result concerns the estimation of the matrix of offsets $Z^* = x^*(\overline{x^*})^\top$ from the observation of the matrix A. This result is then used to estimate (up to a global phase shift) the angular vector $x^* = (e^{-\iota\theta_i})_{i=1}^d$. Our first result follows from Theorem 5.

Theorem 31. *Consider $0 < \epsilon < 1$. If $\sigma \leq \sqrt{\log(\epsilon d^4)}$ then, with probability at least $1 - \exp(-\epsilon\sigma^4 d(d-1)/2)$, it holds true that*

$$\frac{e^{-\frac{\sigma^2}{2}}}{2}\|Z^* - \hat{Z}\|_2^2 \leq \left\langle \mathbb{E}A, Z^* - \hat{Z} \right\rangle \leq \frac{128}{6}\sqrt{\epsilon}\sigma^4 d(d-1). \tag{4.19}$$

Once we have an estimator $\hat{Z}$ for the oracle Z^*, we can extract an estimator $\hat{x}$ for the vector of phases x^* by considering a top eigenvector (i.e. an eigenvector associated with the largest eigenvalue) of $\hat{Z}$. It is then possible to quantify the estimation properties of x^* by $\hat{x}$ using a 'sin-Theta' theorem and Theorem 31.

Corollary 32. *Let $\hat{x}$ be a top eigenvector of $\hat{Z}$ with Euclidean norm $\|\hat{x}\|_2 = \sqrt{d}$. Consider $0 < \epsilon < 1$ and assume that $\sigma \leq \sqrt{\log(\epsilon d^4)}$. We have the existence of a universal constant $c_0 > 0$ (which is the constant in the Davis-Kahan theorem for Hermitian matrices) such that, with probability at least $1 - \exp(-\epsilon\sigma^4 d(d-1)/2)$, it holds true that*

$$\min_{z\in\mathbb{C}:|z|=1} \|\hat{x} - zx^*\|_2 \leq 8c_0\sqrt{2/3}\epsilon^{1/4}e^{\sigma^2/4}\sigma^2\sqrt{d}. \tag{4.20}$$

It follows from Corollary 32 that we can estimate x^* (up to a global rotation $z \in \mathbb{C} : |z| = 1$) with a ℓ_2^d-estimation error of the order of $\sigma^2\sqrt{d}$ with exponential deviations. Given that $\|x^*\|_2 = \sqrt{d}$, this means that a constant proportion of the entries are well estimated when ϵ is taken like a constant. For a value of $\epsilon \sim 1/d^2$, the rate of estimation is like σ^2, we therefore get a much better estimation of x^* but only with constant probability. It is important to recall that $\hat{Z}$ and $\hat{x}$ can be both efficiently computed by solving a SDP problem and then by considering a top eigenvector of its solution (for instance, using the power method).

We finish the section on the angular group synchronization with the additive model as considered in A. BANDEIRA et al. (2016). Our aim is still to put forward our methodology and to show that it has a wide spectrum of applications and that, in particular, it covers also the model introduced in A. BANDEIRA et al. (2016).

4.4.3 The angular group synchronization model with additive noise from A. Bandeira et al. (2016)

As mentioned above, we chose to study a multiplicative noise model in A since it makes more sense from a physical point of view to have an additive noise on the offsets $\delta_{ij} = \theta_i - \theta_j[2\pi]$

(this additive noise becoming multiplicative by passing to the exponential in A). However, in A. Bandeira et al. (2016), the authors considered a model with additive noise on $Z^* = x^*(\overline{x^*})^\top$. In this 'additive' model, we observe $C = Z^* + \sigma W$, where W is a complex Wigner matrix and $\sigma > 0$ is the noise level. The MLE $\tilde{x}$ is solution to the problem

$$\tilde{x} \in \underset{x \in \mathbb{C}^d, |x_i|=1 \forall i}{\operatorname{argmax}} \; x^T C x, \tag{4.21}$$

which can be hard to compute in practice. Using the same approach as above, a SDP relaxation can be obtained by removing a rank one constraint, yielding the SDP estimator

$$\tilde{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmax}} \langle C, X \rangle, \tag{4.22}$$

where $\mathcal{C} := \{Z \in \mathbb{H}_d : \; \operatorname{diag}(Z) = 1_d \text{ and } Z \succeq 0\}$. Statistical properties of $\tilde{Z}$ have been obtained in A. Bandeira et al. (2016). We recall this result now.

Theorem 33 (Theorem 2.1 in A. Bandeira et al. (2016))**.** *Let $\tilde{x}$ be a solution of (4.21). Then with probability at least $1 - \mathcal{O}(d^{-\frac{5}{4}})$, $\min_{z \in \mathbb{C}:|z|=1} \|z\tilde{x} - x^*\|_2 \leq 12\sigma$. Moreover, if $\sigma \leq (1/18)d^{1/4}$, then (4.22) has a unique solution which is the rank one matrix $\tilde{x}\tilde{x}^\top$.*

Our methodology (here Theorem 5 is applied) may also be used to handle the 'additive' noise model from A. Bandeira et al. (2016). We consider the same SDP estimator $\tilde{Z}$ as defined in (4.22) and we obtain the following result (the proof is postponed in Section 7.5).

Theorem 34. *Let $\bar{x}$ be a top eigenvector of $\tilde{Z}$. With probability at least $1 - 5\exp(-d/2)$,*

$$\min_{z \in \mathbb{C}:|z|=1} \|z\bar{x} - x^*\|_2 \leq 40 c_0 \sigma$$

where c_0 is the constant appearing in the Davis-Kahan theorem for Hermitian matrices.

Compared with Theorem 2.1 from A. Bandeira et al. (2016), the estimation rate that we get for estimator $\bar{x}$ is the same (up to an absolute constant) as the one obtained for the MLE $\tilde{x}$ in Theorem 33, it is of the order of σ. Note however that our result for $\bar{x}$ holds without any restriction of the noise level σ, whereas in Theorem 33 one needs $\sigma \leq (1/18)d^{1/4}$ to get this result for $\bar{x}$. Note also that our result holds with exponentially large probability, whereas the one in Theorem 33 holds only with polynomial deviation. From a statistical perspective, our result improves the one from A. Bandeira et al. (2016). However, the main interest of Theorem 33 is not on the statistical performance of $\bar{x}$ but on the sharpness of the SDP relaxation since it shows that the SDP relaxation (4.22) is actually exact when $\sigma \leq (1/18)d^{1/4}$. This is a result that we do not have and that our methodology cannot obtain, since it is designed to prove only an estimation bound. But from a statistical point of view, it does not improve the estimation rate to know that the SDP relaxation is exact: our result shows that the SDP relaxation is doing as good as MLE, without proving that they are the same (up to global phase).

A proof of Theorem 34 is given in Section 6.2.3. We actually provide three estimation bounds for $\bar{x}$ in this proof. We are doing so because our aim is to show how a general methodology works in various examples. This methodology relies on the computation of a fixed point ($r^*_G(\delta)$ from (4) here). Hence, understanding how to bound this fixed point is part of the objective of this work. We therefore use the angular group synchronization problem with additive noise as a playground to show three different ways to upper bound such a fixed point. Using the three computations, we actually obtain the following three upper bounds

$$\min_{z \in \mathbb{C}:|z|=1} \|z\bar{x} - x^*\|_2 \leq \begin{cases} 8c_0\sigma\sqrt{d} & \text{with probability at least } 1 - \exp(-d^2/2) \\ c_0\sqrt{36K_G^{\mathbb{C}}\sigma d^{1/4}} & \text{with probability at least } 1 - \exp(-d/2) \\ 40c_0\sigma & \text{with probability at least } 1 - 5\exp(-d/2), \end{cases}$$

where c_0 is the constant appearing in the Davis-Kahan theorem for Hermitian matrices and $K_G^{\mathbb{C}}$ is Grothendieck constant in the complex case. Each of the three bounds above follows from different upper bounds on the complexity fixed point r_G^*. For instance, the second one follows from the 'global' approach, and the third one follows from a decomposition similar to the one from Fei and Chen (2019b) and is the one we used in Theorem 34.

4.5 The Max-Cut problem

4.5.1 Statistical setup

Let $A^0 \in \{0,1\}^{d\times d}$ be the adjacency (symmetric) matrix of an undirected graph $G = (V, E^0)$, where $V := \{1, \dots, d\}$ is the set of the vertices and the set of edges is $E^0 := E \cup E^\top \cup \{(i,i) : A^0_{ii} = 1\}$ where $E := \{(i,j) \in V^2 \ : \ i < j \text{ and } A^0_{ij} = 1\}$ and $E^\top = \{(j,i) : (i,j) \in E\}$. We assume that G has no self loop so that $A^0_{ii} = 0$ for all $i \in V$. A *cut* of G is any subset S of vertices in V. For a cut $S \subset V$, we define its weight by $\mathrm{cut}(G,S) := (1/2)\sum_{(i,j)\in S\times \bar{S}} A^0_{ij}$, that is the number of edges in E between S and its complement $\bar{S} = V\backslash S$. The Max-Cut problem is to find the cut with maximal weight

$$S^* \in \operatorname*{argmax}_{S\subset V} \mathrm{cut}(G,S). \tag{4.23}$$

The Max-Cut problem is a NP-complete problem, but Goemans and Williamson (1995) constructed a 0.878 approximating solution via a SDP relaxation. Indeed, one can write the Max-Cut problem in the following way. For a cut $S \subset V$, we define the membership vector $x \in \{-1,1\}^d$ associated with S by setting $x_i := 1$ if $i \in S$ and $x_i = -1$ if $i \notin S$ for all $i \in V$. We have $\mathrm{cut}(G,S) = (1/4)\sum_{i,j=1}^d A^0_{ij}(1 - x_i x_j) := \mathrm{cut}(G,x)$ and so solving (4.23) is equivalent to solving

$$x^* \in \operatorname*{argmax}_{x\in\{-1,1\}^d} \mathrm{cut}(G,x). \tag{4.24}$$

Since $(x_i x_j)_{i,j} = xx^\top$, the latter problem is also equivalent to solving

$$Z \in \operatorname*{argmax}_{Z\in\mathcal{C}} \left(\frac{1}{4}\sum_{i,j=1}^d A^0_{ij}(1 - Z_{ij})\right), \quad \mathcal{C} := \left\{Z \in \mathbb{R}^{d\times d} : Z \succeq 0, Z_{ii} = 1, \forall i = 1, \dots, d\right\} \tag{4.25}$$

which admits a SDP relaxation by removing the rank-1 constraint. This yields the following SDP relaxation problem of Max-Cut from Goemans and Williamson (1995)

$$Z^* \in \operatorname*{argmin}_{Z\in\mathcal{C}} \langle A^0, Z\rangle, \tag{4.26}$$

Unlike the other examples from the previous sections, the SDP relaxation in (4.26) is not exact, except for bipartite graphs (see Khot and Naor (2009) or Gärtner and Matoušek (2012) for more details). Nevertheless, thanks to the approximation result from Goemans and Williamson (1995), we can use our methodology to estimate Z^* and then deduce an approximate optimal cut. The Max-Cut problem is therefore a good setup for us to test our methodology in a context where the SDP relaxation is not exact, but still widely used in practice. Thus the type of question we want to answer here is: what can we say in a setup where only partial or noisy information is available on $\mathbb{E}[A]$, and when the SDP relaxation associated with $\mathbb{E}[A]$ is also not exact. This differs from the previous setup where exactness of the SDP relaxation holds, and this interesting peculiarity is one of the reasons why we have chosen to present this problem here. Our motivation stems from the observation that, in many situations, the adjacency matrix A^0 is only partially observed, but nevertheless, it might be

interesting to find an approximating solution to the MAX-CUT problem. Let us then introduce a stochastic model for the partial information available on $\mathbb{E}[A]$, the adjacency matrix here.

We observe $A = S \circ A^0 = (s_{ij} A^0_{ij})_{1 \leq i,j \leq d}$ a 'masked' version of A^0, where $S \in \mathbb{R}^{d \times d}$ is symmetric with upper triangular matrix filled with *i.i.d.* Bernoulli entries: for all $i, j \in V$ such that $i \leq j$, $S_{ij} = S_{ji} = s_{ij}$ where $(s_{ij})_{i \leq j}$ is a family of *i.i.d.* Bernoulli random variables with parameter $p \in (1/2, 1)$. Consider $B := -(1/p)A$, so that $\mathbb{E}[B] = -A^0$. We can write Z^* as an oracle since $Z^* \in \operatorname{argmax}_{Z \in \mathcal{C}} \langle \mathbb{E}B, Z^* - Z \rangle$ and so we estimate Z^* via the SDP estimator $\hat{Z} \in \operatorname{argmax}_{Z \in \mathcal{C}} \langle B, Z \rangle$. Our first aim is to quantify the cost we pay by using $\hat{Z}$ instead of Z^* in our final choice of cut. It appears that the fixed point used in Theorem 3 may be used to quantify this loss

$$r^*(\Delta) = \inf \left(r > 0 : \mathbb{P} \left[\sup_{Z \in \mathcal{C} : \langle \mathbb{E}B, Z^* - Z \rangle \leq r} \langle B - \mathbb{E}B, Z - Z^* \rangle \leq (1/2) r \right] \geq 1 - \Delta \right). \tag{4.27}$$

Our second result is an explicit high-probability upper bound on the latter fixed point.

4.5.2 Main results for the Max-Cut problem

In this section, we gather the two results on the estimation of Z^* from $\hat{Z}$ and on the approximate optimality of the final cut constructed from $\hat{Z}$. Let us now explicitly provide the construction of this cut. We consider the same strategy as in GOEMANS and WILLIAMSON (1995). Assume that $\hat{Z}$ has been constructed. Let $\hat{G}$ be a centered Gaussian vector with covariance matrix $\hat{Z}$. Let $\hat{x}$ be the sign vector of $\hat{G}$. Using the statistical properties of $\hat{Z}$, it is possible to prove near optimality of $\hat{x}$.

We denote the optimal values of the MAX-CUT problem associated with the graph G and its SDP relaxation by

$$\mathrm{SDP}(G) := (1/4) \langle A^0, J - Z^* \rangle = \max_{Z \in \mathcal{C}} \frac{1}{4} \sum_{i,j} A^0_{i,j} (1 - Z_{ij}) \text{ and } \textsc{Max-Cut}(G) := \mathrm{cut}(G, S^*),$$

where S^* is a solution of (4.23) and $J = (1)_{d \times d}$. Our first result is to show how the 0.878-approximation result from GOEMANS and WILLIAMSON (1995) is downgraded by the incomplete information we have on the graph (since we only partially observed the adjacency matrix A^0 via the masked matrix A).

Theorem 35. *For all $0 < \Delta < 1$. With probability at least $1 - \Delta$ (with respect to the masked S), it holds true that*

$$\mathrm{SDP}(G) \geq \mathbb{E}\left[\mathrm{cut}(G, \hat{x}) | \hat{Z} \right] \geq 0.878 \mathrm{SDP}(G) - \frac{0.878 r^*(\Delta)}{4}.$$

To make the notation more precise, $\hat{x}$ is the sign vector of $\hat{G}$ which is a centered Gaussian variable with covariance $\hat{Z}$. In that context, $\mathbb{E}\left[\mathrm{cut}(G, \hat{x}) | \hat{Z} \right]$ is the conditional expectation according to $\hat{G}$ for a fixed $\hat{Z}$. Moreover, the probability 'at least $1 - \Delta$' that we obtain is with respect to the random mask, that is to the randomness in A.

Let us now frame Theorem 35 into the following perspective. If we had known the entire adjacency matrix (which is the case when $p = 1$), then we could have used Z^* instead of $\hat{Z}$. In that case, for $x^\star$ the sign vector of $G^\star \sim \mathcal{N}(0, Z^*)$, we know from GOEMANS and WILLIAMSON (1995) that

$$\mathrm{SDP}(G) \geq \mathbb{E}\left[\mathrm{cut}(G, x^\star) \right] \geq 0.878 \mathrm{SDP}(G). \tag{4.28}$$

Therefore, from a trade-off perspective, Theorem 35 characterizes the price we pay for not observing the entire adjacency matrix A^0, but only a masked version A of it. It is an interesting output of Theorem 35 to observe that the fixed point $r^*(\Delta)$ measures, in a quantitative way, this loss. If we were able to identify scenarios of p and d for which $r^*(\Delta) = 0$, that would prove that there is no loss for partially observing A^0 in the MAX-CUT problem. The approach we use to control $r^*(\Delta)$ is the global one, which does not allow for exact reconstruction (that is, to show that $r^*(\Delta) = 0$).

Let us now turn to an estimation result of Z^* by $\hat{Z}$ via an upper bound on $r^*(\Delta)$.

Theorem 36. *There exists a universal constant $C > 0$ such that With probability at least $1 - 4^{-d}$:*

$$\left\langle \mathbb{E}B, Z^* - \hat{Z} \right\rangle \leq r^*(4^{-d}) \leq C \left(2d \sqrt{\frac{(2\log 4)(1-p)(d-1)}{p}} + \frac{8d\log 4}{3} \right).$$

In particular, it follows from the approximation result from Theorem 35 and the high-probability upper bound on $r^*(\Delta)$ from Theorem 36 that, with probability at least $1 - 4^{-d}$

$$\mathbb{E}\left[\text{cut}(G, \hat{x}) | \hat{Z} \right] \geq 0.878\text{SDP}(G) - \frac{0.878}{4} C \left(2d \sqrt{\frac{(2\log 4)(1-p)(d-1)}{p}} + \frac{8d\log 4}{3} \right) \quad (4.29)$$

This result is non-trivial only when the right-hand side term is strictly larger than $0.5 \cdot \text{SDP}(G)$, which is the performance of a random cut. As a consequence, (4.29) shows that one can still do better than randomness even in an incomplete information setup for the MAX-CUT problem when p, d, and $\text{SDP}(G)$ are such that

$$0.378\text{SDP}(G) > \frac{0.878}{4} C \left(2d \sqrt{\frac{(2\log 4)(1-p)(d-1)}{p}} + \frac{8d\log 4}{3} \right).$$

For instance, when p is like a constant, it requires $\text{SDP}(G)$ to be larger than $c_0 d^{3/2}$ (for some absolute constant c_0) and when $p = 1 - 1/d$, it requires $\text{SDP}(G)$ to be at least $c_0 d$ (for some absolute constant c_0).

Remark 4. To get exact recovery, that is $r^*(\Delta) = 0$, in the MAX-CUT problem (which shows that there is no loss for the MAX-CUT problem by observing only a masked version of the adjacency matrix), we have to develop a local approach, as for the Signed Clustering and the Group Synchronization problems. To that end, we would need to solve two problems: firstly, find a curvature for the objective function $Z \to \langle \mathbb{E}B, Z^* - Z \rangle$, and secondly, study the oscillations of the empirical process $Z \to \langle \mathbb{E}B - B, Z^* - Z \rangle$. We leave those two difficult problems for future research.

4.6 The sparse PCA problem

Principal Components analysis (PCA) is one of the most fundamental dimension reduction algorithm as well as one of the most used data visualization tool. It can be efficiently performed via some truncated SVD algorithms on the $N \times d$ data matrix (N being the number of data and d the dimension of the data, that is the number of features) which requires only $\mathcal{O}(k^2 \min(d, N))$ operations to get the first k top eigenvectors (GOLUB & VAN LOAN, 2013; HALKO, MARTINSSON, & TROPP, 2011).

However, principal components are linear mixture of features that may be of very different nature and as so are for most of the time meaningless. This problem becomes more salient for

high-dimensional data (i.e. when $d > N$) where the diversity of features (text, socio-professional categories, geographic location, familiar situation, consumption habits, etc.) may be very large. Moreover, in the high-dimensional setting, PCA no longer provides meaningful estimates of the principal components of the actual covariance matrix Σ as exhibited by the phase transition from Baik, Ben Arous, and Péché (2005).

One way to alleviate both interpretation and inconsistency in the high-dimensional setting is to look for principal components which are linear mixture of a small number of features – that is "sparse" principal component. This problem is known as sparse PCA and was introduced in Johnstone and Lu (2009b) and Johnstone and Lu (2009a). It can be stated as the following optimization problem:

$$\hat{v}_1 \in \underset{\|v\|_2=1, \|v\|_0 \leq k}{\operatorname{argmax}} \|\hat{\Sigma}_N v\|_2, \tag{4.30}$$

where the X_i's are *i.i.d.* centered vectors in $\mathbb{R}^d$ with covariance $\mathbb{E}[X_i X_i^\top] = \Sigma$, $\hat{\Sigma}_N = (1/N)\sum_{i=1}^N (X_i - \bar{X}_N)(X_i - \bar{X}_N)^\top$ is the empirical covariance matrix, $\|v\|_0$ is the size of the support of v and k is some fixed sparsity level.

From an algorithmic point of view there are two major issues in the optimization problem (4.30):

1. the objective function that we want to maximize is convex; and it is notoriously difficult to maximize a convex function even on a convex set
2. because of the sparsity constraint '$\|v\|_0 \leq k$', the constraint set is not convex.

If the sparsity constraint was not there, then (4.30) would be the classical PCA problem for finding a first principal component, that is a top eigenvector of $\hat{\Sigma}_N$. In that case, even though it is a maximization problem of a convex function on a convex set, this problem can be solved efficiently for instance via the power method and is in fact one of the few situation where maximizing a convex function can be performed efficiently.

The extra sparsity constraint in (4.30) somehow emphasis this original issue that the objective function to maximize is convex. One way to overcome this issue is to adapt the power method to this extra constraint, see Journée, Nesterov, Richtárik, and Sepulchre (2010). Another way is via SDP relaxation (d'Aspremont et al., 2007). We will use this latter approach so we present it in the next subsection in more details.

4.6.1 SDP relaxation in sparse PCA

Let $X \in \mathbb{R}^d$ be a centered random vector with distribution P. Let $X_1, \ldots, X_N \in \mathbb{R}^d$ be N independant copies of X. Define $A := (1/N)\sum_{i=1}^N X_i X_i^\top$, the empirical covariance matrix of the X_i's. Let $\Sigma := \mathbb{E}[A] = \mathbb{E}_{X\sim P}[XX^T]$ be their covariance matrix. We are looking for a first principal component with a support of small cardinality, that is for a vector $v^* \in \mathbb{R}^d$ with unit-length and cardinality less than a certain integer $k \leq d$, and such that the variance of the X_i's when projected onto v^* is maximal. This can be written as follows:

$$v^* \in \underset{v\in\mathcal{E}}{\operatorname{argmax}} \mathbb{E}[\langle X, v\rangle^2] \text{ where } \mathcal{E} := \left\{ v \in \mathbb{R}^d : \|v\|_2 = 1, \|v\|_0 \leq k \right\}. \tag{4.31}$$

This problem is known to be NP-hard in general (Magdon-Ismail, 2015), so we are looking to relax it. One way to do this is to replace the cardinality function by the ℓ_1^d-norm. Another way is via the lifting procedure, which is described for example in Lemaréchal and Oustry (2018) and is based on the principle that quadratic objective functions and constraint sets of a vector v can be written as linear objective functions and constraint sets of the symmetric rank one matrix $vv^\top$.

In our case, we first note that $\mathbb{E}[\langle X, v\rangle^2] = \langle \mathbb{E}[A], vv^\top\rangle = \langle \Sigma, vv^\top\rangle$. Then, if $Z = vv^T$ with $v \in S_2^{d-1}$ and $\|v\|_0 \leq k$, we have $\mathrm{Tr}(Z) = \|v\|_2^2 = 1$ and $\|Z\|_0 \leq k^2$. Finding a solution of (4.31) is then equivalent (D'Aspremont et al., 2007; Lemaréchal & Oustry, 2018) to finding a top singular vector of $Z^\star$, where $Z^\star$ is solution of the optimization problem

$$Z^\star \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle \mathbb{E}[A], Z\rangle \text{ where } \mathcal{C}_0 := \left\{ Z \in \mathbb{R}^{d\times d} : Z = vv^T, v \in \mathbb{R}^d, \mathrm{Tr}(Z) = 1, \|Z\|_0 \leq k^2 \right\}.$$

In the latter problem, the objective function has now become a linear one thanks to the lifting approach, however the constraint set is not convex. We are now working on that issue to get a full SDP relaxation of (4.31). First, we may replace the condition "$Z = vv^T$" by the equivalent condition "$Z \succeq 0$ and $\mathrm{rank}(Z) = 1$" in $\mathcal{C}_0$.
However, $\mathcal{C}_0 := \left\{ Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \mathrm{Tr}(Z) = 1, \|Z\|_0 \leq k^2, \mathrm{rank}(Z) = 1 \right\}$ is not convex, because of two non-convex constraints: *the cardinality constraint* "$\|Z\|_0 \leq k^2$" and *the rank constraint* "$\mathrm{rank}(Z) = 1$" that we are just dropping out of $\mathcal{C}_0$. By doing so, we end up with the following convex optimization problem:

$$Z^* \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle \mathbb{E}[A], Z\rangle \text{ where } \mathcal{C} := \{Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \mathrm{Tr}(Z) = 1\}. \tag{4.32}$$

We then see Z^* as an oracle for the linear loss function $Z \to \ell_Z(X) = -\langle XX^\top, Z\rangle$ and its associated risk function $Z \to \mathbb{E}\ell_Z(X)$ over the model $\mathcal{C}$, that is $Z^* \in \mathrm{argmin}_{Z\in\mathcal{C}} P\ell_Z$. This enables us to leverage the methodological tools introduced in Section 3 to derive estimators for Z^* and provide statistical guarantees onto them.

This configuration allows us to refer to the work of Wang, Berthet, and Samworth (2016). The authors study the sparse PCA problem where the distribution of the data $X_1, \ldots, X_N$ belongs to a class $\mathcal{P}$ of distributions that all have a sub-exponential tail; it includes, among others, sub-Gaussian distributions (see equation (4) in Wang et al. (2016) for a definition). In particular, they propose the following ℓ_1-regularized ERM estimator

$$\hat{Z} \in \operatorname*{argmin}_{Z\in\mathcal{C}} \left(\langle \frac{-1}{N} \sum_{i=1}^N X_i X_i^\top, Z\rangle + \lambda \|Z\|_1 \right) \text{ where } \mathcal{C} := \left\{ Z : Z \succeq 0, \mathrm{Tr}(Z) = 1 \right\} \tag{4.33}$$

and provide an algorithm for solving it in polynomial time. We report below their main results for this estimator.

Theorem 37. *[Theorem 5 in Wang et al. (2016)] Let $X_1, \ldots, X_N \in \mathbb{R}^d$ be i.i.d. random vectors with distribution in $\mathcal{P}$ and a covariance matrix satisfying the spiked covariance model: $\mathbb{E}[X_i X_i^\top] = I_d + \theta\beta^*(\beta^*)^\top$, where β^* is a k-sparse vector with unit euclidean norm. Let $\lambda = 4\sqrt{\log(d)/N}$, $\epsilon = \log(d)/(4N)$ and consider $\hat{v}_{\lambda,\epsilon} \in \mathrm{argmax}_{\|v\|_2=1} v^\top \hat{Z}^\epsilon v$, where $\hat{Z}^\epsilon$ is an ϵ-maximizer of $Z \to \langle \frac{1}{N}\sum_{i=1}^N X_i X_i^\top, Z\rangle - \lambda\|Z\|_1$ over the model $\mathcal{C}$ defined in (4.33). Finally, let $\hat{v}^0_{\lambda,\epsilon}$ be the k-sparse vector derived from $\hat{v}_{\lambda,\epsilon}$ by setting all but its largest k coordinates in absolute value to 0. If $4\log(d) \leq N \leq k^2 d^2 \theta^{-2}$ and $0 < \theta \leq k$, then it holds true that:*

$$\mathbb{E}\left[\sqrt{2} \|\hat{v}^0_{\lambda,\epsilon}(\hat{v}^0)^\top_{\lambda,\epsilon} - \beta^*(\beta^*)^\top\|_2 \right] \leq (32\sqrt{2} + 3)\sqrt{\frac{k^2 \log(d)}{N\theta^2}}.$$

We are now using our methodology to propose several estimators and provide our insights on the sparse PCA problem. In particular, we will extend Theorem 37 to the heavy-tailed framework, provide in-deviation results and improve the rate to the optimal one $k^2\log(ed/k)/N$ (thanks to localization). On top of that, we will construct new estimators based on the MOM principle to handle robustness issues in sparse PCA.

4.6.2 Exactness and curvature in the spiked covariance model.

We present here two results that will be of crucial importance in the analysis of our estimators (the proofs are postponed to Section 6.2.5). The first one concerns the exactness in the spiked covariance model. That is, the oracle Z^* as defined by equation (4.32), obtained after a lifting and a convex relaxation of the initial problem, turns out to be a matrix of rank one whose unit-norm leading eigenvector is $\pm\beta^*$.

Lemma 38. *In the spiked covariance model $\Sigma = \theta(\beta^*)(\beta^*)^\top + I_d$ with $\beta^* \in S_2^{d-1}$ and β^* is k-sparse, we have $Z^* = (\beta^*)(\beta^*)^\top$, for Z^* defined in (4.32).*

The second one concerns the curvature of the excess risk function around the oracle Z^*. Following our methodology, we need to understand the behavior of the excess risk around Z^* in order to find a good G function that will be used to be define localized subsets of our model. Then, later, based on the results from Section 4.1 we will compute the Rademacher complexities of these localized subsets and then the local complexity fixed points as introduced in Section 3. The fixed point is then used to establish statistical bounds on our estimators. Finding the 'right' curvature function of the excess risk is therefore important in our approach. The following result provides a curvature of the excess risk 'globally', that is on the entire set $\mathcal{C}$ and not just around Z^* (see the proof in Section 39).

Lemma 39. *In the spiked covariance model $\Sigma = \theta(\beta^*)(\beta^*)^\top + I_d$ with $\beta^* \in S_2^{d-1}$ and β^* is k-sparse, the following holds. For all $Z \in \mathcal{C}$, we have $P\mathcal{L}_Z = \langle \Sigma, Z^* - Z \rangle \geq (\theta/2)\|Z^* - Z\|_2^2$.*

As a consequence, using our terminology, the problem has an excess risk curvature function given by $G : Z \to \|Z\|_2^2$ - where $\|\cdot\|_2$ is the canonical Hilbertian norm in $\mathbb{R}^{d\times d}$. We will therefore use the ℓ_2-norm to the square to define our localized models for the study of all estimators introduced below.

4.6.3 ℓ_1-Regularized ERM estimator

Since the parameter we want to estimate has a sparse structure, the choice of estimators regularized by an appropriate norm will enable us to take advantage of this structural property. We start with a regularized ERM estimator, as presented in Section 3.2.2, where the ℓ_1-norm is used as regularization norm:

$$\hat{Z}_\lambda^{\mathrm{RERM}} \in \operatorname*{argmin}_{Z\in\mathcal{C}} \left(P_N\ell_Z + \lambda\|Z\|_1\right), \quad \text{where} \quad \mathcal{C} := \left\{Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \mathrm{Tr}(Z) = 1\right\} \tag{4.34}$$

and $\ell_Z(X) = -\langle XX^\top, Z\rangle$ and $P_N\ell_Z = (1/N)\sum_{i=1}^N \ell_Z(X_i)$. This puts us in condition to use the results of Section 3.2.2 to provide statistical guarantees on $\hat{Z}_\lambda^{\mathrm{RERM}}$.

Lemma 39 shows that, for any value of $\rho > 0$ and $\delta \in (0,1)$, Assumption 3.3 is satisfied with $A = 2/\theta$ and $G : Z \in \mathbb{R}^{d\times d} \to \|Z\|_2^2$. In order to proceed with our methodology, the next step is then to identify a value of ρ^* which satisfies the $2/\theta$-sparsity equation from Definition 8. This is the purpose of the following Lemma (the proof is given in section 6.2.5.3).

Lemma 40. *Let $A > 0$, $\delta \in (0,1)$, and define $r^*(.) := r^*_{\mathrm{RERM,G}}(A, ., \delta)$. If $\rho \geq 10k\sqrt{r^*(\rho)}$, then ρ satisfies the A-sparsity equation from Definition 8.*

The last step is to compute the local complexity fixed point of Definition 7, which is what we are working on below.

Lemma 41. *Grant Assumption 4.1 with $t = \log(ed/10k)$. Suppose that β^* is k-sparse, with $k \leq ed/200$. Let $A = 2/\theta$ and assume that $N \geq 3\log\left(ed/10k\right)$. Then there exists an absolute*

constant $b > 0$ such that, defining:

$$\rho^* := 200bAk^2\sqrt{\frac{1}{N}\log\left(\frac{ed}{k}\right)} \text{ and } r^*(\rho) := bA\sqrt{\frac{\rho^2}{N}\log\left(\frac{b^2A^2(ed)^4}{N\rho^2}\right)}, \tag{4.35}$$

one has $r^*_{\mathrm{RERM,G}}\left(A, \rho^*, 10k/ed\right) \leq r^*(\rho^*)$ and ρ^* satisfies the A-sparsity equation from Definition 8.

We are now ready to state our main result concerning the ℓ_1-regularized ERM estimator for the sparse PCA problem.

Theorem 42. *Grant Assumption 4.1 with $t = \log(ed/10k)$. Suppose that β^* is k-sparse, with $k \leq ed/200$. Assume that $N \geq 3\log\left(ed/10k\right)$ and that λ satisfies the following inequalities:*

$$\frac{20}{21}b\sqrt{\frac{1}{N}\log\left(\frac{ed}{200^{1/2}\log(200)^{1/4}k}\right)} \leq \lambda \leq \frac{2}{\sqrt{3}}b\sqrt{\frac{1}{N}\log\left(\frac{ed}{200^{2/3}k}\right)} \tag{4.36}$$

where b is the absolute constant introduced in Lemma 41 above. Let $C = 40b$. Then, with probability at least $1 - 10k/ed$, it holds true that:

$$\|\hat{Z}^{\mathrm{RERM}}_\lambda - Z^*\|_1 \leq 10Ck^2\sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)}, \quad \|\hat{Z}^{\mathrm{RERM}}_\lambda - Z^*\|_2 \leq C\sqrt{\frac{k^2}{N\theta^2}\log\left(\frac{ed}{k}\right)}$$

and

$$P\mathcal{L}_{\hat{Z}^{\mathrm{RERM}}_\lambda} \leq \frac{C^2}{2}\frac{k^2}{N\theta}\log\left(\frac{ed}{k}\right).$$

Note that if one is willing to get a better deviation parameter, one can assume N larger than $\Upsilon\log(ed/10k)$, for Υ large enough.

Up to this point, we have introduced an estimator for Z^* and provided a convergence rate with high probability. However, our primary focus is not on Z^* itself, but rather on its unit-norm leading eigenvectors $\pm\beta^*$. The purpose of the upcoming result is to leverage the preceding one in order to establish properties related to β^*.

Corollary 43. *Let $\hat{\beta} \in \mathbb{R}^d$ be a leading unit length eigenvector of $\hat{Z}^{\mathrm{RERM}}_\lambda$. Under the conditions of Theorem 42, there exists an absolute constant $D > 0$ such that with probability at least $1 - 10k/ed$:*

$$\|\hat{\beta}\hat{\beta}^\top - \beta^*(\beta^*)^\top\|_2 \leq D\sqrt{\frac{k^2}{N\theta^2}\log\left(\frac{ed}{k}\right)}.$$

We therefore obtain a convergence rate of magnitude $\left(k^2\log\left(ed/k\right)/(N\theta^2)\right)^{1/2}$, when our dataset is made up of $i.i.d$ random variables whose distribution satisfies Assumption 4.1, which includes the case of $i.i.d$ sub-Gaussian variables but it goes much beyond up to variables with only $\log d$ moments. The result of Wang et al. (2016) is available for a class of distributions, including sub-Gaussian distributions, whose covariance matrix fits within the spiked covariance model. They obtain a convergence rate of magnitude $\left(k^2\log(d)/(N\theta^2)\right)^{1/2}$, although our result holds with polynomial deviation while theirs is in expectation. We also note that our result does not suffer from any restrictive condition concerning θ. We therefore slightly improve the results from Wang et al. (2016); this improvement is of the same order as the one obtained

for the LASSO in Bellec, Lecué, and Tsybakov (2018) and is due to a careful localization argument. This shows that our analysis is precise enough to catch the subtle difference between the $\log d$ rate from Wang et al. (2016) and the $\log(ed/k)$ obtained in Theorem 42. Our result also extends the scope of Theorem 37 to heavy-tailed data since we only require the existence of $\log d$ moments. However, to get this improvement for the Lasso type estimator (4.37), one needs to choose λ depending on k in (4.36), which is unknown in practice. To solve this issue, we could use a Lepskii's adaptation scheme as in Bellec et al. (2018). However, we will not follow this path but rather consider another regularization norm: the $SLOPE$ norm, that allows to get the same results as in Theorem 42 but a choice of λ independent of k. This will also give us the opportunity to run our methodology one more time for a different regularization norm.

4.6.4 *SLOPE* regularized ERM estimator

In this section, we study a regularized ERM estimator of Z^* with the $SLOPE$ norm (introduced in Section 4.1.3, and whose definition is restated below) as the regularization norm. We consider a lexicographical order over $[d]^2$ such that for any $k \in [d]$, the k^2 largest elements in $[d]^2$ belong to $[k]^2$. We fix $t > 0$ (which will be choosen appropriately later) and we define, for $p \leq q$, $b_{pq}(t) =: \sqrt{\log(ed^2/pq) + t}$, and $b_{pq}(t) = b_{qp}(t)$ for $p > q$. For $Z \in \mathbb{R}^{d\times d}$, we define $Z^\sharp$ the matrix obtained from Z by reordering its element in absolute value in non-increasing order, and we finally define its $SLOPE$ norm by:

$$\|Z\|_{SLOPE} := \sum_{p,q=1}^{d} b_{pq} Z^\sharp_{pq}.$$

Our estimator is then:

$$\hat{Z}^{\mathrm{RERM}}_{SLOPE} \in \operatorname*{argmin}_{Z\in\mathcal{C}} \left(P_N \ell_Z + \lambda \|Z\|_{SLOPE}\right) \text{ for } \mathcal{C} := \{Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \mathrm{Tr}(Z) = 1\} \tag{4.37}$$

and a regularization parameter $\lambda > 0$ to be chosen later. This puts us in condition to use the results of Section 3.2.2 to provide statistical guarantees on $\hat{Z}^{\mathrm{RERM}}_{SLOPE}$.

As before, the essence of Lemma 39 in this context is that, for any value of $\rho > 0$ and $\delta \in]0, 1[$, Assumption 3.3 is satisfied with $A = 2/\theta$ and $G : Z \in \mathbb{R}^{d\times d} \to \|Z\|_2^2$. In order to proceed with our methodology, our next step is then to identify a value of ρ^* which satisfies the $2/\theta$-sparsity equation. This is the purpose of the following Lemma.

Lemma 44. *Assume that β^* is k-sparse, for some $k \in [d]$. Let $A > 0$, $\delta \in (0, 1)$ and $t > 0$. Define $\Gamma_k(t) := 3\sum_{\ell=1}^{k} b_{\ell\ell}(t)$. If $\rho \geq 10\Gamma_k(t)\sqrt{r^*_{\mathrm{RERM,G}}(A, \rho, \delta)}$, then ρ satisfies the A-sparsity equation from Definition 8.*

Following the path traced by our methodology, all that remains is to calculate the complexity fixed-point parameter

$$r^*_{\mathrm{RERM,G}}(A, \rho, \delta) = \inf\left(r > 0 : \mathbb{P}\left(\sup_{Z\in\mathcal{C}:\|Z-Z^*\|_{SLOPE}\leq\rho, \|Z-Z^*\|_2\leq\sqrt{r}} |(P - P_N)\mathcal{L}_Z| \leq \frac{r}{3A}\right) \geq 1 - \delta\right).$$

The next Lemma gives us an upper bound for $r^*_{\mathrm{RERM,G}}(A, \rho, \delta)$, when ρ satisfies the sparsity equation of Definition 8.

Lemma 45. *Grant Assumption 4.2 for $t = 2\log(ed^2/k^2)$. Suppose that β^* is k-sparse, with $k \leq d/(e^2\log(d))$. Let $A > 0$, and assume that $N \geq 3\log(ed^2)$. Then, there exists an absolute constant $b > 0$ such that, defining:*

$$\rho^* := 10\Gamma_k^* \frac{bA}{\sqrt{N}} \min\left(10\Gamma_k^*; d\right) \quad \text{and} \quad r^* := \frac{b^2A^2}{N} \min\left(10\Gamma_k^*; d\right)^2$$

one has $r^*_{\mathrm{RERM,G}}(A, \rho^*, 2k^2/(ed^2)) \leq r^*$ and ρ^* satisfies the A-sparsity equation rom Definition 8, where $\Gamma_k^* = \Gamma_k(2\log(ed^2/k^2))$ is the quantity introduced in Lemma 44.

We are now ready to state our main result concerning the $SLOPE$ regularized ERM estimator for the sparse PCA problem.

Theorem 46. *Grant Assumption 4.2 for $t = 2\log(ed^2/k^2)$. Suppose that β^* is k-sparse, with $k \leq \min\left(d/(e^2\log(d)), (e/140\sqrt{2})^2 d\right)$. Assume that $N \geq 3\log(ed^2)$ and that λ satisfies the following inequalities:*

$$\frac{10b}{21\sqrt{N}} < \lambda < \frac{2b}{3\sqrt{N}}, \tag{4.38}$$

where b is the constant previously defined in Lemma 45. Then there exist an absolute constants $C_1 > 0$ such that one has with probability at least $1 - 2k^2/(ed^2)$:

$$\|\hat{Z}^{\mathrm{RERM}}_{SLOPE} - Z^*\|_{SLOPE} \leq C_1 \frac{k^2}{\sqrt{N\theta^2}} \log\left(\frac{ed^2}{k^2}\right),$$

$$\|\hat{Z}^{\mathrm{RERM}}_{SLOPE} - Z^*\|_2 \leq C_1 \sqrt{\frac{k^2}{N\theta^2} \log\left(\frac{ed^2}{k^2}\right)}$$

$$\text{and} \quad \left\langle \Sigma, Z^* - \hat{Z}^{\mathrm{RERM}}_{SLOPE} \right\rangle \leq C_1 \frac{k^2}{N\theta} \log\left(\frac{ed^2}{k^2}\right).$$

We can now use this result to obtain properties about our object of interest, which is not directly Z^*, but its unit-length leading eigenvectors $\pm\beta^*$.

Corollary 47. *Let $\hat{\beta} \in \mathbb{R}^d$ be a leading unit-eigen vector of $\hat{Z}^{\mathrm{RSLOPE}}_{\lambda}$. Under the conditions of Theorem 46, there exists an absolute constant $C > 0$ such that with probability at least $1 - 2k^2/ed^2$:*

$$\|\hat{\beta}\hat{\beta}^\top - \beta^*(\beta^*)^\top\|_2 \leq C\sqrt{\frac{k^2}{N\theta^2} \log\left(\frac{ed^2}{k^2}\right)}.$$

Here again, we obtain a rate of convergence of magnitude $\sqrt{(1/N\theta^2)\log\left(ed^2/k^2\right)}$, holding with polynomial deviation, with no restriction on the value of θ. We note that this result holds with a value of the regularization parameter λ that does not depend on the sparsity level k of β^*.

4.6.5 ℓ_1 regularized minmax MOM estimator.

Here, we consider the case where data may be corrupted with outliers. We place ourselves in the framework of the adversarial contamination, which is described in Assumption 3.4: the dataset $\{X_1, \ldots, X_N\}$ used by the statistician may have been corrupted by an adversary. As a consequence, on top of the structural learning problem, we now have to face a robustness to data contamination problem. To deal with these issues all together, we use a regularized minmax MOM estimator.

We therefore consider an equi-partition of $\{1,\dots,N\}$ into $B_1 \sqcup \cdots \sqcup B_K = [N]$, where $|B_k| = N/K$ for all $k \in [K]$. We consider a ℓ_1-regularized minmax MOM estimator

$$\hat{Z}^{RMOM}_{K,\lambda} \in \operatorname*{argmin}_{Z\in\mathcal{C}} \sup_{Z'\in\mathcal{C}} \left(\mathrm{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\|_1 - \|Z'\|_1)\right),$$

for $\mathcal{C} := \left\{Z \in \mathbb{R}^{d\times d} : 0 \preceq Z \preceq I_d, \mathrm{Tr}(Z) = 1\right\}$ and a regularization parameter λ to be chosen later.

In what follows, we provide some statistical guarantees on $\hat{Z}^{RMOM}_{K,\lambda}$ based on Theorem 24 which is our general result for regularized minmax MOM estimators for a general G function used for localization. Here, following Lemma 39, we will use $G : Z \to (\theta/2)\|Z\|_2^2$ (and $A = 1$) for such a localization function. Following our methodology, once the curvature of the excess risk is chosen, we have to find an upper bound on the local complexity fixed point $r^*_{\mathrm{RMOM},G}(\gamma,\rho)$ from Definition 22. But before that we find a sufficient condition on a radius ρ so that it satisfies the sparsity equation from Definition 17.

Lemma 48. *Consider $\gamma > 0$. If $\rho > 0$ is such that $\rho \geq 10k\sqrt{2/\theta}r^*_{\mathrm{RMOM},G}(\gamma,\rho)$, then ρ satisfies the sparsity equation from Definition 17.*

Now that we know how to grasp a value of ρ that satisfies the sparsity equation, the subsequent task is to compute the fixed-point parameter $r^*_{\mathrm{RMOM},G}(\gamma,\rho)$ as introduced in Definition 22, after which, thanks to Theorem 18, we will be able to provide some statistical bounds on $\hat{Z}^{RMOM}_{K,\lambda}$.

Lemma 49. *Grant assumption 4.1 for $t = 1$. Suppose that β^* is k-sparse, for some $k \in [d]$. Assume that $N \geq 2\log(ed/k)+1$ and that $\theta \leq k$. Define $G : Z \in \mathbb{R}^{d\times d} \to (\theta/2)\|Z\|_2^2$. Consider $\gamma > 0$. There exist absolute constants B and $D > 0$ such that, defining:*

$$\rho^*(\gamma) := \max\left(\sqrt{480}B\frac{k^2}{\gamma}\sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)}; 10Dk\sqrt{\frac{2K}{N\theta^2}}\right)$$

$$\text{and}\quad r^*(\gamma,\rho) := \max\left(\sqrt{\frac{B\rho}{\gamma}}\left(\frac{6}{N}\log\left(\frac{2B(ed)^2}{\gamma\theta\rho}\sqrt{\frac{6}{N}}\right)\right)^{1/4}; D\sqrt{\frac{K}{N\theta}}\right),$$

one has $r^*_{\mathrm{RMOM,G}}(\gamma,\rho^*(\gamma)) \leq r^*(\gamma,\rho^*(\gamma))$ and $\rho^*(\gamma)$ satisfies the sparsity equation from Definition 17. The values of B and D are explicited in Section 6.2.5.12.

We are now ready to state our main result about the ℓ_1-Regularized MOM estimator for the sparse PCA problem.

Theorem 50. *Grant assumption 4.1 for $t = 1$. Suppose that β^* is k-sparse, for some $k \in [d]$. Assume that $N \geq 2\log(ed/k) + 1$ and let K be a divisor of N such that $K \geq 100|\mathcal{O}|$. Let $\gamma = 1/32000$ and $\lambda := 11r^*(\gamma, 2\rho^*(\gamma))/(40\rho^*(\gamma))$, where $r^*(.,.)$ and $\rho^*(.)$ are defined in Lemma 49 above. Then, there exists positive constants C_1, C_2 and C_3 such that, with probability at least $1 - \exp(-72K/625)$, it holds true that:*

$$\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_1 \leq \frac{C_1 k}{\sqrt{N\theta^2}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right),$$

$$\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_2 \leq \frac{C_2}{\sqrt{N\theta^2}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right)$$

$$\text{and}\quad P\mathcal{L}_{\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}} \leq \frac{C_3}{N\theta}\max\left(k^2\log\left(\frac{ed}{k}\right); K\right).$$

Since our primary focus is not on Z^* itself, but its unit-norm leading eigenvector β^*, we are now in the process of providing a result on β^*.

Corollary 51. *Let $\hat{\beta} \in \mathbb{R}^d$ be a leading unit length eigenvector of $\hat{Z}_{K,\lambda}^{\mathrm{RMOM}}$. Under the conditions of Theorem 50, there exists a universal constant $D > 0$ such that with probability at least $1 - \exp(-72K/625)$:*

$$\|\hat{\beta}\hat{\beta}^\top - \beta^*(\beta^*)^\top\|_2 \leq \frac{D}{\sqrt{N\theta^2}} \max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right).$$

In the case where $K \leq k^2 \log\big(ed/k\big)$, we get a rate of convergence of magnitude $\sqrt{k^2/(N\theta^2)\log\big(ed/k\big)}$, with no restrictions on the value of θ. This happens with an exponentially large probability depending on the number of groups K even though we only have $\log d$ moments and a dataset that may have been corrupted by an adversary. A similar analysis of a SLOPE regularization of the minmax MOM estimator will lead to a sparsity parameter free choice of λ.

Since it has been known since Kant that theory without practice is useless, this chapter contains the outcome of numerical experiments on three of the application problems considered: signed clustering, Max-Cut and angular synchronization.

5.1 Signed Clustering

To assess the effectiveness of the SDP relaxation, we consider the following experimental setup. We generate synthetic networks following the signed stochastic block model (SSBM) previously described in Section 4.3.1, with $K = 5$ communities. To quantify the effectiveness of the SDP relaxation, we compare the accuracy of a suite of algorithms from the signed clustering literature, *before* the SDP relaxation (i.e., when we perform these algorithms directly on A) and *after* the SDP relaxation (i.e., when we perform the very same algorithms on $\hat{Z}$). To measure the recovery quality of the clustering results, for a given indicator set $x_1, \ldots, x_K$, we rely on the error rate consider in Chiang et al. (2012), defined as

$$\gamma = \sum_{c=1}^{K} \frac{x_c^T A_{com}^- x_c + x_c^T L_{com}^+ x_c}{n^2}, \tag{5.1}$$

where x_c denotes a cluster indicator vector, A_{com} $(= \mathbb{E}A)$ is the complete K-weakly balanced ground truth network – with 1's on the diagonal blocks corresponding to inter-cluster edges, and -1 otherwise – with $A_{com} = A_{com}^+ - A_{com}^-$, and L_{com}^+ denotes the combinatorial graph Laplacian corresponding to A_{com}^-. Note that $x_c^T A_{com}^- x_c$ counts the number of violations within the clusters (since negative edges should not be placed within clusters) and $x_c^T L_{com}^+ x_c$ counts the number of violations across clusters (since positive edges should not belong to the cut). Overall, (5.1) essentially counts the fraction of intra-cluster and inter-cluster edge violations, with respect to the full ground truth matrix. Note that this definition can also be easily adjusted to work on real data sets, where the ground truth matrix A_{com} is not available, which one can replace with the empirical observation A.

In terms of the signed clustering algorithms compared, we consider the following algorithms from the literature. One straightforward approach is to simply rely on the spectrum of the observed adjacency matrix A. Kunegis et al. (2010) proposed spectral tools for clustering, link prediction, and visualization of signed graphs, by solving a 2-way 'signed' ratio-cut problem based on the combinatorial Signed Laplacian (Hou, 2005) $\bar{L} = \bar{D} - A$, where $\bar{D}$ is a diagonal matrix with $\bar{D}_{ii} = \sum_{i=1}^{n} |A_{ij}|$. The same authors proposed signed extensions for the case of the random-walk Laplacian $\bar{L}_{\text{rw}} = I - \bar{D}^{-1}A$, and the symmetric graph Laplacian $\bar{L}_{\text{sym}} = I - \bar{D}^{-1/2}A\bar{D}^{-1/2}$, the latter of which is particularly suitable for skewed degree distributions. Finally, the last algorithm we considered is BNC of Chiang et al. (2012), who introduced a formulation based on the *Balanced Normalized Cut* objective

$$\min_{\{x_1,\ldots,x_K\}\in\mathcal{I}} \left(\sum_{c=1}^{K} \frac{x_c^T (D^+ - A) x_c}{x_c^T \bar{D} x_c} \right), \tag{5.2}$$

which, in light of the decomposition $D^+ - A = D^+ - (A^+ - A^-) = D^+ - A^+ + A^- = L^+ + A^-$, is effectively minimizing the number of violations in the clustering procedure.

In our experiments, we first compute the error rate γ_{before} of all algorithms on the original SSBM graph (shown in Column 1 of Figure 51), and then we repeat the procedure but with the input to all signed clustering algorithms being given by the output of the SDP relaxation, and denote the resulting recovery error by γ_{after}. The third column of the same Figure 51 shows the difference in errors $\gamma_\delta = \gamma_{before} - \gamma_{after}$ between the first and second columns, while the fourth column contains a histogram of the error differences γ_δ. This altogether illustrates the fact that the SDP relaxation does improve the performance of all signed clustering algorithms, except $\bar{L}$, and could effectively be used as a denoising pre-processing step. One potential reason why the SDP pre-processing step does not improve on the accuracy of $\bar{L}$ could stem from the fact that $\bar{L}$ has a good performance to begin with on examples where the clusters have equal sizes and the degree distribution is homogeneous. It would be interesting to further compare the results in settings with skewed degree distributions, such as the classical Barabási-Albert model (Albert & Barabási, 2002).

5.2 Max-Cut

For the Max-Cut problem, we consider two sets of numerical experiments. First, we consider a version of the stochastic block model which essentially perturbs a complete bipartite graph

$$\mathbf{B} = \begin{vmatrix} \mathbf{0}_{n_1 \times n_1} & \mathbf{1}_{n_1 \times n_2} \\ \mathbf{1}_{n_2 \times n_1} & \mathbf{0}_{n_2 \times n_2} \end{vmatrix}, \tag{5.3}$$

where $\mathbf{1}_{n_1 \times n_2}$ (respectively, $\mathbf{0}_{n_1 \times n_2}$) denotes an $n_1 \times n_2$ matrix of all ones, respectively, all zeros. In our experiments, we set $n_1 = n_2 = \frac{n}{2}$, and fix $n = 500$. We perturb $\mathbf{B}$ by deleting edges across the two partitions, and inserting edges within each partition. More specifically, we generated the *full* adjacency matrix A^0 from $\mathbf{B}$ by adding edges independently with probability η within each partition (i.e., along the diagonal blocks in (5.3)). Finally, we denote by A the masked version we observe, $A = A^0 \circ S$, where S denotes the adjacency matrix of an Erdős-Rényi(n, δ) graph. The graph shown in Figure 52 is an instance of the above generative model. Note that, for small values of η, we expect the maximum cut to occur across the initial partition

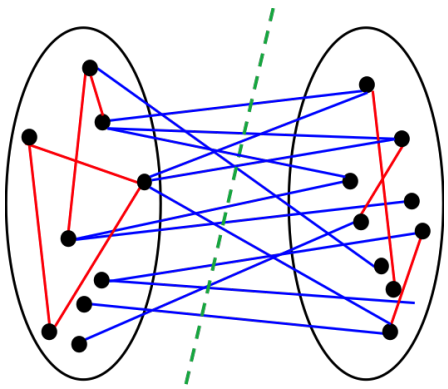


Figure 52: Illustration of MAX-CUT in the setting of a perturbation of a complete bipartite graph.

$\mathcal{P}_{\mathbf{B}}$ in the clean bipartite graph $\mathbf{B}$, which we aim to recover as we sparsify the observed graph A. The heatmap in the left of Figure 53 shows the Adjusted Rand Index (ARI) between the initial partition $\mathcal{P}_{\mathbf{B}}$ and the partition of the MAX-CUT SDP relaxation in (4.26), as we vary the noise parameter η and the sparsity δ. As expected, for a fix level of noise η, we are able to

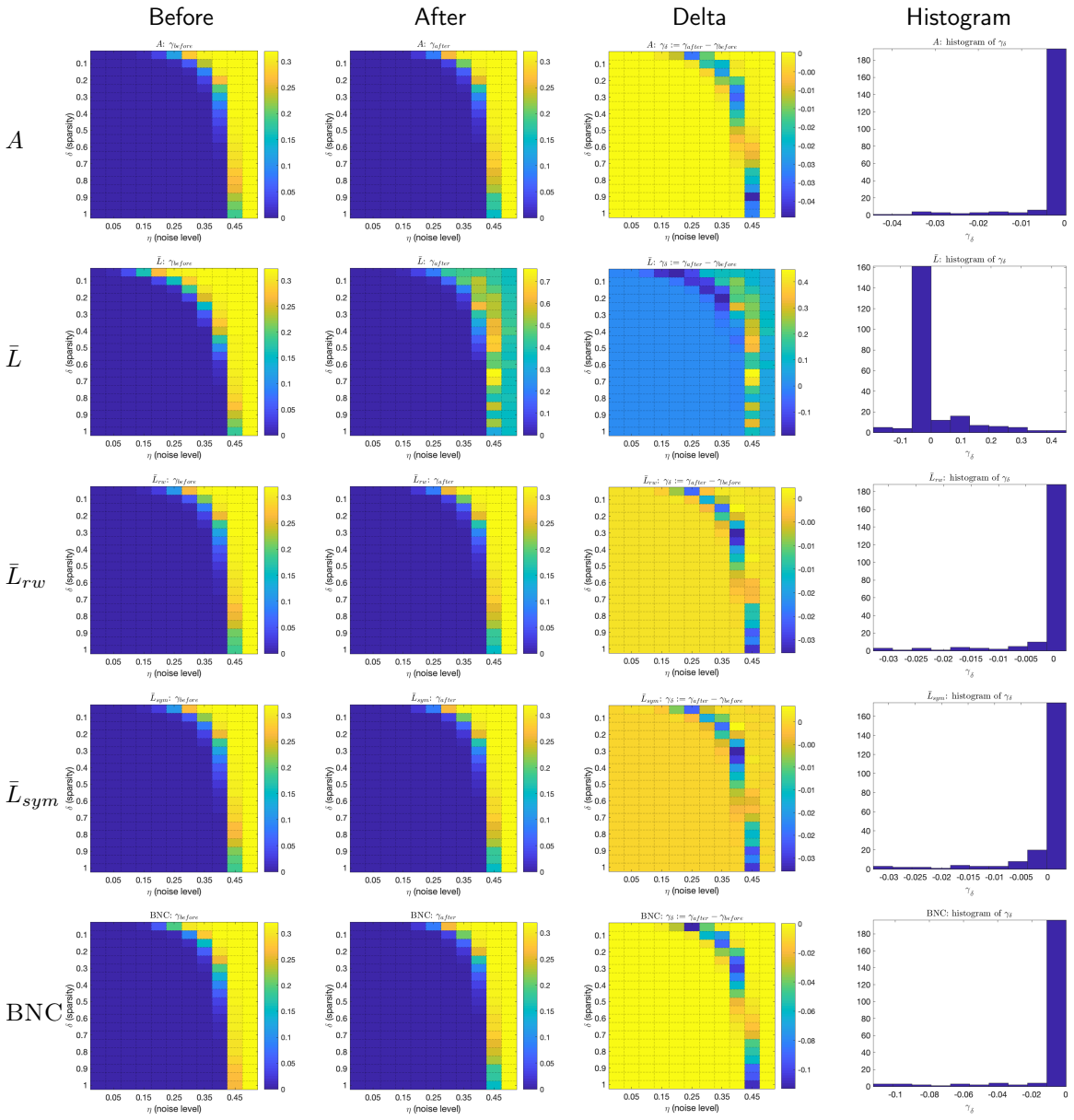
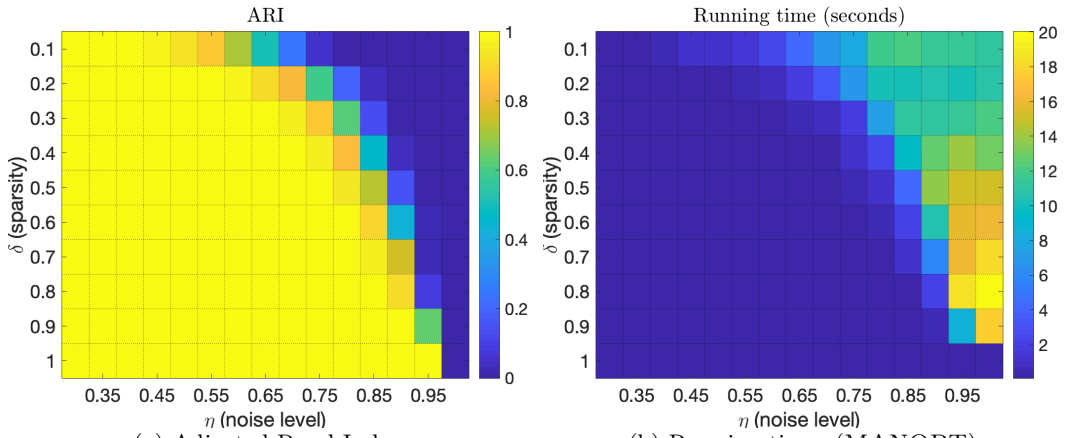


Figure 51: Summary of results for the Signed Clustering problem. The first column denotes the recovery error **before** the SDP relaxation step, meaning that we consider a number of signed clustering algorithms from the literature which we apply directly the initial adjacency matrix A. The second column contains the results when applying the same suite of algorithms **after** the SDP relaxation. The third column shows the difference in errors between the first and second columns, while the fourth column contains a histogram of the delta errors. This altogether illustrates the fact the SDP relaxation does improve the performance of all signed clustering algorithms except $\bar{L}$. Results are averaged over 20 runs.

recover the hypothetically optimal MAX-CUT, for suitable levels of the sparsity parameter. The heatmap in the right of Figure 53 shows the computational running time, as we vary the two parameters, showing that the MANOPT solver takes the longest to solve dense noisy problems, as one would expect.



(a) Adjusted Rand Index. (b) Running times (MANOPT).

Figure 53: Numerical results for MAX-CUT on a perturbed complete bipartite graph, as we vary the noise level η and the sampling sparsity δ. Results are averaged over 20 runs.

In the second set of experiments shown in Figure 54, we consider a graph A^0 chosen at random from the collection[1] of graphs known in the literature as the GSET, where we vary the sparsity level δ, and show the MAX-CUT value attained on the original full graph A^0, but using the MAX-CUT partition computed by the SDP relaxation (4.26) on the sparsified graph A.

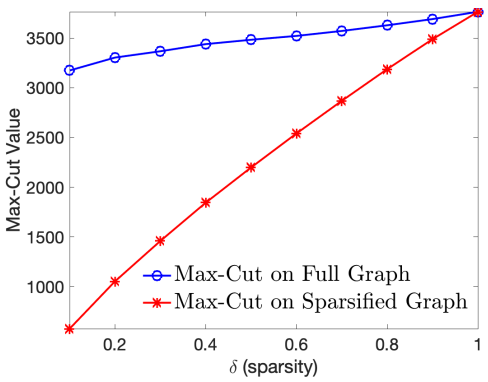


Figure 54: Max-Cut results for the G53 benchmark graph (from the GSET collection) with $n = 1000$ nodes and average degree ≈ 12. Results are averaged over 20 runs.

5.3 Angular Synchronization

For the angular synchronization problem, we consider the following experimental setup, by assessing the quality of the recovered angular solution from the SDP relaxation, as we vary the two parameters of interest. In the x-axis in the plots from Figures 55 and 56 we vary the noise level σ, under two different noise models, Gaussian and outliers. On the y-axis, we vary the sparsity of the sampling graph.

We measure the quality of the recovered angles via the Mean Squared Error (MSE), defined as follows. Since a solution can only be recovered up to a global shift, one needs an MSE error

[1] http://web.stanford.edu/~yyye/yyye/Gset/

that mods out such a degree of freedom. The following MSE is also more broadly applicable for the case when the underlying group is the orthogonal group $O(d)$, as opposed to just $SO(2)$ as in the present work, where one can replace the unknown angles $\theta_1, \ldots, \theta_d$ with their respective representation as 2×2 rotation matrices $h_1, \ldots, h_d \in \mathrm{O}(2)$. To that end, we look for an optimal orthogonal transformation $\hat{O} \in \mathrm{O}(2)$ that minimizes the sum of squared distances between the estimated orthogonal transformations and the ground truth measurements

$$\hat{O} = \underset{O \in O(2)}{\operatorname{argmin}} \sum_{i=1}^{d} \|h_i - O\hat{h}_i\|_F^2, \tag{5.4}$$

where $\hat{h}_1, \ldots, \hat{h}_d \in \mathrm{O}(2)$ denote the 2×2 rotation matrix representation of the estimated angles $\hat{\theta}_1, \ldots, \hat{\theta}_d$. In other words, $\hat{O}$ is the optimal solution to the alignment problem between two sets of orthogonal transformations, in the least-squares sense. Following the analysis of Singer and Shkolnisky (2011), and making use of properties of the trace, one arrives at

$$\begin{aligned} \sum_{i=1}^{d} \|h_i - O\hat{h}_i\|_F^2 &= \sum_{i=1}^{d} \operatorname{Trace}\left[\left(h_i - O\hat{h}_i\right)\left(h_i - O\hat{h}_i\right)^T\right] \\ &= \sum_{i=1}^{d} \operatorname{Trace}\left[2I - 2O\hat{h}_i h_i^T\right] = 4d - 2 \operatorname{Trace}\left[O \sum_{i=1}^{d} \hat{h}_i h_i^T\right]. \end{aligned} \tag{5.5}$$

If we let Q denote the 2×2 matrix

$$Q = \frac{1}{d} \sum_{i=1}^{d} \hat{h}_i h_i^T, \tag{5.6}$$

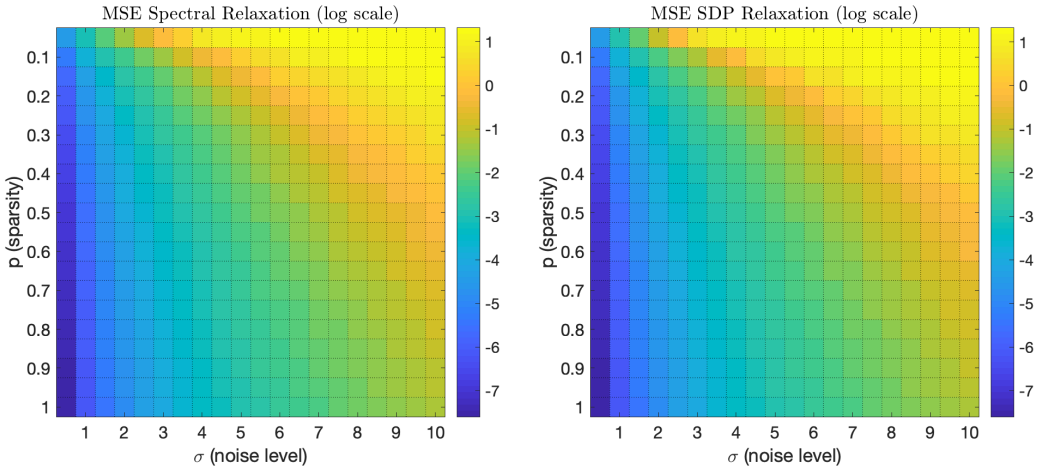
it follows from (5.5) that the MSE is given by minimizing

$$\frac{1}{d} \sum_{i=1}^{d} \|h_i - O\hat{h}_i\|_F^2 = 4 - 2Tr(OQ). \tag{5.7}$$

In Arun, Huang, and Bolstein (1987) it is proven that $Tr(OQ) \leq Tr(VU^TQ)$, for all $O \in O(3)$, where $Q = U\Sigma V^T$ is the singular value decomposition of Q. Therefore, the MSE is minimized by the orthogonal matrix $\hat{O} = VU^T$ and is given by

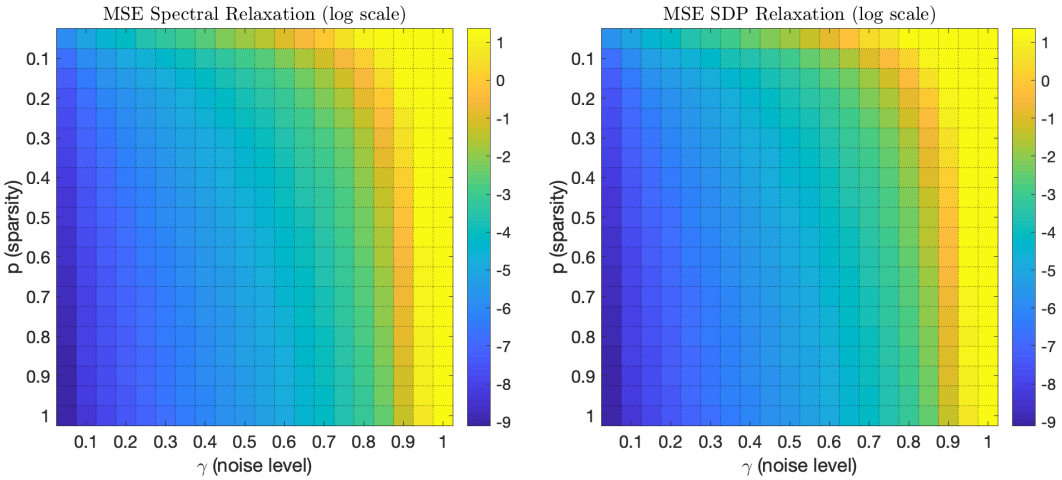
$$\mathrm{MSE} \overset{\mathrm{def}}{=} \frac{1}{d} \sum_{i=1}^{d} \|h_i - \hat{O}\hat{h}_i\|_F^2 = 4 - 2 \operatorname{Trace}(VU^TU\Sigma V^T) = 4 - 2(\sigma_1 + \sigma_2), \tag{5.8}$$

where σ_1, σ_2 are the singular values of Q. Therefore, whenever Q is an orthogonal matrix for which $\sigma_1 = \sigma_2 = 1$, the MSE vanishes. Indeed, the numerical experiments (on a log scale) in Figures 55 and 56 confirm that for noiseless data, the MSE is very close to zero. Furthermore, as one would expect, under favorable noise regimes and sparsity levels, we have almost perfect recovery, both by the SDP and the spectral relaxations, under both noise models.



(a) Spectral relaxation. (b) SDP relaxation (solved via MANOPT).

Figure 55: Recovery rates (MSE (5.8) - the lower the better) for angular synchronization with $n = 500$, under the Gaussian noise model, as we vary the noise level σ and the sparsity p of the measurement graph. Results are averaged over 20 runs.



(a) Spectral relaxation. (b) SDP relaxation (solved via MANOPT).

Figure 56: Recovery rates (MSE (5.8) - the lower the better) for angular synchronization with $n = 500$, under the Outlier noise model, as we vary the noise level γ and the sparsity p of the measurement graph. Results are averaged over 20 runs.

This chapter gathers all the proofs from the previous chapters – general excess risk and estimation bounds as well as applications.

6.1 Proofs of Chapter 3

We define the regularized excess risk $\mathcal{L}_Z^\lambda := \mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|)$, and the regularized loss $\ell_Z^\lambda := \ell_Z + \lambda\|.\|$ for all $Z \in \mathcal{C}$.

6.1.1 Proofs of Section 3.2.1: the ERM estimator

6.1.1.1 Proof of Theorem 3

Denote $r^* = r^*(\delta)$. Assume first that $r^* > 0$ (the case $r^* = 0$ is analyzed later). Let Ω^* be the event onto which for all $Z \in \mathcal{C}$ if $\langle \mathbb{E}A, Z^* - Z\rangle \leq r^*$ then $\langle A - \mathbb{E}A, Z - Z^*\rangle \leq (1/2)r^*$. By Definition of r^*, we have $\mathbb{P}[\Omega^*] \geq 1 - \delta$.

Let $Z \in \mathcal{C}$ be such that $\langle \mathbb{E}A, Z^* - Z\rangle > r^*$ and define Z' such that

$$Z' - Z^* = r^*\langle \mathbb{E}A, Z^* - Z\rangle^{-1}(Z - Z^*).$$

We have $\langle \mathbb{E}A, Z^* - Z'\rangle = r^*$ and $Z' \in \mathcal{C}$ because $\mathcal{C}$ is star-shaped in Z^*. Therefore, on the event Ω^*, $\langle A - \mathbb{E}A, Z' - Z^*\rangle \leq (1/2)r^*$ and so $\langle A - \mathbb{E}A, Z - Z^*\rangle \leq (1/2)\langle \mathbb{E}A, Z^* - Z\rangle$. It therefore follows that on the event Ω^*, if $Z \in \mathcal{C}$ is such that $\langle \mathbb{E}A, Z^* - Z\rangle > r^*$ then

$$\langle A, Z - Z^*\rangle \leq (-1/2)\langle \mathbb{E}A, Z^* - Z\rangle < -r^*/2,$$

which implies that $\langle A, Z - Z^*\rangle < 0$ and therefore Z does not maximize $Z \to \langle A, Z\rangle$ over $\mathcal{C}$. As a consequence, we necessarily have $\langle \mathbb{E}A, Z^* - \hat{Z}\rangle \leq r^*$ on the event Ω^* (which holds with probability at least $1 - \delta$).

Let us now assume that $r^* = 0$. There exists a decreasing sequence $(r_n)_n$ of positive real numbers tending to $r^* = 0$ such that for all $n \geq 0$, $\mathbb{P}[\Omega_n] \geq 1 - \delta$ where Ω_n is the event $\Omega_n = \{\psi(r_n) \leq \theta/2\}$ where for all $r > 0$,

$$\psi(r) = \frac{1}{r} \sup_{Z \in \mathcal{C}: \langle \mathbb{E}A, Z^* - Z\rangle \leq r} \langle A - \mathbb{E}A, Z - Z^*\rangle.$$

Since $\mathcal{C}$ is star-shapped in Z^*, ψ is a non-increasing function and so $(\Omega_n)_n$ is a decreasing sequence (i.e. $\Omega_{n+1} \subset \Omega_n$ for all $n \geq 0$). It follows that $\mathbb{P}[\cap_n \Omega_n] = \lim_n \mathbb{P}[\Omega_n] \geq 1 - \delta$. Let us now place ourselves on the event $\cap_n \Omega_n$. For all n, since Ω_n holds and $r_n > 0$, we can use the same argument as in first case to conclude that $\langle \mathbb{E}A, Z^* - \hat{Z}\rangle \leq r_n$. Since the latter inequality is true for all n (on the event $\cap_n \Omega_n$) and $(r_n)_n$ tends to zero, we conclude that $\langle \mathbb{E}A, Z^* - \hat{Z}\rangle \leq 0 = r^*$.

6.1.1.2 Proof of Theorem 5

Consider $\delta \in (0,1)$. Following the same lines as in the proof of Theorem 3, replacing r^* by r^*_G, we show that on an event Ω that holds with probability at least $1-\delta$, we have $\left\langle \mathbb{E}A, Z^* - \hat{Z} \right\rangle \leq r^*_G(\delta)$. Assumption 3.1 then allows us to conclude that with the same probability $1-\delta$, $\left\langle \mathbb{E}A, Z^* - \hat{Z} \right\rangle \geq G(Z^* - \hat{Z})$, which concludes the proof.

6.1.1.3 Proof of Theorem 6

Consider $r^* = r^*_G(\delta)$. First assume that $r^* > 0$. Let $Z \in \mathcal{C}$ be such that $G(Z^* - Z) > r^*$. Consider $f : \lambda \in [0,1] \to G(\lambda(Z^* - Z))$. We have $f(0) = G(0) = 0$, $f(1) = G(Z^* - Z) > r^*$ and f is continuous. Therefore, there exists $\lambda_0 \in (0,1)$ such that $f(\lambda_0) = r^*$. We let Z' be such that $Z' - Z^* = \lambda_0(Z - Z^*)$. Since $\mathcal{C}$ is star-shapped in Z^* and $\lambda_0 \in [0,1]$ we have $Z' \in \mathcal{C}$. Moreover, $G(Z^* - Z') = r^*$. As a consequence, on the event Ω^* such that for all $Z \in \mathcal{C}$ if $G(Z^* - Z) \leq r^*$ then $\left\langle A - \mathbb{E}A, Z - Z^* \right\rangle \leq (1/2)r^*$, we have $\left\langle A - \mathbb{E}A, Z' - Z^* \right\rangle \leq (1/2)r^*$. The latter and Assumption 3.2 imply that, on Ω^*:

$$(1/2)r^* \geq \left\langle A, Z' - Z^* \right\rangle + \left\langle \mathbb{E}A, Z^* - Z' \right\rangle \geq \left\langle A, Z' - Z^* \right\rangle + G(Z^* - Z') \geq \left\langle A, Z' - Z^* \right\rangle + r^*$$

and so $\left\langle A, Z' - Z^* \right\rangle \leq -r^*/2$. Finally, using the definition of Z', we obtain

$$\left\langle A, Z - Z^* \right\rangle = (1/\lambda_0)\left\langle A, Z' - Z^* \right\rangle \leq -r^*/(2\lambda_0) < 0.$$

In particular, Z cannot be a maximizer of $Z \to \left\langle A, Z \right\rangle$ over $\mathcal{C}$ and so necessarily, on the event Ω^*, $G(Z^* - \hat{Z}) \leq r^*$.

Let us now consider the case where $r^* = 0$. Using the same approach as in the proof of Theorem 3, we only have to check that the function

$$\psi : r > 0 \to \frac{1}{r} \sup_{Z \in \mathcal{C} : G(Z^* - Z) \leq r} \left\langle A - \mathbb{E}A, Z - Z^* \right\rangle$$

is non-increasing. Consider $0 < r_1 < r_2$. Without loss of generality, we may assume that there exists some $Z_2 \in \mathcal{C}$ such that $G(Z^* - Z_2) \leq r_2$ and $\psi(r_2) = \left\langle A - \mathbb{E}A, Z_2 - Z^* \right\rangle / r_2$. If $G(Z^* - Z_2) \leq r_1$ then $\psi(r_2) \leq (r_1/r_2)\psi(r_1) \leq \psi(r_1)$. If $G(Z^* - Z_2) > r_1$ then there exists $\lambda_0 \in (0,1)$ such that for $Z_1 = Z^* + \lambda_0(Z_2 - Z^*)$ we have $G(Z^* - Z_1) = r_1$ and $Z_1 \in \mathcal{C}$. Moreover, $r_1 = G(\lambda_0(Z^* - Z_2)) \leq \lambda_0 G(Z^* - Z_2) \leq \lambda_0 r_2$ and so $\lambda_0 \geq r_1/r_2$. It follows that

$$\psi(r_2) = \frac{1}{r_2}\left\langle A - \mathbb{E}A, Z_2 - Z^* \right\rangle = \frac{1}{\lambda_0 r_2}\left\langle A - \mathbb{E}A, Z_1 - Z^* \right\rangle \leq \frac{r_1}{\lambda_0 r_2}\psi(r_1) \leq \psi(r_1),$$

where we used that $\psi(r) > 0$ for all $r > 0$ because $Z^* \in \{Z \in \mathcal{C} : G(Z^* - Z) \leq r\}$ for all $r > 0$.

6.1.2 Proofs of Section 3.2.2: the regularized ERM estimator

6.1.2.1 Proof of Theorem 9

Let $\delta \in (0,1)$. Let $A > 0$ and $\rho^* > 0$ be such that Assumption 3.3 holds, and assume that $\rho^* > 0$ satisfies the A-sparsity equation from Definition 8. Let $\gamma := 1/(3A)$. In the rest of the proof, we write $r^*(.)$ for $r^*_{\mathrm{RERM,G}}(A, ., \delta)$. Let us define

$$\mathcal{B} := \left\{ Z \in \mathcal{C} : \|Z - Z^*\| \leq \rho^* \text{ and } G(Z - Z^*) \leq r^*(\rho^*) \right\}.$$

Consider the following event:

$$\Omega := \left\{ \forall Z \in \mathcal{B}, \quad |(P - P_N)\mathcal{L}_Z| \leq \gamma r^*(\rho^*) \right\}.$$

By definition of $r^*(.)$, Ω holds with probability at least $1 - \delta$. Let us now prove the statistical bounds announced in Theorem 9 on the event Ω.

Suppose that $\hat{Z} \in \mathcal{B}$. This means that $\|\hat{Z} - Z^*\| \leq \rho^*$ and $G(\hat{Z} - Z^*) \leq r^*(\rho^*)$. Moreover, on Ω it also means that $|(P - P_N)\mathcal{L}_{\hat{Z}}| \leq \gamma r^*(\rho^*)$, and then:

$$\begin{aligned}
P\mathcal{L}_{\hat{Z}} &= (P - P_N)\mathcal{L}_{\hat{Z}} + P_N\mathcal{L}_{\hat{Z}} \\
&\leq \gamma r^*(\rho^*) + P_N\mathcal{L}_{\hat{Z}} \\
&= \gamma r^*(\rho^*) + P_N(\mathcal{L}^{\lambda}_{\hat{Z}} - \lambda(\|\hat{Z}\| - \|Z^*\|)) \\
&= \gamma r^*(\rho^*) + P_N\mathcal{L}^{\lambda}_{\hat{Z}} + \lambda(\|Z^*\| - \|\hat{Z}\|) \\
&\overset{(i)}{\leq} \gamma r^*(\rho^*) + \lambda\|\hat{Z} - Z^*\| \\
&\leq \gamma r^*(\rho^*) + \lambda\rho^* \\
&\overset{(ii)}{\leq} 3\gamma r^*(\rho^*) = \frac{r^*(\rho^*)}{A},
\end{aligned}$$

where (i) holds since $P_N\mathcal{L}^{\lambda}_{\hat{Z}} \leq 0$ by definition of $\hat{Z}$ and (ii) holds because of the choice of λ given in (3.7).

Then, if we can show that $\hat{Z} \in \mathcal{B}$, we will have the desired bounds on Ω. Since we know that $P_N\mathcal{L}^{\lambda}_{\hat{Z}} \leq 0$, it is sufficient to prove that for any $Z \in \mathcal{C}\backslash\mathcal{B}$, $P_N\mathcal{L}^{\lambda}_{Z} > 0$.

Let $Z \in \mathcal{C}\backslash\mathcal{B}$. Because $\mathcal{C}$ is star-shaped in Z^* and by the regularity properties assumed for G, we have the existence of $Z_0 \in \partial\mathcal{B}$, the border of $\mathcal{B}$, and $\alpha > 1$ such that $Z - Z^* = \alpha(Z_0 - Z^*)$. The border of $\mathcal{B}$, that we denoted by $\partial\mathcal{B}$ is the set of all $Z \in \mathcal{C}$ such that either $\|Z - Z^*\| = \rho^*$ and $G(Z - Z^*) \leq r^*(\rho^*)$ or $\|Z - Z^*\| \leq \rho^*$ and $G(Z - Z^*) = r^*(\rho^*)$. By linearity of the loss function, we have $P_N\mathcal{L}_Z = \alpha P_N\mathcal{L}_{Z_0}$. Moreover, we have by the triangular inequality that

$$\begin{aligned}
\|Z\| - \|Z^*\| &= \|\alpha Z_0 - (\alpha - 1)Z^*\| - \|Z^*\| \\
&\geq \alpha\|Z_0\| - (\alpha - 1)\|Z^*\| - \|Z^*\| \geq \alpha(\|Z_0\| - \|Z^*\|)
\end{aligned}$$

and so

$$\begin{aligned}
P_N\mathcal{L}^{\lambda}_{Z} &= P_N\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \\
&\geq \alpha P_N\mathcal{L}_{Z_0} + \lambda\alpha(\|Z_0\| - \|Z^*\|) = \alpha P_N\mathcal{L}^{\lambda}_{Z_0}.
\end{aligned} \tag{6.1}$$

We showed that for any $Z \in \mathcal{C}\backslash\mathcal{B}$, there exist $Z_0 \in \partial\mathcal{B}$ and $\alpha > 1$ such that $P_N\mathcal{L}^{\lambda}_{Z} > \alpha P_N\mathcal{L}^{\lambda}_{Z_0}$. Hence, we only have to show that $Z \to P_N\mathcal{L}^{\lambda}_{Z}$ is positive on the border of $\mathcal{B}$ to show that it is positive over $\mathcal{C}\backslash\mathcal{B}$.

Let $Z_0 \in \partial\mathcal{B}$. Two cases arise: either $\|Z_0 - Z^*\| = \rho^*$ and $G(Z - Z^*) \leq r^*(\rho^*)$, or $\|Z_0 - Z^*\| \leq \rho^*$ and $G(Z - Z^*) = r^*(\rho^*)$.

First case: We assume that $\|Z_0 - Z^*\| = \rho^*$ and $G(Z - Z^*) \leq r^*(\rho^*)$, that is $Z_0 \in H_{\rho^*,A}$. Let $V \in H$ be such that $\|Z^* - V\| \leq \rho^*/20$ and $\Phi \in \partial\|.\|(V)$. We have:

$$\begin{aligned}
\|Z_0\| - \|Z^*\| &\geq \|Z_0\| - \|V\| - \|Z^* - V\| \\
&\geq \left\langle \Phi, Z_0 - V \right\rangle - \|Z^* - V\| \quad (\text{ since } \Phi \in \partial\|.\|(V)) \\
&= \left\langle \Phi, Z_0 - Z^* \right\rangle - \left\langle \Phi, V - Z^* \right\rangle - \|Z^* - V\| \\
&\geq \left\langle \Phi, Z_0 - Z^* \right\rangle - 2\|Z^* - V\| \quad (\text{ since } \left\langle \Phi, U \right\rangle \leq \|U\| \text{ for any } U \in H) \\
&\geq \left\langle \Phi, Z_0 - Z^* \right\rangle - \frac{\rho^*}{10}.
\end{aligned}$$

This is true for any $\Phi \in \underset{V \in Z^* + \frac{\rho^*}{20}}{\cup} \partial\|.\|(V) = \Gamma_{Z^*}(\rho^*)$. Then taking the sup over $\Gamma_{Z^*}(\rho^*)$ gives:

$$\|Z_0\| - \|Z^*\| \geq \sup_{\Phi \in \Gamma_{Z^*}(\rho^*)} \left\langle \Phi, Z_0 - Z^* \right\rangle - \frac{\rho^*}{10}$$

and then taking the infimum over $H_{\rho^*,A}$ gives:

$$\begin{aligned}\|Z_0\| - \|Z^*\| &\geq \inf_{Z_0 \in H_{\rho^*,A}} \|Z_0\| - \|Z^*\| \\ &\geq \inf_{Z_0 \in H_{\rho^*,A}} \sup_{\Phi \in \Gamma_{Z^*}(\rho^*)} \langle \Phi, Z_0 - Z^* \rangle - \frac{\rho^*}{10} \\ &= \Delta(\rho^*, A) - \frac{\rho^*}{10} \geq \frac{7}{10}\rho^*,\end{aligned}$$

where the last inequality holds since ρ^* is supposed to satisfy the A-sparsity equation. Then, we have:

$$P_N \mathcal{L}^{\lambda}_{Z_0} = P_N \mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \geq P_N \mathcal{L}_{Z_0} + \frac{7}{10}\lambda\rho^* = P\mathcal{L}_{Z_0} - (P - P_N)\mathcal{L}_{Z_0} + \frac{7}{10}\lambda\rho^*.$$

But on Ω, we have $(P - P_N)\mathcal{L}_{Z_0} \leq \gamma r^*(\rho^*)$ since $Z_0 \in \mathcal{B}$, and we know by definition of Z^* that $P\mathcal{L}_{Z_0} \geq 0$. Then we conclude that:

$$P_N \mathcal{L}^{\lambda}_{Z_0} \geq \frac{7}{10}\lambda\rho^* - \gamma r^*(\rho^*) > 0$$

where the last inequality is due to the choice of λ given in (3.7).

Second case: Now we assume that $\|Z_0 - Z^*\| \leq \rho^*$ and $G(Z - Z^*) = r^*(\rho^*)$. We have:

$$\begin{aligned}P_N \mathcal{L}^{\lambda}_{Z_0} &= P_N \mathcal{L}_{Z_0} - \lambda(\|Z^*\| - \|Z_0\|) \\ &\geq P\mathcal{L}_{Z_0} - (P - P_N)\mathcal{L}_{Z_0} - \lambda\|Z^* - Z_0\| \\ &\geq P\mathcal{L}_{Z_0} - (P - P_N)\mathcal{L}_{Z_0} - \lambda\rho^*.\end{aligned}$$

But we know from Assumption 3.3 that $P\mathcal{L}_{Z_0} \geq A^{-1}G(Z_0 - Z^*)$, and on Ω we have $(P - P_N)\mathcal{L}_{Z_0} \leq \gamma r^*(\rho^*)$. Then we get:

$$P_N \mathcal{L}^{\lambda}_{Z_0} \geq A^{-1}G(Z_0 - Z^*) - \gamma r^*(\rho^*) - \lambda\rho^* = A^{-1}r^*(\rho^*) - \gamma r^*(\rho^*) - \lambda\rho^* > 0,$$

where the last inequality comes from the choice of λ given in (3.7).

Then, we proved that $P_N \mathcal{L}^{\lambda}_{Z_0} > 0$ for any $Z_0 \in \partial(\mathcal{B})$ and as we said before, this implies that $P_N \mathcal{L}^{\lambda}_{Z}$ is positive over $\mathcal{C}\backslash\mathcal{B}$. Since $P_N \mathcal{L}^{\lambda}_{\hat{Z}} < 0$ we conclude that on Ω, $\hat{Z}$ necessarily belongs to $\mathcal{B}$, which proves the bounds announced in Theorem 9.

6.1.3 Proofs of Section 3.3.1: the minmax MOM estimator

6.1.3.1 Proof of Theorem 11

The proof of this theorem is broken down into two steps. First, we identify an event Ω on which the estimator $\hat{Z}^{\mathrm{MOM}}_{\mathrm{K}}$ has the desired properties. Then, we show that this event holds with high probability. For the sake of simplicity, in the rest of the proof we write r^* for $r^*_{\mathrm{MOM,ER}}(\gamma)$ and $\hat{Z}$ for $\hat{Z}^{\mathrm{MOM}}_{K}$. Let $\gamma = 1/6400$, and consider the set $\mathcal{C}_\gamma := \left\{Z \in \mathcal{C} : P\mathcal{L}_Z \leq (r^*)^2\right\}$. Define the event Ω_K as follows:

$$\Omega_K := \left\{\forall Z \in \mathcal{C}_\gamma, \exists J \subset [K] : |J| > K/2 \text{ and } \forall k \in J, |(P_{B_k} - P)\mathcal{L}_Z| \leq (r^*)^2/4\right\}.$$

We start with showing that on Ω_K, the estimator $\hat{Z}$ satisfies the excess risk bound announced in Theorem 11.

Lemma 52. *On the event Ω_K, $P\mathcal{L}_{\hat{Z}} \leq (r^*)^2$.*

Proof. Let $Z \in \mathcal{C}\backslash\mathcal{C}_\gamma$. Let $\alpha := (r^*)^{-2}P\mathcal{L}_Z > 1$, and let $Z_0 := Z^* + \alpha^{-1}(Z - Z^*)$. By the star-shaped property of $\mathcal{C}$, $Z_0 \in \mathcal{C}$, and by linearity of ℓ, $P\mathcal{L}_{Z_0} = \alpha^{-1}P\mathcal{L}_Z = (r^*)^2$, so that $Z_0 \in \mathcal{C}_\gamma$. Then, on Ω_K, there exists strictly more than $K/2$ blocks B_k on which $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leq (r^*)^2/4$, that is $P_{B_k}\mathcal{L}_{Z_0} \geq P\mathcal{L}_{Z_0} - (r^*)^2/4 = (3/4)(r^*)^2$ and so $P_{B_k}\mathcal{L}_Z = \alpha P_{B_k}\mathcal{L}_{Z_0} \geq \alpha(3/4)(r^*)^2$ because $\alpha > 1$. This holds on strictly more than half of the blocks B_k, therefore $\mathrm{Med}(-P_{B_k}\mathcal{L}_Z : k \in [K]) \geq -(3/4)(r^*)^2$ and this holds for all $Z \in \mathcal{C}\backslash\mathcal{C}_\gamma$, hence, we have

$$\sup_{Z\in\mathcal{C}\backslash\mathcal{C}_\gamma} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq -(3/4)(r^*)^2. \tag{6.2}$$

Moreover, on Ω_K, for $Z \in \mathcal{C}_\gamma$, there exists strictly more than $K/2$ blocks B_k on which $-P_{B_k}\mathcal{L}_Z \leq (r^*)^2/4 - P\mathcal{L}_Z \leq (r^*)^2/4$, since $P\mathcal{L}_Z \geq 0$ by definition of Z^*. Therefore, we have

$$\sup_{Z\in\mathcal{C}_\gamma} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq (r^*)^2/4. \tag{6.3}$$

But by definition of $\hat{Z}$, we have:

$$\begin{aligned}
\mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) &\leq \sup_{Z\in\mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \\
&= \max\left(\sup_{Z\in\mathcal{C}_\gamma} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z), \sup_{Z\in\mathcal{C}\backslash\mathcal{C}_\gamma} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z)\right) \\
&\leq \frac{(r^*)^2}{4}
\end{aligned}$$

that is, $\mathrm{MOM}_K(\ell_{Z^*} - \ell_{\hat{Z}}) \geq -(1/4)(r^*)^2 > -(3/4)(r^*)^2$. From (6.2) we conclude that, necessarily, $\hat{Z} \in \mathcal{C}_\gamma$, that is, $P\mathcal{L}_{\hat{Z}} \leq (r^*)^2$. □

At this point, we proved that on the event Ω_K, the estimator $\hat{Z}$ satisfies the statistical bounds announced in Theorem 11. Now it remains to prove that Ω_K holds with high probability.

Lemma 53. *Assume that $|\mathcal{O}| \leq K/100$. Then Ω_K holds with probability at least $1-\exp(-72K/625)$.*

Proof. Consider

$$\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t\geq 1\}} + 2(t - (1/2))\mathbb{1}_{\left\{1/2\leq t\leq 1\right\}},$$

so that for any $t \in \mathbb{R}$,

$$\mathbb{1}_{\{t\geq 1\}} \leq \phi(t) \leq \mathbb{1}_{\left\{t\geq 1/2\right\}}.$$

For $k \in [K]$, let $W_k := \{X_i : i \in B_k\}$ and $F_Z(W_k) = (P_{B_k} - P)\mathcal{L}_Z$. We also define the counterparts of these quantities constructed with the non-corrupted vectors: $\widetilde{W}_k := \left\{\widetilde{X}_i : i \in B_k\right\}$ and $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - P)\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := (K/N)\sum_{i\in B_k} \mathcal{L}_Z(\widetilde{X}_i)$. Define

$$\psi : Z \in \mathbb{R}^{d\times d} \to \sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(W_k)|\leq (r^*)^2/4\right\}}.$$

We show now that, with high probability, if $Z \in \mathcal{C}_\gamma$, then $\psi(Z) > K/2$. In the contaminated framework, it is sufficient to prove that, with high probability, for all $Z \in \mathcal{C}_\gamma$,

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|> \frac{(r^*)^2}{4}\right\}} \leq \frac{49K}{100}. \tag{6.4}$$

Indeed, consider $Z \in \mathcal{C}_\gamma$ such that (6.4) holds. Then, there exist at least $(1 - 49/100)K = (51/100)K$ blocks B_k on which $|F_Z(\widetilde{W}_k)| \leq (r^*)^2/4$. On the other hand, we know that $|\mathcal{O}| \leq K/100$, so that among the $(51/100)K$ previous blocks, at most $K/100$ contain corrupted data. The other $(50/100)K = K/2$ contain only non-corrupted data, so we have $F_Z(\widetilde{W}_k) = F_Z(W_k)$ on these blocks. We conclude that $\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(W_k)| \leq (r^*)^2/4\right\}} > K/2$, that is $\psi(Z) > K/2$, if (6.4) holds.

Let $Z \in \mathcal{C}_\gamma$. We have:

$$
\begin{aligned}
&\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} \\
&= \sum_{k\in[K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right)\right] \\
&= \sum_{k\in[K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right\}}\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right) \\
&\leq \sum_{k\in[K]} \left(\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right) \\
&\leq \sup_{Z\in\mathcal{C}_\gamma} \left(\sum_{k\in[K]} \phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right).
\end{aligned}
\tag{6.5}
$$

We start with bounding the last sum in the previous inequality. For each $k \in [K]$, it follows from Markov's inequality and the definition of r^* that

$$
\begin{aligned}
\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right) &\leq \frac{64}{(r^*)^4}\mathbb{E}\left[F_Z(\widetilde{W}_k)^2\right] = \frac{64}{(r^*)^4}\left(\frac{K}{N}\right)\mathrm{Var}(\mathcal{L}_Z(\widetilde{X})) \\
&= \frac{64}{(r^*)^4}\left(V_K(r^*)\right)^2 \leq \frac{1}{200}.
\end{aligned}
$$

Plugging that into (6.5), we get:

$$
\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} \leq \frac{K}{200} + \sup_{Z\in\mathcal{C}_\gamma} \left(\sum_{k\in[K]} \Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\left[\Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right).
\tag{6.6}
$$

We now we have to bound this last term. Using Mc Diarmind inequality (Theorem 6.2 in Boucheron et al. (2013a) for $t = 12/25$), we get that with probability at least $1 - \exp(-72K/625)$, for all $Z \in \mathcal{C}_\gamma$,

$$
\sum_{k\in[K]} \phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) \tag{6.7}
$$

$$
\leq \frac{12}{25}K + \mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma} \sum_{k\in[K]} \phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]. \tag{6.8}
$$

Let now $\epsilon_1, \ldots, \epsilon_K$ be Rademacher variables independent from the $\widetilde{X}_i$'s. By the symmetrization

Lemma, we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right] \leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]. \tag{6.9}$$

As ϕ is 2-Lipschitz with $\phi(0)=0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), Theorem 4.12) to get that:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right] \leq 8\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k\frac{F_Z(\widetilde{W}_k)}{(r^*)^2}\right] = \frac{8}{(r^*)^2}\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-P)\mathcal{L}_Z\right]. \tag{6.10}$$

Now, let $(\sigma_i)_{i=1,\dots,N}$ be a family of Rademacher variables independent from the $\widetilde{X}_i$'s and the ϵ_i's. For any $k\in[K]$ and any $i\in[N]$, the variables $\epsilon_k\sigma_i\mathcal{L}_Z(X_i)$ and $\sigma_i\mathcal{L}_Z(X_i)$ have the same distribution, so that we get, using the symmetrization Lemma:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-P)\mathcal{L}_Z\right] \leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\frac{K}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right] = 2KE(r^*) \leq 2K\gamma(r^*)^2.$$

Combining this with (6.7), (6.9) and (6.10), we finally get that, with probability at least $1-\exp(-72K/625)$

$$\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right] \leq \left(\frac{12}{25}+32\gamma\right)K. \tag{6.11}$$

Plugging that into (6.6), we conclude that with probability at least $1-\exp(-72K/625)$, one has

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r^*)^2}{4}\right\}} \leq \left(\frac{1}{200}+\frac{12}{25}+32\gamma\right)K \leq \frac{49}{100}K$$

from our choice of parameters. This allows to affirm that Ω_K holds with probability at least $1-\exp(-72K/625)$, which concludes the proof.

□

6.1.3.2 Proof of Theorem 13.

The proof is divided into two parts. First, we identify an event Ω_K on which the estimator has the desired statistical properties. Second, we prove that this event holds with high probability. For the sake of simplicity, we write $\hat{Z}$ for $\hat{Z}_K^{\mathrm{MOM}}$ and r^* for $r^*_{\mathrm{MOM},L_2}(\gamma)$ with $\gamma=1/3200$. Let $0<A<1$ be such that Assumption 3.5 holds. We define $\nu=A^2/\gamma$, $\tau=(2A)^{-1}$, $C_{K,A}=\max\left((r^*)^2,\nu K/N\right)$ and $\mathcal{B}_{K,A}:=\left\{Z\in\mathcal{C}:\|Z-Z^*\|_{L_2}\leq\sqrt{C_{K,A}}\right\}$ - where the L_2-norm is defined as $Z\to\|Z\|_{L_2}=\mathbb{E}[\langle\widetilde{X},Z\rangle^2]^{1/2}$. We consider the following event:

$$\Omega_K:=\left\{\forall Z\in\mathcal{B}_{K,A},\exists J\subset\{1,\dots,K\}:|J|>\frac{K}{2}\text{ and }\forall k\in J,\left|(P_{B_k}-P)\mathcal{L}_Z\right|\leq\tau C_{K,A}\right\}.$$

We show in the next three lemmas that, on Ω_K, $\hat{Z}$ satisfies the statistical bounds announced in Theorem .13. Then the fourth lemma will prove that Ω_K holds with large probability, the one announced in Theorem .13.

Lemma 54. *If there exists $\eta > 0$ such that:*

$$\sup_{Z \in \mathcal{C} \setminus \mathcal{B}_{K,A}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) < -\eta \quad \text{and} \quad \sup_{Z \in \mathcal{B}_{K,A}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq \eta \tag{6.12}$$

then $\|\hat{Z} - Z^*\|_{L_2}^2 \leq C_{K,A}$.

Proof. Assume that (6.12) holds. Then:

$$\inf_{Z \in \mathcal{C} \setminus \mathcal{B}_{K,A}} \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) > \eta. \tag{6.13}$$

Moreover, if we define $Z \to T_K(Z) = \sup_{Z' \in \mathcal{C}} \mathrm{MOM}_K(\ell_Z - \ell_{Z'})$, then:

$$T_K(Z^*) = \max\left(\sup_{Z \in \mathcal{C} \setminus \mathcal{B}_{K,A}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z), \sup_{Z \in \mathcal{B}_{K,A}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \right) \leq \eta. \tag{6.14}$$

By definition of $\hat{Z}$, we have $T_K(\hat{Z}) = \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_Z) \leq T_K(Z^*) \leq \eta$. But by (6.13), any $Z \in \mathcal{C} \setminus \mathcal{B}_{K,A}$ satisfies:

$$T_K(Z) \geq \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) \geq \inf_{Z \in \mathcal{C} \setminus \mathcal{B}_{K,A}} \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) \geq \eta$$

which allows us to conclude that, necessarily, $\hat{Z} \in \mathcal{B}_{K,A}$, i.e. $\|Z^* - \hat{Z}\|_{L_2}^2 \leq C_{K,A}$. □

Lemma 55. *Assume that $K \geq 100|\mathcal{O}|$. Then on Ω_K, (6.12) holds with $\eta = \tau C_{K,A}$.*

Proof. Let $Z \in \mathcal{C}$ be such that $\|Z - Z^*\|_{L_2} > \sqrt{C_{K,A}}$. By the star-shaped property of $\mathcal{C}$, there exists $Z_0 \in \mathcal{C}$ and $\alpha > 1$ such that $\|Z_0 - Z^*\|_{L_2} = \sqrt{C_{K,A}}$ and $Z - Z^* = \alpha(Z_0 - Z^*)$. Now, for each block B_k we have by the linearity of the loss function:

$$P_{B_k} \mathcal{L}_Z = \alpha P_{B_k} \mathcal{L}_{Z_0}. \tag{6.15}$$

As $Z_0 \in \mathcal{B}_{K,A}$, on Ω_K there exist strictly more than $K/2$ blocks on which $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leq \tau C_{K,A}$. Moreover, since $\|Z_0 - Z^*\|_{L_2} = \sqrt{C_{K,A}}$, we get from Assumption 3.5 that $P\mathcal{L}_{Z_0} \geq A^{-1}\|Z_0 - Z^*\|_{L_2}^2 = A^{-1}C_{K,A}$. Then, on these blocks, $P_{B_k}(\ell_{Z_0} - \ell_{Z^*}) \geq P\mathcal{L}_{Z_0} - \tau C_{K,A} \geq (A^{-1} - \tau)C_{K,A}$, which implies that $P_{B_k}(\ell_{Z^*} - \ell_{Z_0}) \leq -(A^{-1} - \tau)C_{K,A} \leq -\tau C_{K,A}$, since we have $\tau = (2A)^{-1}$. From (6.15) we conclude that, on Ω_K, there exist sricly more than $K/2$ blocks B_k on which $P_{B_k}(\ell_{Z^*} - \ell_Z) \leq -\alpha\tau C_{K,A} \leq -\tau C_{K,A}$, since $\alpha \geq 1$. This is true for all $Z \in \mathcal{C} \setminus \mathcal{B}_{K,A}$; in other words, we have

$$\sup_{Z \in \mathcal{C} \setminus \mathcal{B}_{K,A}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq -\tau C_{K,A}$$

Moreover, on Ω_K, for any $Z \in \mathcal{B}_{K,A}$, there exist stricly more than $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leq \tau C_{K,A}$, so that $P_{B_k}(\ell_Z - \ell_{Z^*}) \geq -\tau C_{K,A} + P(\ell_Z - \ell_{Z^*}) \geq -\tau C_{K,A}$, since $P(\ell_Z - \ell_{Z^*}) \geq 0$ by definition of Z^*. Then, we have $P_{B_k}(\ell_{Z^*} - \ell_Z) \leq \tau C_{K,A}$ on stricly more than $K/2$ blocks, which implies that $\mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq \tau C_{K,A}$. This being true for any $Z \in \mathcal{B}_{K,A}$, we conclude that (6.12) holds with $\eta = \tau C_{K,A}$. □

Lemma 56. *Grant Assumption 3.5 and assume that $K \geq 100|\mathcal{O}|$. On Ω_K, $P\mathcal{L}_{\hat{Z}} \leq 2\tau C_{K,A}$.*

Proof. Assume that Ω_K holds. From Lemmas 54 and 55 , $\|\hat{Z}-Z^*\|_{L_2}^2 \leq C_{K,A}$, that is $\hat{Z} \in \mathcal{B}_{K,A}$. Therefore, on strictly more than $K/2$ blocks B_k, we have $\left|(P_{B_k}-P)\mathcal{L}_{\hat{Z}}\right| \leq \tau C_{K,A}$, and then on these blocks:

$$P\mathcal{L}_{\hat{Z}} \leq P_{B_k}\mathcal{L}_{\hat{Z}} + \tau C_{K,A}. \tag{6.16}$$

In addition, by definition of $\hat{Z}$ and (6.14) (for $\eta = \tau C_{K,A}$):

$$\mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) \leq \sup_{Z\in\mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq \tau C_{K,A}$$

which implies the existence of $K/2$ blocks (at least) on which:

$$P_{B_k}\mathcal{L}_{\hat{Z}} \leq \tau C_{K,A} \tag{6.17}$$

As a consequence, there exist at least one block B_k on which (6.16) and (6.17) holds simultaneously. On this block, we have: $P\mathcal{L}_{\hat{Z}} \leq \tau C_{K,A} + \tau C_{K,A} = 2\tau C_{K,A}$, which concludes the proof. □

At this point, we proved that on the event Ω_K, the estimator $\hat{Z}$ has the statistical properties announced in Theorem 13. In the final lemma, we show that Ω_K holds with high probability.

Lemma 57. *Assume that $|\mathcal{O}| \leq K/100$. Then Ω_K holds with probability at least $1-\exp(-72K/625)$.*

Proof. Consider

$$\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t\geq 1\}} + 2(t-1/2)\mathbb{1}_{\left\{1/2\leq t\leq 1\right\}},$$

so that for any $t \in \mathbb{R}$,

$$\mathbb{1}_{\{t\geq 1\}} \leq \phi(t) \leq \mathbb{1}_{\left\{t\geq 1/2\right\}}.$$

For $k \in [K]$, let $W_k := \{X_i : i \in B_k\}$ and $F_Z(W_k) = (P_{B_k} - P)\mathcal{L}_Z$. We also define the counterparts of these quantities constructed with the non-corrupted vectors: $\widetilde{W}_k := \left\{\widetilde{X}_i : i \in B_k\right\}$ and $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - P)\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := (K/N)\sum_{i\in B_k}\mathcal{L}_Z(\widetilde{X}_i)$. Define

$$\psi(Z) = \sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(W_k)|\leq \tau C_{K,A}\right\}}.$$

We are now showing that, with high probability, if $Z \in \mathcal{B}_{K,A}$, then $\psi(Z) > K/2$. In the adversarial corruption setup, it is enough to prove that the following inequality occurs with high probability: for all $Z \in \mathcal{B}_{K,A}$,

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\right\}} \leq \frac{49K}{100}. \tag{6.18}$$

Indeed, consider $Z \in \mathcal{C}$ such that (6.18) holds. Then, there exist at least $(1-49/100)K = (51/100)K$ blocks B_k on which $|F_Z(\widetilde{W}_k)| \leq \tau C_{K,A}$. On the other hand, we know that $|\mathcal{O}| \leq K/100$, so that among the $(51/100)K$ previous blocks, at most $K/100$ contain corrupted data. The other $(50/100)K = K/2$ contain only non-corrupted data, so we have $F_Z(\widetilde{W}_k) = F_Z(W_k)$ on these blocks. We conclude that $\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(W_k)|\leq\tau C_{K,A}\right\}} > K/2$, that is $\psi(Z) > K/2$, if (6.18) holds.

Then, we only have to show that (6.18) holds uniformly over all $Z \in \mathcal{B}_{K,A}$ with high probability. This is what we do now. Let $Z \in \mathcal{B}_{K,A}$. We have:

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\right\}} \\
&= \sum_{k\in[K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right)\right] \\
&= \sum_{k\in[K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{\tau C_{K,A}}{2}\right\}}\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) \\
&\leq \sum_{k\in[K]} \left(\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\left[\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) \\
&\leq \sup_{Z\in\mathcal{B}_{K,A}} \left(\sum_{k\in[K]} \Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\left[\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right).
\end{aligned} \tag{6.19}$$

We start with bounding the last sum in the previous inequality. For each $k \in [K]$, it follows from Markov's inequality, the definition of $C_{K,A}$ and the linearity of the loss function that

$$\begin{aligned}
\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) &\leq \frac{4}{(\tau C_{K,A})^2}\mathbb{E}\left[F_Z(\widetilde{W}_k)^2\right] \\
&= \frac{4}{(\tau C_{K,A})^2}\left(\frac{K}{N}\right)\mathrm{Var}(\mathcal{L}_Z(\widetilde{X})) \\
&\leq \frac{4}{(\tau C_{K,A})^2}\left(\frac{K}{N}\right)\mathbb{E}[\mathcal{L}_Z(\widetilde{X})^2] \\
&= \frac{4}{(\tau C_{K,A})^2}\frac{K}{N}\|Z - Z^*\|_{L_2}^2 \\
&\leq \frac{4}{(\tau C_{K,A})^2}\frac{K}{N}C_{K,A} \leq \frac{4}{\tau^2\nu} = \frac{1}{200}.
\end{aligned}$$

Plugging the latter result into (6.19), we get:

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\right\}} \leq \frac{K}{200} + \sup_{Z\in\mathcal{B}_{K,A}} \left(\sum_{k\in[K]} \Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\left[\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right). \tag{6.20}$$

We now have to bound this last term. Using Mc Diarmind inequality (Theorem 6.2 in Boucheron et al. (2013a) for taking $t = 12/25$), we get that with probability at least $1 - \exp(-72K/625)$, for all $Z \in \mathcal{B}_{K,A}$,

$$\begin{aligned}
&\sum_{k\in[K]} \phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) \\
&\qquad \leq \frac{12}{25}K + \mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}} \sum_{k\in[K]} \phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right].
\end{aligned}$$

Let now $\epsilon_1, \dots, \epsilon_K$ be Rademacher variables independent from the $\widetilde{X}_i$'s. By the symmetrization Lemma, we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)-\mathbb{E}\left[\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right]\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]$$

As ϕ is 2-Lipschitz with $\phi(0)=0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), Theorem 4.3) to get that:

$$\begin{aligned}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]&\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\frac{F_Z(\widetilde{W}_k)}{\tau C_{K,A}}\right]\\&=2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\frac{(\widetilde{P_{B_k}}-P)\mathcal{L}_Z}{\tau C_{K,A}}\right]\end{aligned}$$

Now, let $(\sigma_i)_{i=1,\dots,K}$ be a family of Rademacher variables independant from the $\widetilde{X}_i$'s and the ϵ_k's. Using the symmetrization Lemma one more time, we get

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\frac{(\widetilde{P_{B_k}}-P)\mathcal{L}_Z}{C_{K,A}}\right]\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\frac{K}{N}\sum_{i=1}^N\sigma_i\frac{\mathcal{L}_Z(\widetilde{X}_i)}{C_{K,A}}\right].$$

To bound this last term, we consider two cases: either $C_{K,A}=(r^*)^2$ or $C_{K,A}=\nu K/N$. In the first case, by definition of r^* we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^N\sigma_i\frac{\mathcal{L}_Z(\widetilde{X}_i)}{C_{K,A}}\right]\leq\frac{1}{C_{K,A}}\gamma(r^*)^2N=\gamma N.$$

In the second case, we decompose the supremum into two parts:

$$\begin{aligned}&\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^N\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\\&\qquad=\max\left(\sup_{Z\in\mathcal{B}_{K,A}:\|Z-Z^*\|_{L_2}\leq r^*}\sum_{i=1}^N\sigma_i\mathcal{L}_Z(\widetilde{X}_i),\sup_{Z\in\mathcal{B}_{K,A}:r^*\leq\|Z-Z^*\|_{L_2}\leq\sqrt{\frac{\nu K}{N}}}\sum_{i=1}^N\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right).\end{aligned}$$

Let $Z\in\mathcal{B}_{K,A}$ be such that $r^*\leq\|Z-Z^*\|_{L_2}\leq\sqrt{\frac{\nu K}{N}}$. Since $\mathcal{C}$ is star-shapped in Z^*, there exists $Z_0\in\mathcal{C}$ such that $\|Z_0-Z^*\|_{L_2}=r^*$ and $Z-Z^*=\kappa(Z_0-Z^*)$ for some $\kappa\geq 1$, so that $\kappa=\|Z-Z^*\|_{L_2}/\|Z_0-Z^*\|_{L_2}\leq(\nu K/N)(r^*)^{-1}$. Moreover, we have by linearity of $\mathcal{L}$ that $\mathcal{L}_{Z_0}=\kappa\mathcal{L}_Z$. Therefore, we obtain

$$\begin{aligned}\sup_{Z\in\mathcal{B}_{K,A}:r^*\leq\|Z-Z^*\|_{L_2}\leq\sqrt{\frac{\nu K}{N}}}\sum_{i=1}^N\sigma_i\mathcal{L}_Z(\widetilde{X}_i)&\leq\sup_{1\leq\kappa\leq\frac{1}{r^*}\sqrt{\frac{\nu K}{N}}}\sup_{Z_0\in\mathcal{B}_{K,A}:\|Z_0-Z^*\|_{L_2}\leq r^*}\sum_{i=1}^N\sigma_i\kappa\mathcal{L}_{Z_0}(\widetilde{X}_i)\\&=\sqrt{\frac{\nu K}{N}}\frac{1}{r^*}\sup_{Z_0\in\mathcal{B}_{K,A}:\|Z_0-Z^*\|_{L_2}\leq r^*}\sum_{i=1}^N\sigma_i\mathcal{L}_{Z_0}(\widetilde{X}_i).\end{aligned}$$

Since $C_{K,A}=\nu K/N\geq(r^*)^2$, we get, using the definition of r^*:

$$\begin{aligned}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^N\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right]&\leq\sqrt{\frac{\nu K}{N}}\frac{1}{r^*}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}:\|Z-Z^*\|_{L_2}\leq r^*}\sum_{i=1}^N\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right]\\&\leq\sqrt{\frac{\nu K}{N}}\frac{1}{r^*}\gamma(r^*)^2N\leq C_{K,A}\gamma N.\end{aligned}$$

Finally, we get that whatever the value of $C_{K,A}$ is:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^{N}\sigma_i\frac{\mathcal{L}_Z(\widetilde{X}_i)}{C_{K,A}}\right]\leq\gamma N.$$

Combining all these inequalities, we finally get that, with probability at least $1-\exp(-12K/625)$, for all $Z\in\mathcal{B}_{K,A}$,

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\right\}}\leq\frac{12}{25}K+\frac{1}{200}K+\frac{8\gamma}{\tau}K\leq\frac{49}{400}K,$$

from our choice of parameters. This concludes the proof. □

6.1.3.3 Proof of Theorem 15

The proof of this theorem follows the same lines as the one of the last Theorem 11 and 13: we start with identifying an event on which our estimator has the desired properties, and then we prove that this event holds with large probability.

For the sake of simplicity, we write $\hat{Z}$ for $\hat{Z}_K^{\mathrm{MOM}}$ and r^* for $r^*_{\mathrm{MOM,G}}(\gamma)$ for $\gamma=1/6400$. Consider A and $G:H\to\mathbb{R}$ such that Assumption 3.6 holds. Define $\mathcal{C}_{\gamma,G}:=\left\{Z\in\mathcal{C}:G(Z-Z^*)\leq(r^*)^2\right\}$. We consider the following event:

$$\Omega_K=\left\{\forall Z\in\mathcal{C}_{\gamma,G},\exists J\subset[N]:|J|>K/2\text{ and }\forall k\in J,\left|(P_{B_k}-P)\mathcal{L}_Z\right|\leq\frac{(r^*)^2}{4}\right\}.$$

We first show that on the event Ω_K, $\hat{Z}$ satisfies the statistical bounds announced in Theorem 15.

Lemma 58. *If there exists $\eta>0$ such that*

$$\sup_{Z\in\mathcal{C}\setminus\mathcal{C}_{\gamma,G}}\mathrm{MOM}_K(\ell_{Z^*}-\ell_Z)<-\eta\quad\text{and}\quad\sup_{Z\in\mathcal{C}_{\gamma,G}}\mathrm{MOM}_K(\ell_{Z^*}-\ell_Z)\leq\eta\tag{6.21}$$

then $G(\hat{Z}-Z^*)\leq(r^*)^2$.

Proof. Assume that (6.21) holds. Then:

$$\inf_{Z\in\mathcal{C}\setminus\mathcal{C}_{\gamma,G}}\mathrm{MOM}_K(\ell_Z-\ell_{Z^*})>\eta.\tag{6.22}$$

Moreover, define $Z\to T_K(Z)=\sup_{Z'\in\mathcal{C}}\mathrm{MOM}_K(\ell_Z-\ell_{Z'})$, we have

$$T_K(Z^*)=\max\left(\sup_{Z\in\mathcal{C}\setminus\mathcal{C}_{\gamma,G}}\mathrm{MOM}_K(\ell_{Z^*}-\ell_Z),\sup_{Z\in\mathcal{C}_{\gamma,G}}\mathrm{MOM}_K(\ell_{Z^*}-\ell_Z)\right)\leq\eta\tag{6.23}$$

and, by definition of $\hat{Z}$, we also have $T_K(\hat{Z})=\sup_{Z\in\mathcal{C}}\mathrm{MOM}_K(\ell_{\hat{Z}}-\ell_Z)\leq\sup_{Z\in\mathcal{C}}\mathrm{MOM}_K(\ell_{Z^*}-\ell_Z)=T_K(Z^*)\leq\eta$. However, by (6.22), any $Z\in\mathcal{C}\setminus\mathcal{C}_{\gamma,G}$ must satisfy

$$T_K(Z)\geq\mathrm{MOM}_K(\ell_Z-\ell_{Z^*})\geq\inf_{Z\in\mathcal{C}\setminus\mathcal{C}_{\gamma,G}}\mathrm{MOM}_K(\ell_Z-\ell_{Z^*})>\eta.$$

Therefore, we necessarily have $\hat{Z}\in\mathcal{C}\cap\mathcal{C}_{\gamma,G}$, that is $G(\hat{Z}-Z^*)\leq(r^*)^2$. □

Lemma 59. *Assume that $A<2$. On the event Ω_K, (6.21) holds with $\eta=(r^*)^2/4$.*

Proof. Let Z be such that $G(Z - Z^*) > (r^*)^2$. By the star-shaped property of $\mathcal{C}$ and the regularity property of G, there exist $Z_0 \in \partial\mathcal{C}_{\gamma,G}$ and $\alpha > 1$ such that $Z = Z^* + \alpha(Z_0 - Z^*)$. Since $G(Z_0 - Z^*) = (r^*)^2$, we have by Assumption 3.6 that $P\mathcal{L}_{Z_0} \geq A^{-1}G(Z_0 - Z^*)$. Moreover, on Ω_K, there are at least $K/2$ blocks B_k on which $\left|(P_{B_k} - P)\mathcal{L}_{Z_0}\right| \leq (r^*)^2/4$ and so $P_{B_k}\mathcal{L}_{Z_0} \geq P\mathcal{L}_{Z_0} - (r^*)^2/4 \geq A^{-1}G(Z_0 - Z^*) - (r^*)^2/4 \geq (r^*)^2/4$ since we assumed that $A < 2$. Now, by linearity of the loss function, we have on these blocks

$$P_{B_k}\mathcal{L}_Z = \alpha P_{B_k}\mathcal{L}_{Z_0} \geq \alpha(r^*)^2/4 > (r^*)^2/4.$$

We conclude that $\mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) < -(r^*)^2/4$. This being true for any $Z \in \mathcal{C}\backslash\mathcal{C}_{\gamma,G}$ we have:

$$\sup_{Z\in\mathcal{C}\backslash\mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq -\frac{(r^*)^2}{4}.$$

This shows the left-hand side inequality of (6.21) for $\eta = (r^*)^2/4$.

Next, let $Z \in \mathcal{C}$ be such that $G(Z - Z^*) \leq (r^*)^2$. On Ω_K, there are at least $K/2$ blocks B_k on which $\left|(P_{B_k} - P)\mathcal{L}_{Z_0}\right| \leq (r^*)^2/4$, that is $-P_{B_k}\mathcal{L}_Z \leq (r^*)^2/4 - P\mathcal{L}_Z \leq (r^*)^2/4$ since $P\mathcal{L}_Z \geq 0$ by definition of Z^*. Then, $\mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leq (r^*)^2/4$. This holds for all $Z \in \mathcal{C}_{\gamma,G}$, in other words, the right-hand side inequality of (6.21) holds for $\eta = (r^*)^2/4$ and this concludes the proof. □

Lemma 60. *Assume the conditions of Theorem 15 are met. Then, on Ω_K, $P\mathcal{L}_{\hat{Z}} \leq (r^*)^2/2$.*

Proof. From Assumption 3.6 combined with the fact that $A < 2$, we have from Lemmas 58 and 59 that $G(\hat{Z} - Z^*) \leq (r^*)^2$. Then on Ω_K there exist strictly more than $K/2$ blocks B_k on which $\left|(P_{B_k} - P)\mathcal{L}_{\hat{Z}}\right| \leq (r^*)^2/4$, that is:

$$P\mathcal{L}_{\hat{Z}} \leq P_{B_k}\mathcal{L}_{\hat{Z}} + \frac{(r^*)^2}{4}. \tag{6.24}$$

Moreover, by (6.23) and by definition of $\hat{Z}$, we have:

$$\begin{aligned}\mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) &\leq \sup_{Z\in\mathcal{C}} \mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_Z) \\ &\leq \sup_{Z\in\mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) = T_K(Z^*) \leq \eta = \frac{(r^*)^2}{4}.\end{aligned}$$

As a consequence, there exist at least $K/2$ blocks B_k on which $P_{B_k}(\ell_{\hat{Z}} - \ell_{Z^*}) \leq (r^*)^2/4$, that is:

$$P_{B_q}\mathcal{L}_{\hat{Z}} \leq \frac{(r^*)^2}{4}. \tag{6.25}$$

So there must be at least one block B_{k_0} on which (6.24) and (6.25) hold simultaneously. On this block, we have:

$$P\mathcal{L}_{\hat{Z}} \leq P_{B_{k_0}}\mathcal{L}_{\hat{Z}} + \frac{(r^*)^2}{4} \leq \frac{(r^*)^2}{4} + \frac{(r^*)^2}{4} = \frac{(r^*)^2}{2}.$$

□

At this stage of the proof, we have shown that on the event Ω_K, the estimator $\hat{Z}$ has the statistical bounds announced in Theorem 15. The final ingredient is to show that, under the conditions of Theorem 15, Ω_K holds with exponentially large probability. This is the purpose of the next result that can be proved using the same proof as the one of Lemma 53.

Lemma 61. *Assume the conditions of Theorem 15 are met, with $A < 2$. Then Ω_K holds with probability at least $1 - \exp(-72K/625)$.*

6.1.4 Proofs of Section 3.3.2: the Regularized MOM estimator

6.1.4.1 Proof of Theorem 18

The proof is structured in the same way as the previous ones: we identify an event on which $\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}$ has the desired statistical properties, then we show that this event holds with high probability. Let $\gamma = 1/32000$. Consider $\rho^* > 0$ such that ρ^* satisfies the sparsity equation of Definition 17. For the sake of simplicity, all along this proof we write $\hat{Z}$ for $\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}$ and $r^*_b := r^*_{\mathrm{RMOM,ER}}(\gamma, b\rho^*)$ for both $b \in \{1,2\}$. For $b \in \{1,2\}$, we define $\mathcal{B}_b := \left\{Z \in \mathcal{C} : P\mathcal{L}_Z \leq (r^*_b)^2 \text{ and } \|Z - Z^*\| \leq b\rho^*\right\}$. Then we define:

$$\Omega_K = \left\{ \forall b \in \{1,2\}, \forall Z \in \mathcal{B}_b, \exists J \subset [K], |J| > K/2, \forall k \in J, \left|(P_{B_k} - P)\mathcal{L}_Z\right| \leq \frac{(r^*_b)^2}{20} \right\}.$$

Finally, we consider $\lambda := (11/(40\rho^*))(r^*_2)^2$. We begin the proof by showing that on Ω_K, $\hat{Z}$ has the statistical properties announced in Theorem 18.

Lemma 62. *If there exists $\eta > 0$ such that*

$$\sup_{Z \in \mathcal{C}\backslash\mathcal{B}_2} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda\left(\|Z^*\| - \|Z\|\right) < -\eta \tag{6.26}$$

and

$$\sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda\left(\|Z^*\| - \|Z\|\right) \leq \eta \tag{6.27}$$

then $P\mathcal{L}_{\hat{Z}} \leq (r^*_2)^2$ and $\|\hat{Z} - Z^*\| \leq 2\rho^*$.

Proof. For $Z \in \mathcal{C}$, define $S(Z) = \sup_{Z' \in \mathcal{C}} \mathrm{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\| - \|Z'\|)$. For all $Z \in \mathcal{C}\backslash\mathcal{B}_2$ we have:

$$\begin{aligned} S(Z) &\geq \mathrm{MOM}_K(\ell_Z - \ell^*_Z) + \lambda(\|Z\| - \|Z^*\|) \\ &\geq \inf_{Z \in \mathcal{C}\backslash\mathcal{B}_2} \mathrm{MOM}_K(\ell_Z - \ell^*_Z) + \lambda(\|Z\| - \|Z^*\|) > \eta \end{aligned}$$

since (6.26) holds. Moreover, we have by definition of $\hat{Z}$:

$$S(\hat{Z}) \leq S(Z^*) = \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leq \eta$$

since (6.27) holds. This shows that necessarily $\hat{Z} \in \mathcal{B}_2$. □

We are now looking for $\eta > 0$ such that (6.26) and (6.27) hold, which the following Lemma allows us to do.

Lemma 63. *Under the assumptions of Theorem 18 and on the event event Ω_K, (6.26) and (6.27) hold with $\eta = 19(r^*_2)^2/50$.*

Proof. Let $b \in \{1,2\}$. Let $Z \in \mathcal{C}\backslash\mathcal{B}_b$. By the star-shaped property of $\mathcal{C}$, there exist $Z_0 \in \partial\mathcal{B}_b$ and $\alpha > 1$ such that $Z = Z^* + \alpha(Z_0 - Z^*)$. As a consequence, by linearity of the loss function and convexity of the regularization norm, for all $k \in [K]$ we have

$$\begin{aligned} P_{B_k}\mathcal{L}^\lambda_Z &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \\ &= \alpha P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|\alpha Z_0 + (1-\alpha)Z^*\| - \|Z^*\|) \\ &\geq \alpha P_{B_k}\mathcal{L}_{Z_0} + \lambda\alpha(\|Z_0\| - \|Z^*\|) = \alpha P_{B_k}\mathcal{L}^\lambda_{Z_0}. \end{aligned} \tag{6.28}$$

Now, since $Z_0 \in \partial\mathcal{B}_b$, we have either *a)* $P\mathcal{L}_{Z_0} = (r_b^*)^2$ and $\|Z_0 - Z^*\| < b\rho^*$ or *b)* $P\mathcal{L}_{Z_0} < (r_b^*)^2$ and $\|Z_0 - Z^*\| = b\rho^*$.

In the first case *a)*, on Ω_K, there are at least $K/2$ blocks B_k on which $P_{B_k}\mathcal{L}_{Z_0} \geq P\mathcal{L}_{Z_0} - (r_b^*)^2/20 = (19/20)(r_b^*)^2$. Therefore, on these blocs, we have

$$\begin{aligned} P_{B_k}\mathcal{L}^\lambda_{Z_0} &= P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \\ &\geq \frac{19}{20}(r_b^*)^2 - \lambda\|Z_0 - Z^*\| \\ &\geq \frac{19}{20}(r_b^*)^2 - \lambda b\rho^* \\ &= \frac{19}{20}(r_b^*)^2 - \frac{11b}{40}(r_2^*)^2 \geq \begin{cases} 2(r_2^*)^2/5 & \text{for } b = 2 \\ (r_2^*)^2/5 & \text{for } b = 1, \end{cases} \end{aligned} \tag{6.29}$$

where we used in the case $b = 1$ that $r_1^* \geq r_2^*/\sqrt{2}$ thanks to Proposition 7.1 from the Appendix.

In the second case *b)*, we have $Z_0 \in \bar{H}_{b\rho^*}$ from Definition 17. Since the sparsity equation holds for $\rho = \rho^*$, it also holds for $\rho = b\rho^*$ (see Proposition 7.2 in the Appendix). Let $V \in H$ be such that $\|Z^* - V\| \leq b\rho^*/20$ and $\Phi \in \partial\|.\|(V)$. We have:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geq \|Z_0\| - \|V\| - \|Z^* - V\| \\ &\geq \big\langle \Phi, Z_0 - V \big\rangle - \|Z^* - V\| \quad (\text{ since } \Phi \in \partial\|.\|(V)) \\ &= \big\langle \Phi, Z_0 - Z^* \big\rangle - \big\langle \Phi, V - Z^* \big\rangle - \|Z^* - V\| \\ &\geq \big\langle \Phi, Z_0 - Z^* \big\rangle - 2\|Z^* - V\| \quad (\text{ since } \big\langle \Phi, U \big\rangle \leq \|U\| \text{ for any } U \in H) \\ &\geq \big\langle \Phi, Z_0 - Z^* \big\rangle - \frac{b\rho^*}{10}. \end{aligned}$$

This is true for any $\Phi \in \bigcup_{V \in Z^* + b\rho^*/20} \partial\|.\|(V) = \Gamma_{Z^*}(b\rho^*)$. Then taking the sup over $\Gamma_{Z^*}(b\rho^*)$ gives:

$$\|Z_0\| - \|Z^*\| \geq \sup_{\Phi \in \Gamma_{Z^*}(b\rho^*)} \big\langle \Phi, Z_0 - Z^* \big\rangle - \frac{b\rho^*}{10}$$

and then taking the infimum over $\bar{H}_{b\rho^*}$ gives:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geq \inf_{Z_0 \in \bar{H}_{b\rho^*}} \|Z_0\| - \|Z^*\| \\ &\geq \inf_{Z_0 \in \bar{H}_{b\rho^*}} \sup_{\Phi \in \Gamma_{Z^*}(2\rho^*)} \big\langle \Phi, Z_0 - Z^* \big\rangle - \frac{b\rho^*}{10} \\ &= \Delta(b\rho^*) - \frac{b\rho^*}{10} \geq \frac{7}{10} b\rho^* \end{aligned} \tag{6.30}$$

where the last inequality holds since $b\rho^*$ satisfies the sparsity equation. Then, $\lambda(\|Z_0\| - \|Z^*\|) \geq (7/10)\lambda b\rho^* = (77/400)b(r_2^*)^2$. Now, since $Z_0 \in \mathcal{B}_b$, on Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leq (r_b^*)^2/20$ and so $P_{B_k}\mathcal{L}_{Z_0} \geq (r_b^*)^2/20$ - because $P\mathcal{L}_{Z_0} \geq 0$. Therefore, on the very same blocks,

$$\begin{aligned} P_{B_k}\mathcal{L}^\lambda_{Z_0} &= P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \\ &\geq -\frac{1}{20}(r_b^*)^2 + \frac{77}{400} b(r_2^*)^2 \geq \begin{cases} 134(r_2^*)^2/400 & \text{for } b = 2 \\ 29(r_2^*)^2/400 & \text{for } b = 1, \end{cases} \end{aligned} \tag{6.31}$$

where we used that $r_1^* \leq r_2^*$ (see Proposition 7.1 in the Appendix). As a consequence, it follows from (6.28), the fact that $\alpha > 1$, (6.29) and (6.31) for $b = 2$ that for all $Z \in \mathcal{C}\backslash\mathcal{B}_2$, on more than $K/2$ blocks B_k: $P_{B_k}\mathcal{L}^\lambda_Z \geq (134/400)(r_2^*)^2$ and so (6.26) holds for $\eta \leq (134/400)(r_2^*)^2$.

Let us now turn to Equation (6.27). Let $Z \in \mathcal{B}_1$. On Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leq (r_1^*)^2/20$. On these blocks B_k, all $P_{B_k}\mathcal{L}_Z^\lambda$'s are such that

$$\begin{aligned} P_{B_k}\mathcal{L}_Z^\lambda &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \\ &\geq P\mathcal{L}_Z - \frac{1}{20}(r_1^*)^2 - \lambda\|Z - Z^*\| \\ &\geq -\frac{1}{20}(r_1^*)^2 - \lambda\rho^* \\ &= -\frac{1}{20}(r_1^*)^2 - \frac{11}{40}(r_2^*)^2 \geq -\frac{13}{40}(r_2^*)^2 \end{aligned} \tag{6.32}$$

because $r_1^* \leq r_2^*$ (see Proposition 7.1 in the Appendix). Next, it follows from (6.28), the fact that $\alpha > 1$, (6.29) for $b = 1$, (6.31) for $b = 1$ and (6.32) that

$$\begin{aligned} \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) &\leq \max\left(\frac{-1}{5}, \frac{-29}{400}, \frac{13}{40}\right) r_2^2 \\ &= \frac{13}{40}(r_2^*)^2 \end{aligned} \tag{6.33}$$

and so (6.27) holds for $\eta \geq 13(r_2^*)^2/40$. As a consequence, (6.26) and (6.27) both hold for $\eta = 132(r_2^*)^2/400$. □

At this stage, we have shown that on the event Ω_K, the estimator $\hat{Z}$ has the statistical properties announced in Theorem 18. In what follows we prove that in the framework of Theorem 18, Ω_K holds with exponentially large probability.

Lemma 64. *Assume that $K \geq 100|\mathcal{O}|$, and let $\rho^* > 0$ be such that it satisfies the sparsity equation from Definition 17. Then, Ω_K holds with probability at least $1 - 2\exp(-72K/625)$.*

Proof. Consider

$$\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t \geq 1\}} + 2(t - 1/2)\mathbb{1}_{\{1/2 \leq t \leq 1\}},$$

so that for any $t \in \mathbb{R}$,

$$\mathbb{1}_{\{t \geq 1\}} \leq \phi(t) \leq \mathbb{1}_{\{t \geq 1/2\}}.$$

For $k \in [K]$, let $W_k := \{X_i : i \in B_k\}$ and $F_Z(W_k) = (P_{B_k} - P)\mathcal{L}_Z$. We also define the counterparts of these quantities constructed with the non-corrupted vectors: $\widetilde{W}_k := \left\{\widetilde{X}_i : i \in B_k\right\}$ and $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - \widetilde{P})\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := \frac{K}{N}\sum_{i \in B_k}\mathcal{L}_Z(\widetilde{X}_i)$ and $\widetilde{P}\mathcal{L}_Z := \mathbb{E}[\mathcal{L}_Z(\widetilde{X}_i)]$. For both $b \in \{1, 2\}$, define

$$Z \to \psi_b(Z) = \sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(W_k)| \leq \frac{(r_b^*)^2}{20}\right\}}.$$

Consider $b \in \{1, 2\}$. We want to show that, with high probability, if $Z \in \mathcal{B}_b$, then $\psi_b(Z) > K/2$ which follows if one can proves that

$$\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{20}\right\}} \leq \frac{49K}{100}. \tag{6.34}$$

Indeed, consider $Z \in \mathcal{B}_b$ such that (6.34) holds. Then, there exist at least $(1 - 49/100)K = 51K/100$ blocks B_k on which $|F_Z(\widetilde{W}_k)| \leq (r_b^*)^2/20$. On the other hand, we know that $|\mathcal{O}| \leq K/100$, so that among the $51K/100$ previous blocks, at most $K/100$ contains corrupted data. The other $50K/100 = K/2$ contain only non-corrupted data, so we have $F_Z(\widetilde{W}_k) = F_Z(W_k)$ on

these block and so $\psi_b(Z) > K/2$.

Let $Z \in \mathcal{B}_b$. We have:

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \\
&= \sum_{k\in[K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right)\right] \\
&= \sum_{k\in[K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{40}\right\}}\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leq \sum_{k\in[K]} \left(\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leq \sup_{Z\in\mathcal{B}_b} \left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right)
\end{aligned} \tag{6.35}$$

We start with bounding the last sum in the previous inequality. For each $k \in [K]$, Markov's inequality and the definition of r_b^* yield to

$$\begin{aligned}
\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) &\leq \frac{1600}{(r_b^*)^4}\mathbb{E}\left[F_Z(\widetilde{W}_k)^2\right] \\
&= \frac{1600}{(r_b^*)^4}\left(\frac{K}{N}\right)\mathrm{Var}(\mathcal{L}_Z(\widetilde{X})) \\
&\leq \frac{1600}{(r_b^*)^4}\left(V_K(r_b^*)\right)^2 \leq \frac{1}{200}
\end{aligned}$$

Plugging this last result into (6.35), we get:

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \leq \frac{K}{200} + \sup_{Z\in\mathcal{B}_b} \left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right). \tag{6.36}$$

We now we have to bound this last term. Using Mc Diarmind inequality (Theorem 6.2 in Boucheron et al. (2013a) with $t = 12/25$), we get that with probability at least $1 - \exp(-72K/625)$, for all $Z \in \mathcal{B}_b$,

$$\begin{aligned}
&\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) \\
&\qquad \leq \frac{12K}{25} + \mathbb{E}\left[\sup_{Z\in\mathcal{B}_b} \sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right].
\end{aligned} \tag{6.37}$$

Let now $\epsilon_1, \ldots, \epsilon_K$ be Rademacher variables independant from the $\widetilde{X}_i$'s. By the symmetrization

Lemma, we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right]$$
$$\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]. \tag{6.38}$$

As ϕ is Lipschitz with $\phi(0)=0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), chapter 4) to get that:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\frac{20F_Z(\widetilde{W}_k)}{(r_b^*)^2}\right]$$
$$=\frac{40}{(r_b^*)^2}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-\widetilde{P})\mathcal{L}_Z\right] \tag{6.39}$$

Now, let $(\sigma_i)_{i=1,\dots,N}$ be a family of Rademacher variables independant from the $\widetilde{X}_i$'s and the ϵ_i's. Using the symmetrization Lemma again, we get:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-\widetilde{P})\mathcal{L}_Z\right]\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\frac{K}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right]$$
$$\leq 2KE(r_b^*,b\rho^*)\leq 2K\gamma(r_b^*)^2.$$

Combining this with (6.37), (6.38) and (6.39), we finally get that, with probability at least $1-\exp(-72K/625)$:

$$\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\leq\left(\frac{12}{25}+160\gamma\right)K. \tag{6.40}$$

Plugging that into (6.36), we conclude that, with probability at least $1-\exp(-72K/625)$, for all $Z\in\mathcal{B}_b$,

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}}\leq\left(\frac{1}{200}+\frac{12}{25}+160\gamma\right)K\leq\frac{49}{100}K$$

for our choice of parameters. Now, in order for Ω_K to hold, this inequality must be verified for both $b=1$ and 2. Then, we finally conclude that Ω_K holds with probability $1-2\exp(-72K/625)$, which concludes the proof.

□

6.1.4.2 Proof of Theorem 21

Let $K>0$ be a divisor of N such that $K\geq 100|\mathcal{O}|$. Let $\gamma=1/32000$. Let $A\in(0,1]$ and $\rho^*>0$ be such that Assumption 3.7 holds and satisfying the sparsity equation from Definition 20. Define $\nu=320000A^2$.

For the sake of simplicity, we write all along this proof $\hat{Z}$ for $\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}$. For $b\in\{1,2\}$, we define $r_b^*=r^*_{\mathrm{RMOM},L_2}(\gamma,b\rho^*)$,

$$C_{K,b}:=\max\left(\nu\frac{K}{N},(r_b^*)^2\right)=C_K(\gamma,b\rho^*,A),$$

and the localized models $\mathcal{B}_{K,b} := \left\{Z \in \mathcal{C} : \|Z - Z^*\| \leq b\rho^* \text{ and } \|Z - Z^*\|_{L_2} \leq \sqrt{C_{K,b}}\right\}$ - we recall that the L_2-norm associated with the good data $\widetilde{X}$ is defined as $\|Z\|_{L_2} = \mathbb{E}[\langle \widetilde{X}, Z\rangle^2]^{1/2}$. With these notation, we have $\lambda := (11/(40\rho^*))C_{K,2}$. Finally, we define the event onto which $\hat{Z}$ will have the desired properties:

$$\Omega_K = \left\{ \forall b \in \{1,2\}, \forall Z \in \mathcal{B}_{K,b}, \quad \sum_{k=1}^{K} \mathbb{1}_{\left\{|(P_{B_k} - P)\mathcal{L}_Z| \leq \frac{C_{K,b}}{20}\right\}} > \frac{K}{2} \right\}.$$

First, we show that on Ω_K, $\hat{Z}$ has the statistical properties announced in Theorem 21. Then, we show that Ω_K holds with high probability.

Lemma 65. *If there exists $\eta > 0$ such that*

$$\sup_{Z \in \mathcal{C}\setminus\mathcal{B}_{K,2}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) < -\eta \tag{6.41}$$

and

$$\sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leq \eta, \tag{6.42}$$

then $\|Z - Z^*\| \leq 2\rho^*$ and $\|Z - Z^*\|_{L_2} \leq \sqrt{C_{K,2}}$.

Proof. Assume that such an η exists. For $Z \in \mathcal{C}$, define $S(Z) = \sup_{Z' \in \mathcal{C}} \mathrm{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\| - \|Z'\|)$. For $Z \in \mathcal{C}\setminus\mathcal{B}_{K,2}$ we have:

$$\begin{aligned} S(Z) &\geq \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) + \lambda(\|Z\| - \|Z^*\|) \\ &\geq \inf_{Z \in \mathcal{C}\setminus\mathcal{B}_{K,2}} \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) + \lambda(\|Z\| - \|Z^*\|) > \eta \end{aligned}$$

since (6.41) holds. Moreover, we have by definition of $\hat{Z}$:

$$S(\hat{Z}) \leq S(Z^*) = \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leq \eta,$$

since (6.42) holds. This shows that necessarily $\hat{Z} \in \mathcal{B}_{K,2}$. □

We are now looking for $\eta > 0$ such that (6.41) and (6.42) hold. In the following result we identify such a η on the event Ω_K.

Lemma 66. *Under the conditions of Theorem 21 and on the event Ω_K, (6.41) and (6.42) hold with $\eta = 33C_{K,2}/100$.*

Proof. Consider $b \in \{1,2\}$ and $Z \in \mathcal{C}\setminus\mathcal{B}_{K,b}$. From the star-shaped property of $\mathcal{C}$, we have the existence of $Z_0 \in \partial\mathcal{B}_{K,b}$ and $\alpha > 1$ such that $Z = Z^* + \alpha(Z_0 - Z^*)$. As a consequence, by linearity of the loss function and convexity of the regularization norm, for all $k \in [K]$ we have

$$\begin{aligned} P_{B_k}\mathcal{L}_Z^\lambda &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \\ &= \alpha P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|\alpha Z_0 + (1-\alpha)Z^*\| - \|Z^*\|) \\ &\geq \alpha P_{B_k}\mathcal{L}_{Z_0} + \lambda\alpha(\|Z_0\| - \|Z^*\|) = \alpha P_{B_k}\mathcal{L}_{Z_0}^\lambda. \end{aligned} \tag{6.43}$$

Now, since $Z_0 \in \partial\mathcal{B}_{K,b}$, we have either *a)* $\|Z_0 - Z^*\|_{L_2} = \sqrt{C_{K,b}}$ and $\|Z_0 - Z^*\| < b\rho^*$ or *b)* $\|Z_0 - Z^*\|_{L_2} < \sqrt{C_{K,b}}$ and $\|Z_0 - Z^*\| = b\rho^*$.

In the first case *a)*, on Ω_K, there are at least $K/2$ blocks B_k on which $P_{B_k}\mathcal{L}_{Z_0} \geq P\mathcal{L}_{Z_0} - C_{K,b}/(20)$. But from Assumption 3.7, we have in this case that $AP\mathcal{L}_{Z_0} \geq \|Z_0 - Z^*\|_{L_2}^2 = C_{K,b}$,

so that, on the same blocks of data, $P_{B_k}\mathcal{L}_{Z_0} \geq (1/A)C_{K,b} - (1/20)C_{K,b} \geq (19/20)C_{K,b}$, since we assumed that $0 < A \leq 1$. Therefore, on these blocs, we have

$$\begin{aligned} P_{B_k}\mathcal{L}^{\lambda}_{Z_0} &= P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \\ &\geq \frac{19}{20}C_{K,b} - \lambda\|Z_0 - Z^*\| \\ &\geq \frac{19}{20}C_{K,b} - \lambda b\rho^* \\ &= \frac{19}{20}C_{K,b} - \frac{11b}{40}C_{K,2}. \end{aligned}$$

But thanks to Proposition 7.1 from the Appendix, we have that $r_1^* \geq r_2^*/\sqrt{2}$, from which we deduce that $C_{K,1} \geq C_{K,2}/2$. As a consequence, on the previous blocks, we have

$$P_{B_k}\mathcal{L}^{\lambda}_{Z_0} \geq \begin{cases} 7C_{K,2}/40 & \text{for } b = 1 \\ 16C_{K,2}/40 & \text{for } b = 2. \end{cases} \tag{6.44}$$

In the second case *b)*, we have $Z_0 \in \widetilde{H}_{b\rho^*,A}$ from Definition 20. Since the sparsity equation is satisfied by ρ^*, it is also satisfied by $b\rho^*$ as well (see Proposition 7.2 in the Appendix). Let $V \in H$ be such that $\|Z^* - V\| \leq b\rho^*/20$ and $\Phi \in \partial\|.\|(V)$. We have:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geq \|Z_0\| - \|V\| - \|Z^* - V\| \\ &\geq \big\langle \Phi, Z_0 - V \big\rangle - \|Z^* - V\| \quad (\text{ since } \Phi \in \partial\|.\|(V)) \\ &= \big\langle \Phi, Z_0 - Z^* \big\rangle - \big\langle \Phi, V - Z^* \big\rangle - \|Z^* - V\| \\ &\geq \big\langle \Phi, Z_0 - Z^* \big\rangle - 2\|Z^* - V\| \quad (\text{ since } \big\langle \Phi, U \big\rangle \leq \|U\| \text{ for any } U \in H) \\ &\geq \big\langle \Phi, Z_0 - Z^* \big\rangle - \frac{b\rho^*}{10}. \end{aligned}$$

This is true for any $\Phi \in \underset{V \in Z^* + b\rho^*/20}{\cup} \partial\|.\|(V) = \Gamma_{Z^*}(b\rho^*)$. Then taking the sup over $\Gamma_{Z^*}(b\rho^*)$ gives:

$$\|Z_0\| - \|Z^*\| \geq \sup_{\Phi \in \Gamma_{Z^*}(b\rho^*)} \big\langle \Phi, Z_0 - Z^* \big\rangle - \frac{b\rho^*}{10}$$

and then taking the infimum over $\widetilde{H}_{b\rho^*,A}$ gives:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geq \inf_{Z_0 \in \widetilde{H}_{b\rho^*,A}} \|Z_0\| - \|Z^*\| \\ &\geq \inf_{Z_0 \in \widetilde{H}_{b\rho^*,A}} \sup_{\Phi \in \Gamma_{Z^*}(2\rho^*)} \big\langle \Phi, Z_0 - Z^* \big\rangle - \frac{b\rho^*}{10} \\ &= \Delta(b\rho^*) - \frac{b\rho^*}{10} \geq \frac{7}{10}b\rho^* \end{aligned} \tag{6.45}$$

where the last inequality holds since $b\rho^*$ satisfies the sparsity equation. Then, $\lambda(\|Z_0\| - \|Z^*\|) \geq (7/10)\lambda b\rho^* = (77/400)bC_{K,2}$. Now, since $Z_0 \in \mathcal{B}_{K,b}$, on Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leq C_{K,b}/(20)$ and so $P_{B_k}\mathcal{L}_{Z_0} \geq -C_{K,b}/(20)$ (because $P\mathcal{L}_{Z_0} \geq 0$). Therefore, on the very same blocks,

$$\begin{aligned} P_{B_k}\mathcal{L}^{\lambda}_{Z_0} &= P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \\ &\geq -\frac{1}{20}C_{K,b} + \frac{77b}{400}C_{K,2} \geq \begin{cases} 57(r_2^*)^2/400 & \text{for } b = 1 \\ 134(r_2^*)^2/400 & \text{for } b = 2, \end{cases} \end{aligned} \tag{6.46}$$

where we used that $C_{K,1} \leq C_{K,2}$ because $r_1^* \leq r_2^*$ (see Proposition 7.1 in the Appendix). As a consequence, it follows from (6.43), the fact that $\alpha > 1$, (6.44) and (6.46) for $b = 2$ that, for

all $Z \in \mathcal{C} \backslash \mathcal{B}_{K,2}$, on more than $K/2$ blocks B_k: $P_{B_k}\mathcal{L}_Z^\lambda \geq (134/400)C_{K,2}$ and so (6.41) holds for $\eta < (134/400)C_{K,2}$.

Let us now turn to Equation (6.42). Let $Z \in \mathcal{B}_{K,1}$. On Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leq C_{K,1}/20$. On these blocks B_k, all $P_{B_k}\mathcal{L}_Z^\lambda$'s are such that

$$\begin{aligned} P_{B_k}\mathcal{L}_Z^\lambda &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \\ &\geq P\mathcal{L}_Z - \frac{1}{20}C_{K,1} - \lambda\|Z - Z^*\| \\ &\geq -\frac{1}{20}C_{K,1} - \lambda\rho^* \\ &= -\frac{1}{20}C_{K,1} - \frac{11}{40}C_{K,2} \geq -\frac{13}{40}C_{K,2}, \end{aligned} \tag{6.47}$$

where we used the fact that, thanks to Proposition 7.1 in the Appendix, $C_{K,1} \leq C_{K,2}$. Next, it follows from (6.43), the fact that $\alpha > 1$, (6.44) and (6.43) for $b = 1$ and (6.47) that

$$\begin{aligned} \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) &\leq \max\left(\frac{-7}{40}, \frac{-57}{400}, \frac{13}{40}\right) C_{K,2} \\ &= \frac{13}{40}C_{K,2} \end{aligned} \tag{6.48}$$

and so (6.42) holds for $\eta \geq 13C_{K,2}/40$. As a consequence, (6.41) and (6.42) both hold for $\eta = 132C_{K,2}/400$.

□

From Lemmas 65 and 66, we conclude that on the event Ω_K, $\hat{Z} \in \mathcal{B}_{K,2}$. We use this information to upper bound the excess risk of $\hat{Z}$ in the following result.

Lemma 67. *Under the conditions of Theorem 21 and on the event Ω_K, we have*

$$P\mathcal{L}_{\hat{Z}} \leq \frac{27}{100}C_{K,2}.$$

Proof. From Lemmas 65 and 66, we have that $\hat{Z} \in \mathcal{B}_{K,2}$. On Ω_K, this implies the existence of stricly more than $K/2$ blocks B_k on which

$$P\mathcal{L}_{\hat{Z}} \leq P_{B_k}\mathcal{L}_{\hat{Z}} + \frac{C_{K,2}}{20}. \tag{6.49}$$

Now, by definition of $\hat{Z}$, (6.42) and Lemma 66 we get

$$\begin{aligned} \mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) + \lambda(\|\hat{Z}\| - \|Z^*\|) &\leq \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \\ &\leq 33\frac{C_{K,2}}{100}. \end{aligned}$$

This means that there exist at least $K/2$ blocks B_k on which $P_{B_k}\mathcal{L}_{\hat{Z}} + \lambda(\|\hat{Z}\| - \|Z^*\|) \leq 33C_{K,2}/100$. Since $\lambda(\|Z^*\| - \|\hat{Z}\|) \leq \lambda\|Z^* - \hat{Z}\| \leq 2\lambda\rho^* = 11C_{K,2}/20$, we have on these blocks

$$P_{B_k}\mathcal{L}_{\hat{Z}} \leq 33\frac{C_{K,2}}{100} + 11\frac{C_{K,2}}{20} = 22\frac{C_{K,2}}{100}. \tag{6.50}$$

Therefore, there exist at least a block B_{k_0} on which (6.49) and (6.50) hold simultaneously. On this block, we can write

$$P\mathcal{L}_{\hat{Z}} \leq P_{B_{k_0}}\mathcal{L}_{\hat{Z}} + \frac{C_{K,2}}{20} \leq 22\frac{C_{K,2}}{100} + \frac{C_{K,2}}{20} = 27\frac{C_{K,2}}{100}.$$

□

At this stage, we have shown that on the event Ω_K, the regularized minmax MOM-estimator $\hat{Z}$ has the statistical properties announced in Theorem 21. In what follows, we prove that, in the framework of Theorem 21, Ω_K holds with exponentially large probability.

Proposition 6.1. Consider ρ^* that satisfies the sparsity equation from Definition 20. Assume that $K \geq 100|\mathcal{O}|$. Then, Ω_K holds with probability at least $1 - 2\exp(-72K/625)$.

Proof. Consider $b \in \{1, 2\}$. Define

$$\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t \geq 1\}} + 2(t - 1/2)\mathbb{1}_{\left\{1/2 \leq t \leq 1\right\}},$$

so that for any $t \in \mathbb{R}$,

$$\mathbb{1}_{\{t \geq 1\}} \leq \phi(t) \leq \mathbb{1}_{\left\{t \geq 1/2\right\}}.$$

For $k \in [K]$, let $W_k := \{X_i : i \in B_k\}$ and $F_Z(W_k) = (P_{B_k} - P)\mathcal{L}_Z$. We also define the counterparts of these quantities constructed with the non-corrupted vectors: $\widetilde{W}_k := \left\{\widetilde{X}_i : i \in B_k\right\}$ and $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - P)\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := (K/N)\sum_{i \in B_k} \mathcal{L}_Z(\widetilde{X}_i)$. For $b \in \{1, 2\}$, define

$$\psi_b(Z) = \sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(W_k)| \leq \frac{C_{K,b}}{20}\right\}}.$$

We would like to show that, if $Z \in \mathcal{B}_{K,b}$, then $\psi_b(Z) > K/2$ with high probability. As we showed in the proof of Lemma 53, in our framework this is true if we show that with high probability, for all $Z \in \mathcal{B}_{K,b}$,

$$\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} \leq \frac{49K}{100}, \tag{6.51}$$

and this is what we do now. Let $Z \in \mathcal{B}_{K,b}$. We have:

$$\begin{aligned}
&\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} \\
&= \sum_{k \in [K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right)\right] \\
&= \sum_{k \in [K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right\}}\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) \\
&\leq \sum_{k \in [K]} \left(\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) \\
&\leq \sup_{Z \in \mathcal{B}_{K,b}} \left(\sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right)
\end{aligned} \tag{6.52}$$

We start with bounding the last sum in the previous inequality. For each $k \in [K]$, Markov's

inequality and the definition of $C_{K,b}$ yield to

$$\begin{aligned}
\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) &\leq \left(\frac{40}{C_{K,b}}\right)^2 \mathbb{E}\left[|F_Z(\widetilde{W}_k)|^2\right] \\
&= \left(\frac{40}{C_{K,b}}\right)^2 \left(\frac{K}{N}\right)^2 \mathbb{E}\left[\left(\sum_{i\in B_k} \mathcal{L}_Z(\widetilde{X}_i) - \mathbb{E}\left[\mathcal{L}_Z(\widetilde{X}_i)\right]\right)^2\right] \\
&\leq \left(\frac{40}{C_{K,b}}\right)^2 \frac{K}{N}\|Z - Z^*\|_{L_2}^2 \\
&\leq \left(\frac{40}{C_{K,b}}\right)^2 \frac{K}{N} C_{K,b} \leq 40^2\nu^{-1} = \frac{1}{200}.
\end{aligned}$$

Plugging this last result into (6.52), we get:

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} \\
&\quad\leq \frac{K}{200} + \sup_{Z\in\mathcal{B}_{K,b}} \left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right). \qquad (6.53)
\end{aligned}$$

We now have to bound this last term. Using Mc Diarmind inequality (Theorem 6.2 in Boucheron et al. (2013a) with $t = 12/25$), we get that with probability at least $1 - \exp(-72K/625)$, for all $Z \in \mathcal{B}_{K,b}$,

$$\begin{aligned}
&\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) \leq \\
&\quad\frac{12K}{25} + \mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}} \sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]. \qquad (6.54)
\end{aligned}$$

Let now $\epsilon_1, \ldots, \epsilon_K$ be Rademacher variables independant from the $\widetilde{X}_i$'s. By the symmetrization Lemma, we have:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}} \sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right] \\
&\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}} \sum_{k\in[K]} \epsilon_k \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]. \qquad (6.55)
\end{aligned}$$

As ϕ is Lipschitz with $\phi(0) = 0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), chapter 4) to get that:

$$\begin{aligned}
\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}} \sum_{k\in[K]} \epsilon_k \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right] &\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}} \sum_{k\in[K]} \epsilon_k \frac{20F_Z(\widetilde{W}_k)}{C_{K,b}}\right] \\
&= \frac{40}{C_{K,b}} \mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}} \sum_{k\in[K]} \epsilon_k (\widetilde{P_{B_k}} - \widetilde{P})\mathcal{L}_Z\right] \qquad (6.56)
\end{aligned}$$

Now, let $(\sigma_i)_{i=1,\dots,N}$ be a family of Rademacher variables independant from the $\widetilde{X}_i$'s and the ϵ_i's. Using the symmetrization Lemma again, we get:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-\widetilde{P})\mathcal{L}_Z\right] \leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,b}}\frac{K}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right]$$
$$\leq 2K(r_b^*)^2 \leq 2K\gamma C_{K,b}.$$

Combining this with (6.54), (6.55) and (6.56), we finally get that, with probability at least $1-\exp(-72K/625)$:

$$\sup_{Z\in\mathcal{B}_{K,b}}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\leq\left(\frac{12}{25}+160\gamma\right)K \tag{6.57}$$

Plugging that into (6.53), we conclude that, with probability at least $1-\exp(-72K/625)$, for all $Z\in\mathcal{B}_{K,b}$,

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{C_{K,b}}{20}\right\}}\leq\left(\frac{1}{200}+\frac{12}{25}+160\gamma\right)K\leq\frac{49}{100}K$$

for our choice of parameters. Now, in order for Ω_K to hold, this inequality must be verified for both $b=1$ and 2. Then, we finally conclude that Ω_K holds with probability $1-2\exp(-72K/625)$, which concludes the proof.

□

6.1.4.3 Proof of Theorem 24

The proof is structured in the same way as the previous ones: we identify an event on which $\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}$ has the desired statistical properties, then we show that this event holds with high probability. We place ourselves under the conditions of Theorem 24, i.e., we assume the existence of $A\in(0,1]$ such that Assumption 3.8 holds, $\gamma=1/32000$ and ρ^* which satisfies the sparsity equation from Definition 23. For $b\in\{1,2\}$ we define $r_b^*=r^*_{\mathrm{RMOM,G}}(\gamma,2\rho^*)$ and

$$\mathcal{B}_b:=\left\{Z\in\mathcal{C}:G(Z-Z^*)\leq(r_b^*)^2 \text{ and } \|Z-Z^*\|\leq b\rho^*\right\}.$$

With these notation, $\lambda=(11/(40\rho^*))r_2^*$. We consider the event

$$\Omega_{K,G}=\left\{\forall b\in\{1,2\},\forall Z\in\mathcal{B}_b,\ \sum_{k=1}^{K}\mathbb{1}\left(\left|(P_{B_k}-P)\mathcal{L}_Z\right|\leq\frac{1}{20}(r_b^*)^2\right)>\frac{K}{2}\right\}.$$

For the sake of simplicity, in the rest of the proof we write $\hat{Z}=\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}$.

Lemma 68. *If there exists $\eta>0$ such that*

$$\sup_{Z\in\mathcal{C}\setminus\mathcal{B}_2}\mathrm{MOM}_K(\ell_{Z^*}-\ell_Z)+\lambda(\|Z^*\|-\|Z\|)<-\eta \tag{6.58}$$

and

$$\sup_{Z\in\mathcal{C}}\mathrm{MOM}_K(\ell_{Z^*}-\ell_Z)+\lambda(\|Z^*\|-\|Z\|)\leq\eta \tag{6.59}$$

then $\|Z-Z^*\|\leq 2\rho^*$ and $G(Z-Z^*)\leq r^*_{\mathrm{RMOM,G}}(\gamma,2\rho^*)^2$.

Proof. Let η be such that (6.58) and (6.59) hold. For all $Z \in \mathcal{C}$, define $S(Z) = \sup_{Z' \in \mathcal{C}} \mathrm{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\| - \|Z'\|)$. It follows from (6.58) that for all $Z \in \mathcal{C} \backslash \mathcal{B}_2$,

$$S(Z) \geq \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) + \lambda(\|Z\| - \|Z^*\|) > \eta$$

Moreover, it follows from the definition of $\hat{Z}$ and (6.59) that

$$S(\hat{Z}) \leq S(Z^*) = \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leq \eta$$

This shows that necessarily $\hat{Z} \in \mathcal{B}_2$. □

Lemma 69. *Under the conditions of Theorem 24 and on the event $\Omega_{K,G}$, (6.58) and (6.59) hold with $\eta = (33/100)(r_2^*)^2$.*

Proof. Let $b \in \{1, 2\}$. Let $Z \in \mathcal{C} \backslash \mathcal{B}_b$. By the star-shaped property of $\mathcal{C}$ and the regularity property of G, there exist $Z_0 \in \partial \mathcal{B}_b$ and $\alpha > 1$ such that $Z = Z^* + \alpha(Z_0 - Z^*)$. As a consequence, by linearity of the loss function and convexity of the regularization norm, for all $k \in [K]$ we have

$$\begin{aligned} P_{B_k} \mathcal{L}_Z^\lambda &= P_{B_k} \mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \\ &= \alpha P_{B_k} \mathcal{L}_{Z_0} + \lambda(\|\alpha Z_0 + (1-\alpha) Z^*\| - \|Z^*\|) \\ &\geq \alpha P_{B_k} \mathcal{L}_{Z_0} + \lambda \alpha (\|Z_0\| - \|Z^*\|) = \alpha P_{B_k} \mathcal{L}_{Z_0}^\lambda. \end{aligned} \tag{6.60}$$

Now, since $Z_0 \in \partial \mathcal{B}_b$, we have either *a)* $G(Z_0 - Z^*) = (r_b^*)^2$ and $\|Z_0 - Z^*\| < b\rho^*$ or *b)* $G(Z_0 - Z^*) < (r_b^*)^2$ and $\|Z_0 - Z^*\| = b\rho^*$.

In the first case *a)*, on $\Omega_{K,G}$, there are at least $K/2$ blocks B_k on which $P_{B_k} \mathcal{L}_{Z_0} \geq P\mathcal{L}_{Z_0} - (r_b^*)^2/(20)$. But we also have from Assumption 3.8 that $AP\mathcal{L}_{Z_0} \geq G(Z_0 - Z^*) = (r_b^*)^2$, so that $P_{B_k} \mathcal{L}_{Z_0} \geq (1/A)(r_b^*)^2 - (1/20)(r_b^*)^2 \geq (19/20)(r_b^*)^2$, since we assumed that $A \leq 1$. Therefore, on these blocs, we have

$$\begin{aligned} P_{B_k} \mathcal{L}_{Z_0}^\lambda &= P_{B_k} \mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \\ &\geq \frac{19}{20}(r_b^*)^2 - \lambda \|Z_0 - Z^*\| \\ &\geq \frac{19}{20}(r_b^*)^2 - \lambda b \rho^* \\ &= \frac{19}{20}(r_b^*)^2 - \frac{11b}{40}(r_2^*)^2 \geq \begin{cases} (r_2^*)^2/5 & \text{for } b = 1 \\ 2(r_2^*)^2/5 & \text{for } b = 2, \end{cases} \end{aligned} \tag{6.61}$$

where we used in the case $b = 1$ that $r_1^* \geq r_2^*/\sqrt{2}$ thanks to Proposition 7.1 from the Appendix.

In the second case *b)*, we have $Z_0 \in \bar{H}_{b\rho^*}$ from Definition 23. Since the sparsity equation holds for $\rho = \rho^*$, it also holds for $\rho = b\rho^*$ (see Proposition 7.2 in the Appendix). Let $V \in H$ be such that $\|Z^* - V\| \leq b\rho^*/20$ and $\Phi \in \partial \|.\|(V)$. We have:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geq \|Z_0\| - \|V\| - \|Z^* - V\| \\ &\geq \langle \Phi, Z_0 - V \rangle - \|Z^* - V\| \quad (\text{ since } \Phi \in \partial \|.\|(V)) \\ &= \langle \Phi, Z_0 - Z^* \rangle - \langle \Phi, V - Z^* \rangle - \|Z^* - V\| \\ &\geq \langle \Phi, Z_0 - Z^* \rangle - 2\|Z^* - V\| \quad (\text{ since } \langle \Phi, U \rangle \leq \|U\| \text{ for any } U \in H) \\ &\geq \langle \Phi, Z_0 - Z^* \rangle - \frac{b\rho^*}{10}. \end{aligned}$$

This is true for any $\Phi \in \underset{V \in Z^* + b\rho^*/20}{\cup} \partial\|.\|(V) = \Gamma_{Z^*}(b\rho^*)$. Then taking the sup over $\Gamma_{Z^*}(b\rho^*)$ gives:

$$\|Z_0\| - \|Z^*\| \geq \sup_{\Phi \in \Gamma_{Z^*}(b\rho^*)} \langle \Phi, Z_0 - Z^* \rangle - \frac{b\rho^*}{10}$$

and then taking the infimum over $\bar{H}_{b\rho^*,A}$ gives:

$$\begin{aligned}
\|Z_0\| - \|Z^*\| &\geq \inf_{Z_0 \in \bar{H}_{b\rho^*,A}} \|Z_0\| - \|Z^*\| \\
&\geq \inf_{Z_0 \in \bar{H}_{b\rho^*,A}} \sup_{\Phi \in \Gamma_{Z^*}(2\rho^*)} \langle \Phi, Z_0 - Z^* \rangle - \frac{b\rho^*}{10} \\
&= \Delta(b\rho^*) - \frac{b\rho^*}{10} \geq \frac{7}{10} b\rho^* \qquad (6.62)
\end{aligned}$$

where the last inequality holds since $b\rho^*$ satisfies the sparsity equation. Then, $\lambda(\|Z_0\| - \|Z^*\|) \geq (7/10)\lambda b\rho^* = (77/400)b(r_2^*)^2$. Now, since $Z_0 \in \mathcal{B}_b$, on $\Omega_{K,G}$ there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leq (r_b^*)^2/(20)$ and so $P_{B_k}\mathcal{L}_{Z_0} \geq -(r_b^*)^2/(20)$ (because $P\mathcal{L}_{Z_0} \geq 0$). Therefore, on the very same blocks,

$$\begin{aligned}
P_{B_k}\mathcal{L}_{Z_0}^\lambda &= P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \\
&\geq -\frac{1}{20}(r_b^*)^2 + \frac{77}{400} b(r_2^*)^2 \geq \begin{cases} 57(r_2^*)^2/(400) & \text{for } b = 1 \\ 134(r_2^*)^2/(400) & \text{for } b = 2, \end{cases} \qquad (6.63)
\end{aligned}$$

where we used that $r_1^* \leq r_2^*$ (see Proposition 7.1 in the Appendix). As a consequence, it follows from (6.60), the fact that $\alpha > 1$, (6.61) and (6.63) for $b = 2$ that, for all $Z \in \mathcal{C} \backslash \mathcal{B}_2$, on more than $K/2$ blocks B_k: $P_{B_k}\mathcal{L}_Z^\lambda \geq (134/400)(r_2^*)^2$ and so (6.58) holds for $\eta < (134/400)(r_2^*)^2$.

Let us now turn to Equation (6.59). Let $Z \in \mathcal{B}_1$. On $\Omega_{K,G}$ there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leq (r_1^*)^2/(20)$. On these blocks B_k, all $P_{B_k}\mathcal{L}_Z^\lambda$'s are such that

$$\begin{aligned}
P_{B_k}\mathcal{L}_Z^\lambda &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \\
&\geq P\mathcal{L}_Z - \frac{1}{20}(r_1^*)^2 - \lambda\|Z - Z^*\| \qquad (6.64) \\
&\geq -\frac{1}{20}(r_1^*)^2 - \lambda\rho^* \\
&= -\frac{1}{20}(r_1^*)^2 - \frac{11}{40}(r_2^*)^2 \geq -\frac{13}{40}(r_2^*)^2 \qquad (6.65)
\end{aligned}$$

because $r_1^* \leq r_2^*$ (see Proposition 7.1 in the Appendix). Next, it follows from (6.60), the fact that $\alpha > 1$, (6.61) and (6.63) for $b = 1$ and (6.64) that

$$\begin{aligned}
\sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) &\leq \max\left(\frac{-1}{5}, \frac{-57}{400}, \frac{13}{40}\right) r_2^2 \\
&= \frac{13}{40}(r_2^*)^2 \qquad (6.66)
\end{aligned}$$

and so (6.59) holds for $\eta \geq 13(r_2^*)^2/(40)$. As a consequence, (6.58) and (6.59) both hold for $\eta = 132(r_2^*)^2/(400)$. □

From Lemmas 68 and 69, we conclude that on the event $\Omega_{K,G}$, $\hat{Z} \in \mathcal{B}_2$, that is $\|\hat{Z} - Z^*\| \leq 2\rho^*$ and $G(\hat{Z} - Z^*) \leq (r_2^*)^2$. The following lemma gives us an upper bound on the excess risk $P\mathcal{L}_{\hat{Z}}$.

Lemma 70. *Under the conditions of Theorem 24, and on the event $\Omega_{K,G}$, we have*

$$P\mathcal{L}_{\hat{Z}} \leq (93/100)(r_2^*)^2.$$

Proof. From Lemmas 68 and 69, we get that on $\Omega_{K,G}$, $\hat{Z} \in \mathcal{B}_2$. This implies the existence of stricly more than $K/2$ blocks B_k on which $\left|(P_{B_k} - P)\mathcal{L}_{\hat{Z}}\right| \leq (r_2^*)^2/(20)$, that is:

$$P\mathcal{L}_{\hat{Z}} \leq P_{B_k}\mathcal{L}_{\hat{Z}} + (r_2^*)^2/(20). \tag{6.67}$$

Moreover, by (6.59), the definition of $\hat{Z}$ and (69), we have:

$$\begin{aligned}
\mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) + \lambda\left(\|\hat{Z}\| - \|Z^*\|\right) &\leq \sup_{Z\in\mathcal{C}} \ \mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_Z) + \lambda\left(\|\hat{Z}\| - \|Z\|\right) \\
&\leq \sup_{Z\in\mathcal{C}} \ \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda\left(\|Z^*\| - \|Z\|\right) \\
&\leq \frac{33}{100}(r_2^*)^2.
\end{aligned}$$

As a consequence, there exist at least $K/2$ blocks B_k on which

$$\begin{aligned}
P_{B_k}\mathcal{L}_{\hat{Z}} &\leq \frac{33}{100}(r_2^*)^2 - \lambda\left(\|\hat{Z}\| - \|Z^*\|\right) \\
&\leq \frac{33}{100}(r_2^*)^2 + \lambda\|\hat{Z} - Z^*\| \\
&\leq \frac{33}{100}(r_2^*)^2 + 2\lambda\rho^* = \frac{88}{100}(r_2^*)^2.
\end{aligned} \tag{6.68}$$

So there must be at least a block B_{k_0} on which (6.67) and (6.68) hold simultaneously. On this block, we have

$$P\mathcal{L}_{\hat{Z}} \leq P_{B_{k_0}}\mathcal{L}_{\hat{Z}} + \frac{1}{20}(r_2^*)^2 \leq \frac{88}{100}(r_2^*)^2 + \frac{1}{20}(r_2^*)^2 = \frac{93}{100}(r_2^*)^2.$$

□

At this stage, we have shown that on the event $\Omega_{K,G}$, the estimator $\hat{Z}$ has the statistical properties announced in Theorem 24. In what follows we prove that under the conditions of Theorem 24, $\Omega_{K,G}$ holds with exponentially large probability.

Lemma 71. *Assume that $K \geq 100|\mathcal{O}|$, and let $\rho^* > 0$ be such that it satisfies the sparsity equation from Definition 23. Then, Ω_K holds with probability at least $1 - 2\exp(-72K/625)$.*

Proof. Consider

$$\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t\geq 1\}} + 2(t - 1/2)\mathbb{1}_{\left\{1/2\leq t\leq 1\right\}},$$

so that for any $t \in \mathbb{R}$,

$$\mathbb{1}_{\{t\geq 1\}} \leq \phi(t) \leq \mathbb{1}_{\left\{t\geq 1/2\right\}}.$$

For $k \in [K]$, let $W_k := \{X_i : i \in B_k\}$ and $F_Z(W_k) = (P_{B_k} - P)\mathcal{L}_Z$. We also define the counterparts of these quantities constructed with the non-corrupted vectors: $\widetilde{W}_k := \left\{\widetilde{X}_i : i \in B_k\right\}$ and $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - \widetilde{P})\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := \frac{K}{N}\sum_{i\in B_k}\mathcal{L}_Z(\widetilde{X}_i)$ and $\widetilde{P}\mathcal{L}_Z := \mathbb{E}[\mathcal{L}_Z(\widetilde{X}_i)]$. For both $b \in \{1, 2\}$, define

$$Z \to \psi_b(Z) = \sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(W_k)|\leq \frac{(r_b^*)^2}{20}\right\}}.$$

Let $b \in \{1, 2\}$. We want to show that, with high probability, if $Z \in \mathcal{B}_b$, then $\psi_b(Z) > K/2$. As we showed in the proof of Lemma 53, in our framework this is equivalent to proving that the following inequality occurs with high probability:

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|> \frac{(r_b^*)^2}{20}\right\}} \leq \frac{49K}{100}, \tag{6.69}$$

and this is what we do now. Let $Z \in \mathcal{B}_b$. We have:

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \\
&= \sum_{k\in[K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right)\right] \\
&= \sum_{k\in[K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{40}\right\}}\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leq \sum_{k\in[K]} \left(\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leq \sup_{Z\in\mathcal{B}_b} \left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right).
\end{aligned} \tag{6.70}$$

We start with bounding the last sum in the previous inequality. For each $k \in [K]$, Markov's inequality and the definition of r_b^* yield to

$$\begin{aligned}
\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) &\leq \frac{1600^2}{(r_b^*)^4}\mathbb{E}\left[F_Z(\widetilde{W}_k)^2\right] \\
&= \frac{1600^2}{(r_b^*)^4}\left(\frac{K}{N}\right)\mathrm{Var}(\mathcal{L}_Z(\widetilde{X})) \\
&\leq \frac{1600^2}{(r_b^*)^4}\left(V_K(r_b^*)\right)^2 \leq \frac{1}{200}
\end{aligned}$$

Plugging this last result into (6.70), we get:

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \leq \frac{K}{200} + \sup_{Z\in\mathcal{B}_b} \left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right). \tag{6.71}$$

We now have to bound this last term. Using Mc Diarmind inequality (Theorem 6.2 in Boucheron et al. (2013a) with $t = 12/25$), we get that with probability at least $1 - \exp(-72K/625)$, for all $Z \in \mathcal{B}_b$,

$$\begin{aligned}
&\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) \leq \\
&\qquad \frac{12K}{25} + \mathbb{E}\left[\sup_{Z\in\mathcal{B}_b} \sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right].
\end{aligned} \tag{6.72}$$

Let now $\epsilon_1, \ldots, \epsilon_K$ be Rademacher variables independant from the $\widetilde{X}_i$'s. By the symmetrization

Lemma, we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right]$$
$$\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]. \tag{6.73}$$

As ϕ is Lipschitz with $\phi(0)=0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), chapter 4) to get that:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\frac{20F_Z(\widetilde{W}_k)}{(r_b^*)^2}\right]$$
$$=\frac{40}{(r_b^*)^2}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-\widetilde{P})\mathcal{L}_Z\right] \tag{6.74}$$

Now, let $(\sigma_i)_{i=1,\dots,N}$ be a family of Rademacher variables independant from the $\widetilde{X}_i$'s and the ϵ_i's. Using the symmetrization Lemma again, we get:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-\widetilde{P})\mathcal{L}_Z\right]\leq 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\frac{K}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right]$$
$$\leq 2KE_G(r_b^*,b\rho^*)\leq 2K\gamma(r_b^*)^2.$$

Combining this with (6.72), (6.74) and (6.75), we finally get that, with probability at least $1-\exp(-72K/625)$:

$$\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\leq\left(\frac{12}{25}+160\gamma\right)K. \tag{6.75}$$

Plugging that into (6.71), we conclude that, with probability at least $1-\exp(-72K/625)$, for all $Z\in\mathcal{B}_b$,

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}}\leq\left(\frac{1}{200}+\frac{12}{25}+160\gamma\right)K\leq\frac{49}{100}K$$

for our choice of parameters. Now, in order for $\Omega_{K,G}$ to hold, this inequality must be verified for both $b=1$ and 2. Then, we finally conclude that $\Omega_{K,G}$ holds with probability $1-2\exp(-72K/625)$, which concludes the proof.

□

6.2 Proofs of Chapter 4

6.2.1 Stochastic processes

6.2.1.1 Proof of Theorem 25

The proof of Theorem 25 relies on several lemmas. We first recall that $kB_1\cap B_2\subset 2conv(U_{k^2}\cap\mathcal{S}_2^{d\times d})$ where U_{k^2} is the set of all matrices in $\mathbb{R}^{d\times d}$ with k^2 non zero entries (see, for instance,

equation (3.1) in Mendelson, Pajor, and Tomczak-Jaegermann (2007)) and so for all $A \in \mathbb{R}^{d\times d}$,

$$\|A\| \leq 2 \sup_{I\subset[d]\times[d]:|I|=k^2} \left(\sum_{(p,q)\in I} A_{pq}^2\right)^{1/2}.$$

We therefore need to find a high probability upper bound on the ℓ_2 norm of the k^2 largest entries of $\hat{\Sigma}_N - \Sigma$. To that end, we start with the following result.

Lemma 72. *Let $(z_{pq} : p, q \in [d])$ be real-valued random variables (not necessarily independent) and $\lambda, t \geq 1$ be two positive constants. We assume that for $r = 2\log(ed/k) + t$, we have $\|z_{pq}\|_{L_r} \leq \lambda\sqrt{r}$ for all $p, q \in [d]$. Then, with probability at least $1 - \exp(-t)$,*

$$\sup_{I\subset[d]\times[d]:|I|=k^2} \left(\sum_{(p,q)\in I} z_{pq}^2\right)^{1/2} \leq e^2\lambda\sqrt{2k^2\left(\log(ed/k) + t\right)}.$$

Moreover:

$$\mathbb{E}\left[\sup_{I\subset[d]\times[d]:|I|=k^2} \left(\sum_{(p,q)\in I} z_{pq}^2\right)^{1/2}\right] \leq e^2\lambda\sqrt{6k^2\log(ed/k)}$$

Proof. We define for all $p, q \in [d]$,

$$Z_{pq} = z_{pq}I(|z_{pq}| \leq e\lambda\sqrt{r}) \text{ and } Y_{pq} = z_{pq}I(|z_{pq}| > e\lambda\sqrt{r})$$

so that we have $|z_{pq}|^t = |Z_{pq}|^t + |Y_{pq}|^t$. As a consequence and by convexity of $x \in \mathbb{R}^+ \to x^{t/2}$, we have for all $I \subset [d] \times [d]$

$$\left(\frac{1}{|I|}\sum_{(p,q)\in I} z_{pq}^2\right)^{t/2} \leq \frac{1}{|I|}\sum_{(p,q)\in I} |z_{pq}|^t = \frac{1}{|I|}\sum_{(p,q)\in I} |Z_{pq}|^t + \frac{1}{|I|}\sum_{(p,q)\in I} |Y_{pq}|^t. \tag{6.76}$$

Let $I \subset [d] \times [d]$ be such that $|I| = k^2$. We have

$$\frac{1}{|I|}\sum_{(p,q)\in I} |Z_{pq}|^t \leq (e\lambda\sqrt{r})^t. \tag{6.77}$$

For the second term in the right hand side inequality of (6.76), we have

$$\frac{1}{|I|}\sum_{(p,q)\in I} |Y_{pq}|^t \leq \frac{1}{|I|}\sum_{(p,q)\in[d]\times[d]} |Y_{pq}|^t$$

and for $\theta := r/t$ and all $p, q \in [d]$, we have

$$\begin{aligned}\mathbb{E}[|Y_{pq}|^t] = \mathbb{E}\left[|z_{pq}|^t I(|z_{pq}| > e\lambda\sqrt{r})\right] &\leq \mathbb{E}\left[|z_{pq}|^{t\theta}\right]^{1/\theta} \mathbb{P}\left[|z_{pq}| > e\lambda\sqrt{r}\right]^{1-1/\theta} \\ &\leq (\lambda\sqrt{r})^t \left(\frac{\|z_{pq}\|_{L_r}}{e\lambda\sqrt{r}}\right)^{r-t} \leq (\lambda\sqrt{r})^t e^{r-t} = (\lambda\sqrt{r})^t \frac{k^2}{e^2d^2}.\end{aligned}$$

It follows that

$$\mathbb{E}\sup_{I\subset[d]\times[d]:|I|=k^2} \frac{1}{|I|}\sum_{(p,q)\in I} |Y_{pq}|^t \leq \frac{d^2}{k^2}(\lambda\sqrt{r})^t \frac{k^2}{e^2d^2} = \frac{(\lambda\sqrt{r})^t}{e^2}.$$

Hence, using (6.76), (6.77) and the last inequality, we get $\mathbb{E}\mathcal{Z}^t \leq (e\lambda\sqrt{r})^t + (\lambda\sqrt{r})^t/e^2 \leq 2(e\lambda\sqrt{r})^t$ where

$$\mathcal{Z} := \sup_{I\subset[d]\times[d]:|I|=k^2} \left(\frac{1}{|I|}\sum_{(p,q)\in I} z_{pq}^2\right)^{1/2}.$$

As a consequence, $\|\mathcal{Z}\|_{L_t} \leq e\lambda\sqrt{2r}$ and so, for $t \geq 2$ we get by Markov's inequality that $\mathcal{Z} \leq e^2\lambda\sqrt{2r}$ with probability at least $1-\exp(-t)$.
Now, by taking simply $t = 1$ we get:

$$\|\mathcal{Z}\|_{L_1} \leq e\lambda\sqrt{2r} = e\lambda\sqrt{2(2\log(ed/k)+1)} \leq e\lambda\sqrt{6\log(ed/k)}$$

since $k \leq d$. By consequence, $\mathbb{E}[\mathcal{Z}] = \|\mathcal{Z}\|_{L_1} \leq e\lambda\sqrt{6\log(ed/k)}$, which concludes the proof. □

The proof of Theorem 25 will follow from Lemma 72 if one can apply the latter to the variables $z_{pq} = \hat{\Sigma}_{pq} - \Sigma_{pq}$. We therefore have to check that $(\hat{\Sigma}_{pq} - \Sigma_{pq} : p, d \in [d])$ satisfies the assumptions of Lemma 72. In other words, it only remains to show that for all $p, q \in [d]$, $\hat{\Sigma}_{pq} - \Sigma_{pq}$ has $r := 2\log(ed/k) + t$ sub-gaussian moment under Assumption 4.1. To that end we use a version (see Lemma 2.8 in Lecué and Mendelson (2014)) of a result due to Latała taken from Latała et al. (1997) (see Theorem 2 and Remark 2 in Latała et al. (1997)) which states the following:

Lemma 73. *(Latała et al., 1997) There exists an absolute constant c_0 for which the following holds. Let z be a mean-zero random variable and $z_1, \ldots, z_N$ be be N independent copies of z. Consider $p_0 \geq 2$ and assume that there exists $\kappa_1 > 0$ and $\alpha \geq 1/2$ for which $\|z\|_{L_p} \leq \kappa_1 p^\alpha$ for every $2 \leq p \leq p_0$. If $N \geq p_0^{\max\{2\alpha-1,1\}}$ then for every $2 \leq p \leq p_0$,*

$$\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N z_i\right\|_{L_p} \leq c_1(\alpha)\kappa_1\sqrt{p},$$

where $c_1(\alpha) = c_0\exp((2\alpha-1))$.

We use Lemma 73 to prove the following moment growth condition on the $\hat{\Sigma}_{pq} - \Sigma_{pq}, p, q \in [d]$.

Lemma 74. *There exists an absolute constant c_0 such that the following holds. Grant Assumption 4.1 with parameters w and $t \geq 2$. For all $p, q \in [d]$ and all $2 \leq r \leq 2\log(ed/k) + t$, if $N \geq 2\log(ed/k) + t$ then $\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leq (c_0 w^2/\sqrt{N})\sqrt{r}$.*

Proof. Consider $p, q \in [d]$. It follows from Assumption 4.1 and Lemma 73 that for all $p, q \in [d]$ and all $2 \leq r \leq 2\log(ed/k) + t$,

$$\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leq \frac{1}{\sqrt{N}}\|\frac{1}{\sqrt{N}}\sum_{i=1}^N X_{ip}X_{iq} - \mathbb{E}X_{ip}X_{iq}\|_{L_r} \leq \frac{c_0 w^2}{\sqrt{N}}\sqrt{r}.$$

□

6.2.1.2 Proof of Theorem 26

The proof of Theorem 26 relies on several lemmas. We first use a decomposition similar to the one from Lecué and Mendelson (2018). We have

$$\begin{aligned}\|\hat{\Sigma}_N - \Sigma\|_\rho &\leq \min\left(\sup_{Z\in B_2}\sum_{p,q=1}^d Z_{pq}(\hat{\Sigma}_N-\Sigma)_{pq}, \sup_{Z\in\rho B_{SLOPE}}\sum_{p,q=1}^d Z_{pq}(\hat{\Sigma}_N-\Sigma)_{pq}\right)\\ &= \min\left(\sqrt{\sum_{p,q=1}^d(\hat{\Sigma}_N-\Sigma)_{pq}^2}, \rho\sup_{Z\in\rho B_{SLOPE}}\sum_{p,q=1}^d Z^*_{(p,q)}\beta_{pq}\frac{(\hat{\Sigma}_N-\Sigma)^*_{(p,q)}}{\beta_{pq}}\right)\\ &= \min\left(\sqrt{\sum_{p,q=1}^d(\hat{\Sigma}_N-\Sigma)_{pq}^2}, \rho\max_{p,q\in[d]}\frac{(\hat{\Sigma}_N-\Sigma)^*_{(p,q)}}{\beta_{pq}}\right). \end{aligned} \tag{6.78}$$

We already proved a high probability upper bound on the ℓ_2 norm of the k^2 largest entries of $\hat{\Sigma}_N - \Sigma$ in the previous section under a weaker assumption than the one in Assumption 4.2. We just have to use it for $k = d$ to handle the left-hand side term of (6.78). Therefore, with probability at least $1 - \exp(-t)$,

$$\sqrt{\sum_{p,q=1}^d(\hat{\Sigma}_N-\Sigma)_{pq}^2} \leq c_0 w^2\sqrt{\frac{d}{N}}.$$

It only remains to handle the second term in the right-hand side inequality of (6.78). To that end, we start with the following result.

Lemma 75. *Let $\boldsymbol{z} := (z_{pq} : p,q \in [d])$ be real-valued random variables (not necessarily independent) and $\lambda, t \geq 1$ be two positive constants. We denote by $(z^*_{(p,q)} : p,q \in [d])$ the non-increasing sequence (for the same lexicographical order over $[d]^2$ used before) of the rearrangement of the absolute values of the entries of $\boldsymbol{z}$. Consider $p_0, q_0 \in [d]$. We assume that for $r = \log[ed^2/(p_0q_0)] + t$, we have $\|z_{pq}\|_{L_r} \leq \lambda\sqrt{r}$ for all $p,q \in [d]$. Then,*

$$\left\|\frac{z^*_{(p_0,q_0)}}{\beta_{p_0q_0}}\right\|_{L_t} \leq e^2\lambda.$$

Proof. To make the presentation of the proof simpler, we index the entries of $d \times d$ matrices by $[d^2]$. We therefore have d^2 random variables $(z_j)_j$ (not necessarily independent) and $\beta_j = \sqrt{\log(ed^2/j) + t}$ for all $j \in [d^2]$. Consider $j_0 \in [d^2]$ and set $r_0 = \log(ed^2/j_0) + t$. We assume that $\|z_j\|_{L_{r_0}} \leq \lambda\sqrt{r_0}$ for all $p,q \in [d]$. We want to prove that $\|z^*_{j_0}/\beta_{j_0}\|_{L_t} \leq e^2\lambda$. We first remark that

$$\frac{z^*_{j_0}}{\beta_{j_0}} \leq \max_{I\subset[d^2]:|I|=j_0}\frac{1}{\beta_{j_0}|I|}\sum_{j\in I}|z_j| := \mathcal{Z}. \tag{6.79}$$

We define for all $j \in [d^2]$,

$$Z_j = z_j I\left(|z_j| \leq e\lambda\sqrt{r_0}\right) \text{ and } Y_j = z_j I\left(|z_j| > e\lambda\sqrt{r_0}\right).$$

It follows from the convexity of $x \in \mathbb{R}_+ \to x^t$ and the definitions above that

$$\mathbb{E}\mathcal{Z}^t \leq \max_{I\subset[d^2]:|I|=j_0}\frac{1}{\beta_{j_0}^t|I|}\sum_{j\in I}|z_j|^t \leq \left(\frac{e\lambda\sqrt{r_0}}{\beta_{j_0}}\right)^t + \frac{1}{j_0}\sum_{j=1}^{d^2}\frac{\mathbb{E}|Y_j|^t}{\beta_{j_0}^t}. \tag{6.80}$$

Next, for the second term in the right-hand side inequality of (6.80) for $\theta := r_0/t$ and all $j \in [d^2]$, we have

$$\mathbb{E}|Y_j|^t = \mathbb{E}\left[|z_j|^t I(|z_j| > e\lambda\sqrt{r_0})\right] \leq \mathbb{E}\left[|z_j|^{t\theta}\right]^{1/\theta} \mathbb{P}\left[|z_j| > e\lambda\sqrt{r_0}\right]^{1-1/\theta}$$
$$\leq \left(\lambda\sqrt{r_0}\right)^t \left(\frac{\|z_j\|_{L_{r_0}}}{e\lambda\sqrt{r_0}}\right)^{r_0-t} \leq (e\lambda)^t r_0^{t/2} e^{-r_0} = (e\lambda)^t \beta_{j_0}^t e^{-t}\frac{j_0}{ed^2}.$$

We end up in (6.80) with $\mathbb{E}\mathcal{Z}^t \leq (e\lambda)^t + \lambda^t \leq (e^2\lambda)^t$. □

Lemma 76. *Let $\boldsymbol{z} := (z_{pq} : p, q \in [d])$ be real-valued random variables (not necessarily independent) and $\lambda \geq 0, t \geq 3$ be two constants. We denote by $(z^*_{(p,q)} : p, q \in [d])$ the non-increasing sequence (for the same lexicographical order over $[d]^2$ used before) of the rearrangement of the absolute values of the entries of $\boldsymbol{z}$. Consider $r_0 = \log(ed^2) + t$ and assume that $\|z_{pq}\|_{L_r} \leq \lambda\sqrt{r}$ for all $p, q \in [d]$ and $2 \leq r \leq r_0$. Consider $k \in [d]$ and $\gamma \geq 1$. Then, when $t \geq \max\left(2\log(\lceil\log(k^2)\rceil), \gamma\log(ed^2/k^2)\right)$, with probability at least $1 - \exp(-t/2)$,*

$$\max_{p,q\in[d^2]}\left(\frac{z^*_{(p,q)}}{\beta_{pq}}\right) \leq \sqrt{2}e^3\lambda.$$

Proof. We use the same 'vectorial' notation as the one introduced in the proof of Lemma 75. We remark that for all $j \in [d^2]$, we have $(1/\beta_{2j}) \leq \sqrt{2}/\beta_j$ when $t \geq 3$ and for all $j \geq k^2$, $1/\beta_j \leq \sqrt{2}/\beta_{k^2}$ when $t \geq \gamma\log(ed^2/k^2)$, hence,

$$\max_{j\in[d^2]}\left(\frac{z^*_j}{\beta_j}\right) \leq \sqrt{2}\max\left(\frac{z^*_{2^j}}{\beta_{2^j}} : j = 0, 1, \ldots, \lceil\log(k^2)\rceil\right).$$

Il follows from lemma 75 that for all $j = 0, 1, \ldots, \lceil\log(k^2)\rceil$, we have $\|z^*_{2^j}/\beta_{2^j}\|_{L_t} \leq e^2\lambda$ and so by Markov's inequality with probability at least $1-\exp(-t)$, $z^*_{2^j}/\beta_{2^j} \leq e^3\lambda$. The union bound yields that with probability at least $1 - \lceil\log(k^2)\rceil\exp(-t)$, $\max\left(z^*_{2^j}/\beta_{2^j} : j = 0, 1, \ldots, \lceil\log(k^2)\rceil\right) \leq e^3\lambda$.

□

The proof of Theorem 25 will follow from Lemma 76 if one can apply the latter to the variables $z_{pq} = \hat{\Sigma}_{pq} - \Sigma_{pq}$. We therefore have to check that the family of random variables $(\hat{\Sigma}_{pq} - \Sigma_{pq} : p, d \in [d])$ satisfies the assumptions of Lemma 76. In other words, it only remains to show that for all $p, q \in [d]$, $\hat{\Sigma}_{pq} - \Sigma_{pq}$ has $r := \log(ed^2) + t$ sub-gaussian moment under Assumption 4.2. To that end we use a version (see Lemma 2.8 in (Lecué & Mendelson, 2014)) of a result due to Latała taken from Latała et al. (1997) (see Theorem 2 and Remark 2 in Latała et al. (1997)) which states the following:

Lemma 77. *(Latała et al., 1997) There exists an absolute constant c_0 for which the following holds. Let z be a mean-zero random variable and $z_1, \ldots, z_N$ be be N independent copies of z. Consider $p_0 \geq 2$ and assume that there exists $\kappa_1 > 0$ and $\alpha \geq 1/2$ for which $\|z\|_{L_p} \leq \kappa_1 p^\alpha$ for every $2 \leq p \leq p_0$. If $N \geq p_0^{\max\{2\alpha-1,1\}}$ then for every $2 \leq p \leq p_0$,*

$$\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N z_i\right\|_{L_p} \leq c_1(\alpha)\kappa_1\sqrt{p},$$

where $c_1(\alpha) = c_0\exp((2\alpha - 1))$.

We use Lemma 77 to prove the following moment growth condition on the $\hat{\Sigma}_{pq} - \Sigma_{pq}, p, q \in [d]$.

Lemma 78. *There exists an absolute constant c_0 such that the following holds. Grant Assumption 4.2 with parameters w and $t \geq 3$. For all $p, q \in [d]$ and all $2 \leq r \leq 2\log(ed^2) + t$, if $N \geq 2\log(ed^2) + t$ then $\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leq (c_0 w^2/\sqrt{N})\sqrt{r}$.*

Proof. Consider $p, q \in [d]$. It follows from Assumption 4.2 and Lemma 77 that for all $p, q \in [d]$ and all $2 \leq r \leq 2\log(ed^2) + t$,

$$\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leq \frac{1}{\sqrt{N}} \| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_{ip}X_{iq} - \mathbb{E}X_{ip}X_{iq} \|_{L_r} \leq \frac{c_0 w^2}{\sqrt{N}} \sqrt{r}.$$

□

Proof of Theorem 26 We set for all $p, q \in [d]$, $z_{pq} = \hat{\Sigma}_{pq} - \Sigma_{pq}$. It follows from Lemma 78 for $\alpha = 1$ that for all $2 \leq r \leq 2\log(ed^2) + t$, $\|z_{pq}\|_{L_r} \leq \lambda\sqrt{r}$ where $\lambda = c_0 w^2/\sqrt{N}$. The result now follows from Lemma 76. ■

6.2.2 Signed clustering

6.2.2.1 Proof of Equation (4.7)

We recall that the cluster matrix $\bar{Z} \in \{0,1\}^{d\times d}$ is defined by $Z_{ij} = 1$ if $i \sim j$ and $Z_{ij} = 0$ when $i \not\sim j$ and $\alpha = \delta(p + q - 1)$. For all matrix $Z \in [0,1]^{d\times d}$, we have

$$\begin{aligned}
\big\langle Z, \mathbb{E}A - \alpha J \big\rangle &= \sum_{i,j=1}^{d} Z_{ij}(\mathbb{E}A_{ij} - \alpha) = \sum_{(i,j)\in\mathcal{C}^+} Z_{ij}(\mathbb{E}A_{ij} - \alpha) + \sum_{(i,j)\in\mathcal{C}^-} Z_{ij}(\mathbb{E}A_{ij} - \alpha) \\
&= [\delta(2p-1) - \alpha] \sum_{(i,j)\in\mathcal{C}^+ : i\neq j} Z_{ij} + [\delta(2q-1) - \alpha] \sum_{(i,j)\in\mathcal{C}^-} Z_{ij} + (1-\alpha)\sum_{i=1}^{d} Z_{ii} \\
&= \delta(p-q)\left[\sum_{(i,j)\in\mathcal{C}^+} Z_{ij} - \sum_{(i,j)\in\mathcal{C}^-} Z_{ij} \right] + (1-\alpha)\sum_{i=1}^{d} Z_{ii}.
\end{aligned}$$

The latter quantity is maximal for $Z \in [0,1]^{d\times d}$ such that $Z_{ij} = 1$ for $(i,j) \in \mathcal{C}^+$ and $Z_{ij} = 0$ for $(i,j) \in \mathcal{C}^-$, that is when $Z = \bar{Z}$. As a consequence $\{\bar{Z}\} = \operatorname{argmax}_{Z\in[0,1]^{d\times d}} \big\langle Z, \mathbb{E}A - \alpha J \big\rangle$. Moreover, $\bar{Z} \in \mathcal{C} \subset [0,1]^{d\times d}$ so we also have that $\bar{Z}$ is the only solution to the problem $\max_{Z\in\mathcal{C}} \big\langle Z, \mathbb{E}A - \alpha J \big\rangle$ and so $\bar{Z} = Z^*$. ■

6.2.2.2 Proof of Theorem 30

For a problem, such as the signed clustering, where it is possible to characterize the curvature of the excess risk, we start to identify this curvature because the curvature function G, coming out of it, defines the local subsets of $\mathcal{C}$ driving the complexity of the problem. Then, we turn to the stochastic part of the proof, which is entirely summarized into the complexity fixed point $r^*_G(\Delta)$ from (4). Finally, we put the two pieces together and apply the main general result from Theorem 6 to obtain estimation results for the SDP estimator (4.8) in the signed clustering problem.

Curvature equation

In this section, we show that the objective function $Z \in \mathcal{C} \to \big\langle Z, \mathbb{E}A - \alpha J \big\rangle$ satisfies a curvature equation around its maximizer Z^* with respect to the $\ell_1^{d\times d}$-norm given by $G(Z^* - Z) = \theta\|Z^* - Z\|_1$ with parameter $\theta = \delta(p-q)$ (and margin exponent $\kappa = 1$).

Proposition 6.2. For $\theta = \delta(p-q)$, we have for all $Z \in \mathcal{C}$, $\langle \mathbb{E}A - \alpha J, Z^* - Z \rangle = \theta \|Z^* - Z\|_1$.

Proof. Let Z be in $\mathcal{C}$. We have

$$\begin{aligned}\langle Z^* - Z, \mathbb{E}A - \alpha J \rangle &= \sum_{i,j=1}^{d} (Z^* - Z)_{ij}(\mathbb{E}A_{ij} - \alpha) \\ &= \sum_{(i,j)\in\mathcal{C}^+} (Z^*_{ij} - Z_{ij})(\delta(2p-1) - \alpha) + \sum_{(i,j)\in\mathcal{C}^-} (Z^*_{ij} - Z_{ij})(\delta(2q-1) - \alpha) \\ &= \delta(p-q)\left[\sum_{(i,j)\in\mathcal{C}^+} (Z^* - Z)_{ij} - \sum_{(i,j)\in\mathcal{C}^-} (Z^* - Z)_{ij}\right].\end{aligned}$$

Moreover, for all $(i,j) \in \mathcal{C}^+, Z^*_{ij} = 1$ and $0 \le Z_{ij} \le 1$, so $(Z^* - Z)_{ij} = |(Z^* - Z)_{ij}|$. We also have for all $(i,j) \in \mathcal{C}^-$, $(Z^* - Z)_{ij} = -Z_{ij} = -|(Z^* - Z)_{ij}|$ because in that case $Z^*_{ij} = 1$ and $0 \le Z_{ij} \le 1$. Hence,

$$\langle Z^* - Z, \mathbb{E}A - \alpha J \rangle = \delta(p-q)\left[\sum_{(i,j)\in\mathcal{C}^+} |(Z^* - Z)_{ij}| + \sum_{(i,j)\in\mathcal{C}^-} |(Z^* - Z)_{ij}|\right] = \theta\|Z^* - Z\|_1.$$

□

Computation of the complexity fixed point $r^*_G(\Delta)$

Define $W := A - \mathbb{E}A$ the noise matrix of the problem. Since W is symmetric, its entries are not independent. In order to work only with independent random variables, we define the following matrix $\Psi \in \mathbb{R}^{d\times d}$:

$$\Psi_{ij} = \begin{cases} W_{ij} \text{ if } i \le j \\ 0 \text{ otherwise,} \end{cases} \tag{6.81}$$

where 0 entries are considered as independent Bernoulli variables with parameter 0 and therefore, Ψ has independent entries and satisfies the relation $W = \Psi + \Psi^\top$.

In order to obtain upper bounds on the fixed point complexity parameter $r^*_G(\Delta)$ associated with the signed clustering problem, we need to prove a high-probability upper bound on the quantity

$$\sup_{Z\in\mathcal{C}:\|Z-Z^*\|_1\le r} \langle W, Z - Z^* \rangle \tag{6.82}$$

and then find a radius r as small as possible such that the quantity in (6.82) is less than $(\theta/2)r$. We denote $\mathcal{C}_r := \mathcal{C} \cap (Z^* + rB_1^{d\times d}) = \{Z \in \mathcal{C} : \|Z - Z^*\|_1 \le r\}$ where $B_1^{d\times d}$ is the unit $\ell_1^{d\times d}$-ball of $\mathbb{R}^{d\times d}$.

We follow the strategy from Fei and Chen (2019b) by decomposing the inner product $\langle W, Z - Z^* \rangle$ into two parts according to the SVD of Z^*. This observation is a key point in the work of Fei and Chen (2019b) compared to the analysis from Guédon and Vershynin (2016). This allows to perform the localization argument efficiently. Up to a change of index of the nodes, Z^* is a block matrix with K diagonal blocks of 1's. It therefore admits K singular vectors $U_{\bullet k} := I(i \in \mathcal{C}_k)/\sqrt{|\mathcal{C}_k|}$ with multiplicity l_k associated with the singular value l_k for all $k \in [K]$. We can therefore write

$$Z^* = \sum_{k=1}^{K} l_k U_{\bullet k} \otimes U_{\bullet k} = UDU^\top,$$

where $U \in \mathbb{R}^{d\times K}$ has K column vectors given by $U_{\bullet k}, k = 1, \dots, K$ and $D = diag(l_1, \dots, l_K)$.

We define the following projection operator

$$\mathcal{P} : M \in \mathbb{R}^{d\times d} \to UU^T M + MUU^T - UU^T MUU^T,$$

and its orthogonal projection $\mathcal{P}^\perp$ by

$$\mathcal{P}^\perp : M \in \mathbb{R}^{d\times d} \to M - \mathcal{P}(M) = (\mathrm{I}_d - UU^T)M(\mathrm{I}_d - UU^T) = \sum_{k=K+1}^{d} \langle M, U_{\bullet k} \otimes U_{\bullet k}\rangle U_{\bullet k} \otimes U_{\bullet k}$$

where $U_{\bullet k} \in \mathbb{R}^d, k = K+1, \dots, d$ are such that $(U_{\bullet k} : k = 1, \dots, d)$ is an orthonormal basis of $\mathbb{R}^d$.

We use the same decomposition as in (Fei & Chen, 2019b): for all $Z \in \mathcal{C}$,

$$\langle W, Z - Z^*\rangle = \langle W, \mathcal{P}(Z - Z^*) + \mathcal{P}^\perp(Z - Z^*)\rangle = \underbrace{\langle \mathcal{P}(Z - Z^*), W\rangle}_{S_1(Z)} + \underbrace{\langle \mathcal{P}^\perp(Z - Z^*), W\rangle}_{S_2(Z)}.$$

The next step is to control with large probability the two terms $S_1(Z)$ and $S_2(Z)$ uniformly for all $Z \in \mathcal{C} \cap (Z^* + rB_1^{d\times d})$. To that end, we use the two following propositions where we recall that $\rho = \delta \max(1 - \delta(2p-1)^2, 1 - \delta(2q-1)^2)$ and $\nu = \max(2p - 1, 1 - 2q)$. For the sake of clarity, we postpone the proof of Propositions 6.3 and 6.4, based on the work of Fei and Chen (2019b), to Appendix 7.4.

Proposition 6.3. There are absolute positive constants c_0, c_1, c_2 and c_3 such that the following holds. If $\lceil c_1 rK/d\rceil \geq 2eKd\exp(-(9/32)d\rho/K)$ then we have

$$\mathbb{P}\left[\sup_{Z\in\mathcal{C}\cap(Z^*+rB_1^{d\times d})} S_1(Z) \leq c_2 r\sqrt{\frac{K\rho}{d}\log\left(\frac{2eKd}{\lceil \frac{c_1 rK}{d}\rceil}\right)}\right] \geq 1 - 3\left(\frac{\lceil \frac{c_1 rK}{d}\rceil}{2eKd}\right)^{\lceil \frac{c_1 rK}{d}\rceil}.$$

Proposition 6.4. There exists an absolute constant $c_0 > 0$ such that the following holds. When $d\nu\delta \geq \log d$, with probability at least $1 - \exp(-\delta\nu d)$,

$$\sup_{Z\in\mathcal{C}\cap(Z^*+rB_1^{d\times d})} S_2(Z) \leq c_0 Kr\sqrt{\frac{\delta\nu}{d}}.$$

It follows from Proposition 6.3 and Proposition 6.4 that when $d\nu\delta \geq \log d$, for all r such that $\lceil c_1 rK/d\rceil \geq 2eKd\exp(-(9/32)d\rho/K)$ we have, for

$$\Delta = \Delta(r) := \exp(-\delta\nu d) - 3\left(\lceil\frac{c_1 rK}{d}\rceil/(2eKd)\right)^{\lceil \frac{c_1 rK}{d}\rceil},$$

with probability at least $1 - \Delta$,

$$\sup_{Z\in\mathcal{C}\cap(Z^*+rB_1^{d\times d})} \langle W, Z - Z^*\rangle \leq c_0 Kr\sqrt{\frac{\delta\nu}{d}} + c_2 r\sqrt{\frac{K\rho}{d}\log\left(\frac{2eKd}{\lceil \frac{c_1 rK}{d}\rceil}\right)}.$$

Moreover, we have

$$c_0 Kr\sqrt{\frac{\delta\nu}{d}} + c_2 r\sqrt{\frac{K\rho}{d}\log\left(\frac{2eKd}{\lceil \frac{c_1 rK}{d}\rceil}\right)} \leq \frac{\theta}{2}r \tag{6.83}$$

for $\theta = \delta(p - q)$ when $K\sqrt{\nu} \lesssim \sqrt{d\delta}(p - q)$ and $\lceil c_1 rK/d\rceil \geq 2eKd\exp(-\theta^2 d/(K\rho))$. In particular, when $(p - q)^2 d\delta \geq K^2\nu$ and $1 \geq 2eKd\max\left(\exp(-\theta^2 d/(K\rho)), \exp(-(9/32)d\rho/K)\right)$, we

conclude that for all $0 < r \leq d/(c_1 K)$ (6.83) is true. Therefore, one can take $r^*_G(\Delta(0)) = 0$, meaning that we have exact reconstruction of Z^*:
if $(p-q)^2 d\delta \geq K^2\nu$ and $d \gtrsim K \max(\rho/\theta^2 \log(2eK^2\rho/\theta^2), (1/\rho)\log(2eK^2/\rho))$, then, with probability at least $1 - \exp(-\delta\nu d) - 3/(2eKd)$, one has $\hat{Z} = Z^*$.

If $(p-q)^2 d\delta \geq K^2\nu$ and $1 < 2eKd \max\left(\exp(-\theta^2 d/(K\rho)), \exp(-(9/32)d\rho/K)\right)$ then we do not have exact reconstruction anymore, but we see that (6.83) is true for any r such that $2eKd\ \exp(-\theta^2 d/(16c_2^2 K\rho)) \leq \lceil \frac{c_1 rK}{d} \rceil \leq 2eKd\ \exp(-c_0^2 K\delta\nu/(c_2^2\rho))$, which is possible since $(p-q)^2 d\delta \geq K^2\nu$, and then we can conclude that $r^*(\Delta) \leq \frac{2ed^2}{c_1\theta}\exp\left(-\frac{\theta^2 d}{16c_2^2 K\rho}\right)$.

Therefore, it follows from Theorem 6 that

$$\|Z^* - \hat{Z}\|_1 \leq \frac{2ed^2}{c_1\theta}\exp\left(-\frac{\theta^2 d}{16c_2^2 K\rho}\right).$$

6.2.3 Angular synchronization

6.2.3.1 Proof of Equation 4.14

We recall that the offsets are $\delta_{ij} = (\theta_i - \theta_j)[2\pi]$ and we will use the fact that, if g is $\mathcal{N}(0,1)$, then $\mathbb{E}e^{\iota\sigma g} = e^{-\sigma^2/2}$. For all $\gamma_1, \ldots, \gamma_d \in [0, 2\pi)$, we have $\gamma_i - \gamma_j = \delta_{ij}$ for all $i \neq j \in [d]$ if and only if $e^{\sigma^2/2}\mathbb{E}A_{ij}e^{\iota\gamma_j} - e^{\iota\gamma_i} = 0$ for all $i \neq j \in [d]$. We therefore have

$$\operatorname*{argmin}_{x\in\mathbb{C}^d:|x_i|=1}\left\{\sum_{i\neq j}|e^{\sigma^2/2}\mathbb{E}A_{ij}x_j - x_i|^2\right\} = \{(e^{\iota(\theta_i+\theta_0)})_{i=1}^d : \theta_0 \in [0, 2\pi)\}. \tag{6.84}$$

Moreover, for all $x = (x_i)_{i=1}^d \in \mathbb{C}^d$ such that $|x_i| = 1$ for $i = 1, \ldots, d$, we have

$$\sum_{i\neq j}|e^{\sigma^2/2}\mathbb{E}A_{ij}x_j - x_i|^2 = \sum_{i\neq j}|x_i^*\bar{x}_j^* x_j - x_i|^2 = \sum_{i,j=1}^d |x_i^*\bar{x}_j^* - x_i\bar{x}_j|^2$$
$$= 2d^2 - 2\Re\left(\sum_{i,j=1}^d x_i^*\bar{x}_j^* x_j\bar{x}_i\right) = 2d^2 - 2|\langle x^*, x\rangle|^2$$

where $\Re(z)$ denotes the real part of $z \in \mathbb{C}$. On the other side, we have

$$\bar{x}^\top(e^{\sigma^2/2}\mathbb{E}A)x = \sum_{i\neq j}\bar{x}_i x_i^*\bar{x}_j^* x_j + \sum_{i=1}^d \bar{x}_i e^{\sigma^2/2}x_i = d(e^{\sigma^2/2} - 1) + |\langle x^*, x\rangle|^2.$$

Hence, minimizing $x \to \sum_{i\neq j}^d |e^{\sigma^2/2}\mathbb{E}A_{ij}x_j - x_i|^2$ over all $x = (x_i)_i \in \mathbb{C}^d$ such that $|x_i| = 1$ is equivalent to maximize $x \to \bar{x}^\top\mathbb{E}Ax$ over all $x = (x_i)_i \in \mathbb{C}^d$ such that $|x_i| = 1$. This concludes the proof with (6.84).

6.2.3.2 Proof of Equation 4.15

Define $\mathcal{C}' = \{Z \in \mathbb{C}^{d\times d} : |Z_{ij}| \leq 1, \forall i, j \in [d]\}$. We first prove that $\mathcal{C} \subset \mathcal{C}'$. Consider $Z \in \mathcal{C}$. Since $Z \succeq 0$, there exists $X \in \mathbb{C}^{d\times d}$ such that $Z = X\bar{X}^\top$. For all $i \in \{1, \ldots, d\}$, denote by $X_{i\bullet}$ the i-th row vector of X. We have $\|X_{i\bullet}\|_2^2 = \langle X_{i\bullet}, X_{i\bullet}\rangle = Z_{ii} = 1$, since $\operatorname{diag}(Z) = \mathbf{1}_d$. Moreover, for all $i, j \in [d]$, we have $|Z_{ij}| = |\langle X_{i\bullet}, X_{j\bullet}\rangle| \leq \|X_{i\bullet}\|_2\|X_{j\bullet}\|_2 \leq 1$. This proves that $Z \in \mathcal{C}'$ and so $\mathcal{C} \subset \mathcal{C}'$.

Consider $Z' \in \operatorname{argmax}\left\{\Re(\langle \mathbb{E}A, Z\rangle) : Z \in \mathcal{C}'\right\}$. Since $\mathcal{C}'$ is convex and the objective function $Z \to \Re(\langle \mathbb{E}A, Z\rangle)$ is linear (for real coefficients), Z' is one of the extreme points of $\mathcal{C}'$. Extreme

points of $\mathcal{C}'$ are matrices $Z \in \mathbb{C}^{d\times d}$ such that $|Z_{ij}| = 1$ for all $i, j \in [d]$. We can then write each entry of Z' as $Z'_{ij} = e^{\iota\beta_{ij}}$ for some $0 \le \beta_{ij} < 2\pi$ and now we obtain

$$\begin{aligned}\Re(\langle \mathbb{E}A, Z'\rangle) &= \Re\left(\sum_{i,j=1}^{d} \mathbb{E}A_{ij}\overline{Z'_{ij}}\right) = \Re\left(\sum_{i\neq j}^{d} e^{-\sigma^2/2}e^{\iota\delta_{ij}}e^{-\iota\beta_{ij}}\right) + \Re\left(\sum_{i=1}^{d} e^{\iota\delta_{ii}}e^{-\iota\beta_{ii}}\right)\\ &= \sum_{i\neq j} e^{\sigma^2/2}\cos(\delta_{ij}-\beta_{ij}) + \sum_{i}\cos(\delta_{ii}-\beta_{ii}) \le e^{\sigma^2/2}(d^2-d)+d.\end{aligned}$$

The maximal value $e^{\sigma^2/2}(d^2-d)+d$ is attained only for $\beta_{ij} = \delta_{ij}$ for all $i, j \in [d]$, that is for $Z' = (e^{\iota\delta_{ij}})_{i,j=1,\dots,d} = Z^*$. But we have $Z^* \in \mathcal{C}$ and $\mathcal{C} \subset \mathcal{C}'$, so Z^* is the only maximizer of $Z \to \Re(\langle \mathbb{E}A, Z\rangle)$ on $\mathcal{C}$. But for all $Z \in \mathcal{C}$ we have $\langle \mathbb{E}A, Z\rangle = \overline{x^*}^\top Z x^* \in \mathbb{R}$, then Z^* is the only maximizer of $Z \to \langle \mathbb{E}A, Z\rangle$ over $\mathcal{C}$.

6.2.3.3 Proof of Theorem 31

Curvature of the objective function

Proposition 6.5. Consider $\theta = e^{-\sigma^2/2}/2$. Then, for any $Z \in \mathcal{C}$, on has: $\langle \mathbb{E}A, Z^* - Z\rangle \ge \theta\|Z^* - Z\|_2^2$.

Proof. Consider $Z = (z_{ij}e^{\iota\beta_{ij}})_{i,j=1}^d) \in \mathcal{C}$ where $z_{ij} \in \mathbb{R}$ and $0 \le \beta_{ij} < 2\pi$ for all $i, j \in [d]$. Since $Z^*_{ii} = Z_{ii} = 1$ for all $i \in [d]$, we have, on one side, $\langle \mathbb{E}A, Z^* - Z\rangle = e^{-\sigma^2/2}\overline{x^*}^\top(Z^* - Z)x^* \in \mathbb{R}$, and so

$$\begin{aligned}\langle \mathbb{E}A, Z^* - Z\rangle &= \Re\Big(\langle \mathbb{E}A, Z^* - Z\rangle\Big) = \Re\left(\sum_{i,j=1}^{d} \mathbb{E}A_{ij}\overline{(Z^* - Z)_{ij}}\right)\\ &= \Re\left(\sum_{i,j=1}^{d} e^{-\sigma^2/2}e^{\iota\delta_{ij}}(e^{-\iota\delta_{ij}} - z_{ij}e^{-\iota\beta_{ij}})\right)\\ &= e^{-\sigma^2/2}\Re\left(\sum_{i,j=1}^{d} 1 - z_{ij}e^{\iota(\delta_{ij}-\beta_{ij})}\right)\\ &= e^{-\sigma^2/2}\sum_{i,j=1}^{d}(1 - z_{ij}\cos(\delta_{ij}-\beta_{ij})).\end{aligned} \tag{6.85}$$

On the other side, we showed in the proof of (4.15) that $\mathcal{C} \subset \{Z \in \mathbb{C}^{d\times d} : |Z_{ij}| \le 1, \forall i, j \in [d]\}$. So we have $|z_{ij}| \le 1$ for all $i, j \in [d]$ and

$$\begin{aligned}\|Z^* - Z\|_2^2 &= \sum_{i,j=1}^{d}|(Z^* - Z)_{ij}|^2 = \sum_{i,j=1}^{d}|e^{\iota\delta_{ij}} - z_{ij}e^{\iota\beta_{ij}}|^2 = \sum_{i,j=1}^{d}|1 - z_{ij}e^{\iota(-\delta_{ij}+\beta_{ij})}|^2\\ &= \sum_{i,j=1}^{d}(1 - z_{ij}\cos(\beta_{ij}-\delta_{ij}))^2 + z_{ij}^2\sin^2(\beta_{ij}-\delta_{ij})\\ &= \sum_{i,j=1}^{d} 1 - 2z_{ij}\cos(\beta_{ij}-\delta_{ij}) + z_{ij}^2 \qquad (6.86)\\ &\le 2\sum_{i,j=1}^{d}(1 - z_{ij}\cos(\beta_{ij}-\delta_{ij})). \qquad (6.87)\end{aligned}$$

We conclude with (6.85) and (6.86). □

In fact, it follows from the proof of Proposition 6.5 that we have the following equality: for all $Z \in \mathcal{C}$,

$$\langle \mathbb{E}A, Z^* - Z \rangle = \theta \left(\|Z^* - Z\|_2^2 + \| |Z^*|^2 - |Z|^2 \|_1 \right),$$

where $|Z|^2 = (|Z_{ij}|^2)_{1 \le i,j \le d}$ (in particular, $|Z^*|^2 = (1)_{d \times d}$). We therefore know exactly how to characterize the curvature of the excess risk for the angular synchronization problem in terms of the ℓ_2 (to the square) and the ℓ_1 norms. Nevertheless, we will not use the extra term $\| |Z^*|^2 - |Z|^2 \|_1$ in the following.

Computation of the complexity fixed point $r_G^*(\delta)$

It follows from the (global) curvature property of the excess risk for the angular synchronization problem obtained in Proposition 6.5 that for the curvature G function defined by $G(Z^* - Z) = \theta \|Z^* - Z\|_2^2, \forall Z \in \mathcal{C}$, we just have to compute the $r_G^*(\delta)$ fixed point and then apply Theorem 5 in order to obtain statistical properties of $\hat{Z}$ (with respect to both the excess risk and the G function). In this section, we compute the complexity fixed point $r_G^*(\delta)$ for $0 < \delta < 1$.

Following Proposition 6.5, the natural 'local' subsets of $\mathcal{C}$ around Z^* which drive the statistical complexity of the synchronization problem are defined for all $r > 0$ by $\mathcal{C}_r = \{Z \in \mathcal{C} : \|Z - Z^*\|_2 \le r\} = \mathcal{C} \cap (Z^* + rB_2^{d \times d})$.

Consider $Z \in \mathcal{C}_r$. Denote by b_{ij}^R (resp. b_{ij}^I) the real (resp. imaginary) part of $b_{ij} = Z_{ij}^* \overline{Z_{ij} - Z_{ij}^*}$ for all $i, j \in [d]$. Since $\|Z - Z^*\|_2 \le r$ we also have $\sum_{i,j} (b_{ij}^R)^2 + (b_{ij}^I)^2 \le r^2$ and so

$$\begin{aligned}
\langle A - \mathbb{E}A, Z - Z^* \rangle = \langle (S - \mathbb{E}S) \circ Z^*, Z - Z^* \rangle &= 2\Re \left(\sum_{i<j} (S_i j - \mathbb{E}S_{ij}) b_{ij} \right) \\
&= 2 \sum_{i<j} (\cos(\sigma g_{ij}) - \mathbb{E}\cos(\sigma g_{ij})) b_{ij}^R - \sin(\sigma g_{ij}) b_{ij}^I \\
&\le 2r \sqrt{\sum_{i<j} (\cos(\sigma g_{ij}) - \mathbb{E}\cos(\sigma g_{ij}))^2 + (\sin(\sigma g_{ij}))^2} \\
&\le 2r \sqrt{1 - e^{-\sigma^2} + 2e^{-\sigma^2/2} \left(\sum_{i<j} \mathbb{E}\cos(\sigma g_{ij}) - \cos(\sigma g_{ij}) \right)}
\end{aligned}$$

where we used that $\mathbb{E}\cos(\sigma g) = \Re(\mathbb{E}e^{\iota g}) = e^{-\sigma^2/2}$ for $g \sim \mathcal{N}(0,1)$. Now it remains to get a high-probability upper bound on the sum of the centered cosinus of σg_{ij}. To get such a bound, we use Bernstein's inequality: if $Y_1, \dots, Y_N$ are N independent centered random variables such that $|Y_i| \le M$ a.s. for all $i = 1, \dots, N$ then for all $t > 0$, with probability at least $1 - \exp(-t)$,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N Y_i \le \sigma \sqrt{2t} + \frac{2Mt}{3\sqrt{N}}, \tag{6.88}$$

where $\sigma^2 = (1/N) \sum_{i=1}^N var(Y_i)$. In our context, this gives tha for all $t > 0$, with probability at least $1 - \exp(-t)$,

$$\frac{1}{\sqrt{N}} \sum_{i<j} \mathbb{E}\cos(\sigma g_{ij}) - \cos(\sigma g_{ij}) \le \sqrt{2Vt} + \frac{2t}{3\sqrt{N}} \le (1 - e^{-\sigma^2})\sqrt{t} + \frac{2t}{3\sqrt{N}},$$

for $N = d(d-1)/2$ and $V = \mathbb{E}\cos^2(\sigma g) - (\mathbb{E}\cos(\sigma g))^2 = (1/2)(1 - e^{-\sigma^2})^2$ (because $\mathbb{E}\cos^2(\sigma g) = (1/2)\mathbb{E}(1 + \cos(2\sigma g)) = (1/2)(1 + e^{-2\sigma^2})$ when $g \sim \mathcal{N}(0,1)$).

We now have all the ingredients to compute the fixed point $r_G^*(\delta)$ for $0 < \delta < 1$: for $\theta = e^{-\sigma^2/2}/2$ and $t = \log(1/\delta)$,

$$r_G^*(\delta) \leq \frac{4}{\theta}\left(1 - e^{-\sigma^2} + 2e^{-\sigma^2/2}\left((1-e^{-\sigma^2})\sqrt{tN} + \frac{2t}{3}\right)\right) = \frac{32t}{3} + 8(1-e^{-\sigma^2})(e^{\sigma^2/2} + 2\sqrt{tN}).$$

In particular, using $1 - e^{-\sigma^2} \leq \sigma^2$ and for $t = \epsilon\sigma^4 N$ (where $N = d(d-1)/2$) for some $0 < \epsilon < 1$, if $e^{\sigma^2/2} \leq 2\sigma^2\sqrt{\epsilon}N$ then $r_G^*(\delta) \leq (128/3)\sigma^4 N\sqrt{\epsilon}$.

End of the proof of Theorem 31 and Corollary 32: application of Theorem 5 Take $\delta = \exp(-\epsilon\sigma^4 N)$ (for $N = d(d-1)/2$), we have $r_G^*(\delta) \leq (128/3)\sqrt{\epsilon}\sigma^4 N$ when $e^{\sigma^2/2} \leq 2\sqrt{\epsilon}\sigma^2 N$ (which holds for instance when $\sigma \leq \sqrt{\log(\epsilon N^2)}$) and so it follows from Theorem 5 (together with the curvature property of Proposition 6.5 and the computation of the fixed point $r_G^*(\delta)$ just above, that with probability at least $1 - \exp(-\epsilon\sigma^4 d(d-1)/2)$, $\theta\langle Z^* - Z\rangle_2^2 \leq \langle \mathbb{E}A, Z^* - Z\rangle \leq (128/3)\sqrt{\epsilon}\sigma^4 N$, which is the statement of Theorem 31.

Proof of Corollary 32: The oracle Z^* is the rank one matrix $x^*\overline{x^*}^\top$ which has d for largest eigenvalue and associated eigenspace $\{\lambda x^* : \lambda \in \mathbb{C}\}$. In particular, Z^* has a spectral gap $g = d$. Let $\hat{x} \in \mathbb{C}^d$ be a top eigenvector of $\hat{Z}$ with norm $\|\hat{x}\|_2 = \sqrt{d}$. It follows from the Davis-Kahan Theorem (see, for example, Theorem 4.5.5 in (Vershynin, 2018) or Theorem 4 in (Vu, 2010)) that there exists a universal constant $c_0 > 0$ such that

$$\min_{z \in \mathbb{C}:|z|=1} \|\frac{\hat{x}}{\sqrt{d}} - z\frac{x^*}{\sqrt{d}}\|_2 \leq \frac{c_0}{g}\|\hat{Z} - Z^*\|_2,$$

where $g = d$ is the spectral gap of Z^*. We conclude the proof of Corollary 32 using the upper bound on $\|\hat{Z} - Z^*\|_2$ from Theorem 31. ∎

6.2.4 Max-Cut

In this section, we prove the two main results from Section 4.5 using our general methodology for Theorem 36 and the technique from Goemans and Williamson (1995) for Theorem 35.

6.2.4.1 Proof of Theorem 35

The proof of Theorem 35 follows the one from Goemans and Williamson (1995) up to a minor modification due to the fact that we use the SDP estimator $\hat{Z}$ instead of the oracle Z^*. It is based on two tools. The first one is Grothendieck's identity: let $g \sim \mathcal{N}(0, \mathrm{I}_d)$ and $u, v \in \mathcal{S}_2^{d-1}$, we have

$$\mathbb{E}[\mathrm{sign}(\langle g, u\rangle)\mathrm{sign}(\langle g, v\rangle)] = \frac{2}{\pi}\arcsin(\langle u, v\rangle), \tag{6.89}$$

and the identity: for all $t \in [-1, 1]$

$$1 - \frac{2}{\pi}\arcsin(t) = \frac{2}{\pi}\arccos(t) \geq 0.878(1-t). \tag{6.90}$$

We now have enough tools to prove Theorem 35. The right-hand side inequality is trivial since $\text{Max-Cut}(G) \leq \mathrm{SDP}(G)$. For the left-hand side, we denote by $\hat{X}_1, \ldots, \hat{X}_d$ (resp. $X_1^*, \ldots, X_d^*$) the d columns vectors in $\mathcal{S}_2^{d-1}$ of $\hat{Z}$ (resp. Z^*). We also consider the event Ω^* onto which $\langle \mathbb{E}B, Z^* - \hat{Z}\rangle \leq r^*(\delta)$, which hold with probability at least $1 - \delta$ according to Theorem 3. On

the event Ω^*, we have

$$\mathbb{E}\left[\text{cut}(G,\hat{x})|\hat{Z}\right] = \mathbb{E}\left[\frac{1}{4}\sum_{i,j=1}^{d} A^0_{ij}(1-\hat{x}_i\hat{x}_j)\right] = \frac{1}{4}\sum_{i,j=1}^{d} A^0_{ij}\left(1-\mathbb{E}[\text{sign}(\langle \hat{X}_i, g\rangle)\text{sign}(\langle \hat{X}_j, g\rangle)]\right)$$
$$\overset{(i)}{=} \frac{1}{4}\sum_{i,j=1}^{d} A^0_{ij}\left(1-\frac{2}{\pi}\arcsin(\langle \hat{X}_i, \hat{X}_j\rangle)\right) = \frac{1}{2\pi}\sum_{i,j=1}^{d} A^0_{ij}\arccos(\langle \hat{X}_i, \hat{X}_j\rangle)$$
$$\overset{(ii)}{\geq} \frac{0.878}{4}\sum_{i,j=1}^{d} A^0_{ij}(1-\langle \hat{X}_i, \hat{X}_j\rangle)$$
$$= \frac{0.878}{4}\sum_{i,j=1}^{d} A^0_{ij}(1-\langle X^*_i, X^*_j\rangle) + \frac{0.878}{4}\sum_{i,j=1}^{d} A^0_{ij}(\langle X^*_i, X^*_j\rangle - \langle \hat{X}_i, \hat{X}_j\rangle)$$
$$= \frac{0.878}{4}\langle A^0, J-Z^*\rangle + \frac{0.878}{4}\langle A^0, Z^*-\hat{Z}\rangle = 0.878\text{SDP}(G) - \frac{0.878}{4}\langle \mathbb{E}B, Z^*-\hat{Z}\rangle$$
$$\geq 0.878\ \text{SDP}(G) - \frac{0.878}{4}r^*(\delta)$$

where we used (6.89) in *(i)* and (6.90) in *(ii)*.

6.2.4.2 Proof of Theorem 36

For the Max-Cut problem, we do not use any localization argument; we therefore use the (likely sub-optimal) global approach. The methodology is very close to the one used in Guédon and Vershynin (2016) for the community detection problem. In particular, we use both Bernstein and Grothendieck inequalities to compute high-probability upper bound for $r^*(\delta)$. We recall theses two tools now. For Bernstein's inequality, see Equation (6.88) above. The second tool is Grothendieck inequality (Grothendieck, 1956) (see also Pisier (2012) or Theorem 3.4 in Guédon and Vershynin (2016)): if $C \in \mathbb{R}^{d\times d}$ then

$$\sup_{Z\in\mathcal{C}}\langle C, Z\rangle \leq K_G\|C\|_{cut} = K_G \max_{s,t\in\{-1,1\}^d}\sum_{i,j=1}^{d} C_{ij}s_it_j \tag{6.91}$$

where $\mathcal{C} = \{Z \succeq 0 : Z_{ii} = 1, i = 1,\dots,d\}$ and K_G is an absolute constant, called the Grothendieck constant.

In order to apply Theorem 3, we just have to compute the fixed point $r^*(\delta)$. As announced, we use the global approach and Grothendieck inequality (6.91) to get

$$\sup_{Z\in\mathcal{C}:\left\langle \mathbb{E}B, Z^*-Z\right\rangle\leq r}\left\langle B-\mathbb{E}B, Z-Z^*\right\rangle \leq \sup_{Z\in\mathcal{C}}\left\langle B-\mathbb{E}B, Z-Z^*\right\rangle \leq 2K_G\|B-\mathbb{E}B\|_{cut}, \tag{6.92}$$

because $Z^* \in \mathcal{C}$. It follows from Bernstein's inequality (6.88) and a union bound that for all $t > 0$, with probability at least $1-4^d\exp(-t)$,

$$\|B-\mathbb{E}B\|_{cut} = \sup_{s,t\in\{\pm1\}^d}\sum_{1\leq i<j\leq d}(B_{ij}-\mathbb{E}B_{ij})(s_it_j+s_jt_i) \leq 2\sqrt{\frac{(1-p)d(d-1)t}{p}} + \frac{4t}{3}.$$

Therefore, for $t = 2d\log 4$, with probability at least $1-4^{-d}$,

$$r^*(\delta) \leq \|B-\mathbb{E}B\|_{cut} \leq 4K_G\left(2d\sqrt{\frac{(2\log 4)(1-p)(d-1)}{p}} + \frac{8d\log 4}{3}\right)$$

for $\delta = 4^{-d}$. Then the result follows from Theorem 3.

6.2.5 Sparse PCA

6.2.5.1 Proof of Lemma 38

Let $Z \in \mathcal{C}$ and consider its SVD $Z = \sum_i \sigma_i u_i u_i^\top$. We have

$$\begin{aligned}\langle \Sigma, (\beta^*)(\beta^*)^\top - Z\rangle &= \theta\langle (\beta^*)(\beta^*)^\top, (\beta^*)(\beta^*)^\top - Z\rangle + \langle I_d, (\beta^*)(\beta^*)^\top - Z\rangle \\ &\overset{(i)}{=} \theta\langle (\beta^*)(\beta^*)^\top, (\beta^*)(\beta^*)^\top - \sum_i \sigma_i u_i u_i^\top\rangle \\ &= \theta\left(1 - \sum_i \sigma_i \langle u_i, \beta^*\rangle^2\right) \\ &= \theta \sum_i \sigma_i (1 - \langle u_i, \beta^*\rangle^2) \overset{(ii)}{\geq} 0\end{aligned}$$

where we used in *(i)* that $\langle I_d, (\beta^*)(\beta^*)^\top - Z\rangle = \mathrm{Tr}\left((\beta^*)(\beta^*)^\top\right) - \mathrm{Tr}\left(Z\right) = 0$, and in *(ii)* that $|\langle u_i, \beta^*\rangle| \leq 1$ (by Cauchy-Schwart). Hence, $\beta^*(\beta^*)^\top$ is a solution to the problem $\max\Big(\langle \Sigma, Z\rangle, Z \in \mathcal{C}\Big)$. Moreover, using the latter computation, it is straightforward to check that it is unique, that is $\sigma_1 = 1$ and $u_1 u_1^\top = \beta^*(\beta^*)^\top$, otherwise inequality *(ii)* would be strict.

6.2.5.2 Proof of Lemma 39

Let $Z \in \mathcal{C}$ and consider its SVD $Z = \sum_i \sigma_i u_i u_i^\top$. In the proof of Lemma 38, we proved that

$$\langle \Sigma, Z^* - Z\rangle = \theta \sum_i \sigma_i (1 - \langle u_i, \beta^*\rangle^2).$$

On the other-hand, we have

$$\begin{aligned}\|Z^* - Z\|_2^2 &= \mathrm{Tr}\left((Z^* - Z)(Z^* - Z)^\top\right) \\ &= \mathrm{Tr}\left(\left(\sum_i \sigma_i((\beta^*)(\beta^*)^\top - u_i u_i^\top)\right)^2\right) \\ &= \sum_{i,j} \sigma_i \sigma_j \mathrm{Tr}\left((u_i u_i^\top - (\beta^*)(\beta^*)^\top)(u_j u_j^\top - (\beta^*)(\beta^*)^\top)\right) \\ &= \sum_i \sigma_i \left(\sigma_i - 2\langle u_i, \beta^*\rangle^2 + 1\right) \\ &= 2\sum_i \sigma_i (1 - 2\langle u_i, \beta^*\rangle^2) + \left(\sum_i \sigma_i^2 - \sigma_i\right) \\ &= \frac{2}{\theta}\langle \Sigma, Z^* - Z\rangle + \left(\|Z\|_2^2 - \|Z\|_*\right) \leq \frac{2}{\theta}\langle \Sigma, Z^* - Z\rangle.\end{aligned}$$

6.2.5.3 Proof of Lemma 40

It follows from the k-sparsity of β^* that $Z^* = \beta^*(\beta^*)^\top$ is k^2-sparse. Let us denote $I := \mathrm{supp}(Z^*)$: we have $|I| \leq k^2$. Consider $\rho > 0$. To solve the sparsity equation, we will use the following result on the sub-differential of a norm: if $\|.\|$ is a norm over $\mathbb{R}^{d\times d}$, we have for $Z \in \mathbb{R}^{d\times d}$:

$$\partial\|.\|(Z) = \begin{cases} \{\Phi \in S^* : \langle \Phi, Z\rangle = \|Z\|\} & \text{if } Z \neq 0 \\ B^* & \text{if } Z = 0 \end{cases}$$

where S^* (resp. B^*) is the unit-sphere (resp. unit-ball) for the dual norm associated with $\|.\|$, that is $Z \in \mathbb{R}^{d\times d} \to \|Z\|^* = \sup_{\|H\|=1}\langle Z, H\rangle$. Here, we consider the ℓ_1-norm, whose dual norm

is the ℓ_∞ norm.

Since $Z^* \in Z^* + (\rho/20)B$, we have

$$\partial\|.\|_1(Z^*) \subset \Gamma_{Z^*}(\rho) := \bigcup_{V \in Z^* + (\rho/20)B} \partial\|.\|_1(V).$$

Then, there exists $\Phi^* \in \Gamma_{Z^*}(\rho)$ which is norming for Z^*, that is $\|\Phi^*\|_\infty = 1$ and $\langle \Phi^*, Z^* \rangle = \|Z^*\|_1$. Let $Z \in H_{\rho,A} := Z^* + (\rho S_1 \cap \sqrt{r^*(\rho)}B_2)$. For $J \subset [d]^2$, let P_J be the coordinate projection on J. Since the supports of $P_{I^c}Z$ and Z^* are disjoints, we can choose Φ^* such that it is also norming for $P_{I^c}Z$. Then, we have:

$$\begin{aligned}
\langle \Phi^*, Z - Z^* \rangle &= \langle \Phi^*, P_I(Z - Z^*) \rangle + \langle \Phi^*, P_{I^c}(Z - Z^*) \rangle \\
&\geq -|\langle \Phi^*, P_I(Z - Z^*) \rangle| + \|P_{I^c}Z\|_1 \\
&\geq -\|\Phi^*\|_\infty \|P_I(Z - Z^*)\|_1 + \|P_{I^c}Z\|_1 \\
&= -\|P_I(Z - Z^*)\|_1 + \|P_{I^c}(Z - Z^*)\|_1 \\
&= \|Z - Z^*\|_1 - 2\|P_I(Z - Z^*)\|_1 \\
&= \rho - 2\|P_I(Z - Z^*)\|_1.
\end{aligned}$$

Now, we have $\|P_I(Z - Z^*)\|_1 \leq k\|P_I(Z - Z^*)\|_2 \leq k\|Z - Z^*\|_2 \leq k\sqrt{r^*(\rho)}$. We conclude that $\langle \Phi^*, Z - Z^* \rangle \geq \rho - 2k\sqrt{r^*(\rho)}$. Then, $\sup_{\Phi \in \Gamma_{Z^*}(\rho)} \langle \Phi, Z - Z^* \rangle \geq \langle \Phi^*, Z - Z^* \rangle \geq \rho - 2k\sqrt{r^*(\rho)}$. Since this is true for any $Z \in H_{\rho,A}$, we conclude that $c \geq \rho - 2k\sqrt{r^*(\rho)}$, where $P_I(Z - Z^*)$ is the quantity introduced in Definition 8. Then, if we choose ρ such that $\rho \geq 10k\sqrt{r^*(\rho)}$, we have $\Delta(\rho, A) \geq (4/5)\rho$, and the A-sparsity equation is satisfied by such a ρ.

6.2.5.4 Proof of Lemma 41

From Lemma 39, we get that Assumption 3.3 holds with $G : Z \in \mathbb{R}^{d \times d} \to \|Z\|_2^2$ and $A = 2/\theta$, for any $\rho > 0$ and $\delta \in (0, 1)$. Moreover, Assumption 4.1 is granted for $t = \log(ed/10k)$ and $w \geq 0$. Let then $c_0 > 0$ be the constant provided by Theorem 25, and define $b = 3c_0w^2$. Let us define the following function:

$$r : \rho > 0 \to bA\sqrt{\frac{\rho^2}{N}\log\left(\frac{b^2A^2(ed)^4}{N\rho^2}\right)}.$$

We also consider

$$\rho^* := 200Abk^2\sqrt{\frac{1}{N}\log\left(\frac{ed}{k}\right)},$$

as well as $r^* = r(\rho^*)$. We have:

$$\begin{aligned}
100k^2r^* &= 100k^2bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\frac{(ed)^4}{4.10^4k^4\log(\frac{ed}{k})}\right)} \\
&\leq 200k^2bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\left(\frac{ed}{k}\right)^4\right)} = (\rho^*)^2,
\end{aligned} \tag{6.93}$$

so that $\rho^* \geq 10k\sqrt{r^*}$. Let us then define $k^* := \rho^*/\sqrt{r^*}$. Since $k^* > k$, any $2 \leq r \leq 2\log(ed/k^*) + t$ satisfies $2 \leq r \leq 2\log(ed/k) + t$, so that Assumption 4.1 holds with w, t and k^*. We are then in measure to apply Theorem 25 with those parameters. As a consequence, as soon as

$N \geq 2\log(ed/k^*) + t$, one has with probability at least $1 - \exp(-t)$:

$$\|\hat{\Sigma}_N - \Sigma\|_{k^*} \leq c_0 w^2 \sqrt{\frac{(k^*)^2\left(\log\left(\frac{ed}{k^*}\right) + t\right)}{N}} \tag{6.94}$$

where $\|.\|_{k^*}$ is the ℓ_1/ℓ_2 interpolation norm defined in (4.2). Now, we have:

$$\begin{aligned}\sup_{Z \in \mathcal{C} \cap (Z^* + \rho^* B_1 \cap \sqrt{r^*} B_2)} |(P - P_N)\mathcal{L}_Z| &\leq \sqrt{r^*} \sup_{Z \in \mathcal{C} \cap (Z^* + k^* B_1 \cap B_2)} |(P - P_N)\mathcal{L}_Z| \\ &= \sqrt{r^*} \left\| \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top - \mathbb{E}[X_i X_i^\top] \right\|_{k^*} \\ &= \sqrt{r^*} \|\hat{\Sigma}_N - \Sigma\|_{k^*}. \end{aligned} \tag{6.95}$$

Combining it with (6.94), we get that with probability at least $1 - \exp(-t)$:

$$\begin{aligned}\sup_{Z \in \mathcal{C} \cap (Z^* + \rho^* B_1 \cap \sqrt{r^*} B_2)} |(P - P_N)\mathcal{L}_Z| &\leq c_0 w^2 \sqrt{\frac{(\rho^*)^2\left(\log\left(\frac{ed}{k^*}\right) + t\right)}{N}} \\ &\leq c_0 w^2 \sqrt{\frac{2(\rho^*)^2}{N} \log\left(\frac{ed}{10k}\right)} \end{aligned} \tag{6.96}$$

since $k^* \geq 10k$. Now, we have:

$$\begin{aligned} r^* &= bA\sqrt{\frac{(\rho^*)^2}{N} \log\left(\frac{b^2 A^2 (ed)^4}{N(\rho^*)^2}\right)} \\ &= bA\sqrt{\frac{(\rho^*)^2}{N} \log\left(\frac{(ed)^4}{200^2 k^4 \log(ed/10k)}\right)} \\ &\geq bA\sqrt{\frac{(\rho^*)^2}{N} \log\left(\frac{(ed)^3}{200^2 k^3}\right)} \\ &\geq bA\sqrt{\frac{(\rho^*)^2}{N} \log\left(\frac{ed}{10k}\right)}, \end{aligned}$$

where the last inequality holds since we assumed that $k \leq ed/200$. Combining it with (6.96), we conclude that:

$$\sup_{Z \in \mathcal{C} \cap (Z^* + \rho^* B_1 \cap \sqrt{r^*} B_2)} |(P - P_N)\mathcal{L}_Z| \leq \frac{r^*}{3A}$$

which allows us to conclude that $r^*_{RERM,G}(A, \rho^*, e^{-t}) \leq r^*$. Moreover, we have from (6.93) that

$$\rho^* \geq 10k\sqrt{r^*} \geq 10k\sqrt{r^*_{RERM,G}(A, \rho^*, e^{-t})}$$

that is, ρ^* satisfies the A-sparsity equation from Definition 8. These results are valid provided that $N \geq 2\log(ed/k^*) + t$, which is ensured by the assumption that $N \geq 3\log(ed/10k)$, given that $k^* \geq 10k$. This concludes the proof.

6.2.5.5 Proof of Theorem 42

From Lemma 39, we get that Assumption 3.3 holds with $G : Z \in \mathbb{R}^{d \times d} \to \|Z\|_2^2$ and $A = 2/\theta$, for any $\rho > 0$ and $\delta \in (0, 1)$. Moreover, since we assumed that $N \geq 3\log(ed/10k)$, Lemma 41

applies and so, for ρ^* and $r^*(\rho^*)$ (defined in (4.35)) we have $r^*_{\text{RERM,G}}(A, \rho^*, 10k/ed) \leq r^*(\rho^*)$ and ρ^* satisfies the A-sparsity equation from Definition 8. We are then in position to apply Theorem 9, provided that λ satisfies (3.7). Now, we have:

$$\frac{r^*(\rho^*)}{\rho^*} = bA\sqrt{\frac{1}{N}\log\left(\frac{(bA)^2(ed)^4}{N(\rho^*)^2}\right)} = bA\sqrt{\frac{1}{N}\log\left(\frac{(ed)^4}{200^2k^4\log(\frac{ed}{k})}\right)},$$

so that:

$$bA\sqrt{\frac{3}{N}\log\left(\frac{ed}{200^{2/3}k}\right)} \leq \frac{r^*(\rho^*)}{\rho^*} \leq bA\sqrt{\frac{4}{N}\log\left(\frac{ed}{200^{1/2}\log(200)^{1/4}k}\right)},$$

since we assumed that $k \leq ed/200$. As a consequence, (3.7) is satisfied as soon as:

$$\frac{20}{21}b\sqrt{\frac{1}{N}\log\left(\frac{ed}{200^{1/2}\log(200)^{1/4}k}\right)} \leq \lambda \leq \frac{2}{\sqrt{3}}b\sqrt{\frac{1}{N}\log\left(\frac{ed}{200^{2/3}k}\right)}$$

which is the assumption made in (4.36). We only have to check that this authorized interval for λ is not empty, which is ensured as soon as $ed/k \geq 200^{48/47}/\log(200)^{25/47}$, which is granted by the assumption that $k \leq ed/200$.

We are then in measure to apply Theorem 9, which enables us to state that, with probability at least $1 - 10k/ed$:

$$\|\hat{Z}^{\text{RERM}}_\lambda - Z^*\|_1 \leq \rho^* = 200Abk^2\sqrt{\frac{1}{N}\log\left(\frac{ed}{k}\right)} = 400bk^2\sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)},$$

$$\|\hat{Z}^{\text{RERM}}_\lambda - Z^*\|_2 \leq \sqrt{r^*_{\text{RERM,G}}\left(A, \rho^*, 10k/ed\right)} \leq \frac{\rho^*}{10k} = 40b\sqrt{\frac{k^2}{N\theta^2}\log\left(\frac{ed}{k}\right)}$$

$$\text{and} \quad P\mathcal{L}_{\hat{Z}^{\text{RERM}}_\lambda} \leq A^{-1}r^*_{\text{RERM,G}}\left(A, \rho^*, 10k/ed\right) \leq \frac{(\rho^*)^2}{100k^2A} = 800b^2\frac{k^2}{N\theta}\log\left(\frac{ed}{k}\right).$$

This concludes the proof.

6.2.5.6 Proof of Corollary 43

From Theorem 42, we get the existence of a universal constant $C > 0$ such that with probability at least $1 - 20\left(k/ed\right)^{3/4}$, $\|\hat{Z}^{\text{RERM}}_\lambda - Z^*\|_2 \leq C\sqrt{k^2(N\theta^2)\log(ed/k)}$. Now, we can use Davis-Kahan sin-theta theorem (see Corollary 1 in Yu et al. (2014)) to get the existence of a universal constant $c_0 > 0$ such that $\sin(\Theta(\hat{\beta}, \beta^*)) = (1/\sqrt{2})\|\hat{\beta}\hat{\beta}^\top - \beta^*(\beta^*)^\top\|_2 \leq (c_0/g)\|\hat{Z}^{\text{RERM}}_\lambda - Z^*\|_2$ where $g := \lambda_1 - \lambda_2$ (λ_i being the i^{th} largest eigen value of Z^*) is the spectral gap of Z^*. Here, we know that $Z^* = \beta^*(\beta^*)^\top$ is rank one, with 1 as order one eigen value and 0 as order $(d-1)$ eigen value. Then we get $g = 1$, which leads us to the desired result, with $D = \sqrt{2}c_0 \times C$.

6.2.5.7 Proof of Lemma 44

Let A, δ and $t > 0$. In the rest of the proof, we write $r^*_G(.)$ for $r^*_{\text{RERM,G}}(A, ., \delta)$, b_{pq} for $b_{pq}(t)$ and Γ_k for $\Gamma_k(t)$. We consider a lexicographical order on $[d]^2$, $b \in \mathbb{R}^{d\times d}$ and the norm $\|.\|_{SLOPE}$ as they are defined in section 4.6.4.

Let $I := \text{supp}\big((Z^*)^\sharp\big)$ be the set of non-zero coefficients of $(Z^*)^\sharp$. Since $Z^* = \beta^*(\beta^*)^\top$ is k^2-sparse, whe have by construction that $|I| \leq k^2$. Let P_I (resp. P_{I^c}) be the coordinate projection on I (resp. on I^c).

We know that for $Z \neq 0$:

$$\partial\|.\|_{SLOPE}(Z) = \left\{\Phi \in S^*_{SLOPE} : \langle \Phi, Z\rangle = \|Z\|_{SLOPE}\right\},$$

where we denoted S^*_{SLOPE} the unit-sphere of the dual norm of the $SLOPE$ norm. Since $Z^* \in Z^* + \frac{\rho}{20}B_{SLOPE}$, we know that $\partial\|.\|_{SLOPE}(Z^*) \subset \Gamma_{Z^*}(\rho)$. Then:

$$\sup_{\Phi\in\Gamma_{Z^*}(\rho)} \langle \Phi, Z - Z^*\rangle \geq \sup_{\Phi\in\partial\|.\|_{SLOPE}(Z^*)} \langle \Phi, Z - Z^*\rangle.$$

Let σ, π be the permutations of $[d]^2$ such that, for any $(p,q) \in [d]^2$, $(Z^*)^\sharp_{p,q} = |Z^*_{\sigma(p,q)}|$ and $(Z - Z^*)^\sharp_{p,q} = |(Z - Z^*)_{\pi(p,q)}|$. Notice that we have by assumption $\sigma([k]^2) = I$. We then define Φ^* and $\widetilde{\Phi}^*$ as follows: for all $1 \leq p, q \leq d$,

$$\Phi^*_{p,q} = \begin{cases} \operatorname{sgn}(Z^*_{p,q})\, b_{\sigma^{-1}(p,q)} & \text{if } (p,q) \leq (k,k); \\ \operatorname{sgn}((Z - Z^*)_{p,q})\, b_{\pi^{-1}(p,q)} & \text{otherwise} \end{cases}$$

and

$$\widetilde{\Phi}^*_{p,q} = \operatorname{sgn}((Z - Z^*)_{p,q})\, b_{\pi^{-1}(p,q)}.$$

We easily check that such a Φ^* belongs to $\partial||\cdot||_{SLOPE}(Z^*)$ and $\widetilde{\Phi}^*$ to $\partial||\cdot||_{SLOPE}(Z - Z^*)$. Now let $Z \in H_{\rho,A}$. We have:

$$\begin{aligned} \langle \Phi^*, Z - Z^*\rangle &= \left\langle \Phi^*, P_I(Z - Z^*)\right\rangle + \left\langle \Phi^*, P_{I^c}(Z - Z^*)\right\rangle \\ &= \sum_{p,q=1}^{k} \operatorname{sgn}(Z^*_{\sigma(p,q)}) b_{p,q} (Z - Z^*)_{\sigma(p,q)} + \left\langle \Phi^*, P_{I^c}(Z - Z^*)\right\rangle. \end{aligned} \tag{6.97}$$

Regarding the first term, we have:

$$\left|\sum_{p,q=1}^{k} \operatorname{sgn}(Z^*_{\sigma(p,q)}) b_{p,q} (Z - Z^*)_{\sigma(p,q)}\right| \leq \sum_{p,q=1}^{k} b_{p,q} \left|(Z - Z^*)_{\pi(p,q)}\right| = \sum_{p,q=1}^{k} b_{p,q} (Z - Z^*)^\sharp_{p,q},$$

where the first inequality comes from the fact that the operator $(\cdot)^\sharp$ orders the absolute values of $(Z - Z^*)$ in non-increasing order (notice that the inequality holds only for the sum, not for each independent term of the sum). Therefore:

$$\sum_{p,q=1}^{k} \operatorname{sgn}(Z^*_{\sigma(p,q)}) b_{p,q} (Z - Z^*)_{\sigma_{p,q}} \geq -\sum_{p,q=1}^{k} b_{p,q} (Z - Z^*)^\sharp_{p,q}. \tag{6.98}$$

Concerning the second term in (6.97):

$$\begin{aligned} \left\langle \Phi^*, P_{I^c}(Z - Z^*)\right\rangle &= \left\langle \widetilde{\Phi}^*, P_{I^c}(Z - Z^*)\right\rangle = \left\langle \widetilde{\Phi}^*, Z - Z^*\right\rangle - \left\langle \widetilde{\Phi}^*, P_I(Z - Z^*)\right\rangle \\ &= ||Z - Z^*||_{SLOPE} - \sum_{p,q=1}^{k} b_{\pi^{-1}\circ\sigma(p,q)} (Z - Z^*)^\sharp_{\pi^{-1}\circ\sigma(p,q)} \\ &\geq ||Z - Z^*||_{SLOPE} - \sum_{p,q=1}^{k} b_{p,q} (Z - Z^*)^\sharp_{p,q}. \end{aligned} \tag{6.99}$$

Putting (6.97), (6.98) and (6.99) together, we obtain

$$\langle \Phi^*, Z - Z^*\rangle \geq ||Z - Z^*||_{SLOPE} - 2\sum_{p,q=1}^{k} b_{p,q} (Z - Z^*)^\sharp_{p,q} = \rho - 2\sum_{p,q=1}^{k} b_{p,q} (Z - Z^*)^\sharp_{p,q}. \tag{6.100}$$

Now, since $\|Z - Z^*\|_2 \leq \sqrt{r_G^*}$, we can show that for any $k \in [d]$, $(Z - Z^*)^\sharp_{kk} \leq \sqrt{r_G^*}/k$. Indeed, assume the existence of $k_0 \in [d]$ such that $(Z - Z^*)^\sharp_{k_0 k_0} > \frac{\sqrt{r_G^*}}{k_0}$. Then by construction we have that for any $(p, q) \leq (k_0, k_0)$, $(Z - Z^*)^\sharp_{pq} \geq (Z - Z^*)^\sharp_{k_0 k_0}$, so that

$$\|Z - Z^*\|_2^2 = \|(Z - Z^*)^\sharp\|_2^2 \geq \sum_{(p,q)\leq(k_0,k_0)} ((Z - Z^*)^\sharp_{pq})^2 > \sum_{(p,q)\leq(k_0,k_0)} \frac{r_G^*}{k_0^2} = r_G^*,$$

since the k_0^2 largest elements of $(Z - Z^*)^\sharp$ belong to $[k_0]^2$, as a result of which $|\{(p, q) : (p, q) \leq (k_0, k_0)\}| \leq k_0^2$. This is inconsistent with the fact that $\|Z - Z^*\|_2 \leq \sqrt{r_G^*}$.

As a consequence, we have:

$$\begin{aligned}
\sum_{p,q=1}^{k} b_{pq}(Z - Z^*)^\sharp_{pq} &= \sum_{\ell=1}^{k-1} \sum_{(\ell,\ell)\leq(p,q)<(\ell+1,\ell+1)} b_{pq}(Z - Z^*)^\sharp_{\ell\ell} + b_{kk}(Z - Z^*)^\sharp_{kk} \\
&\leq \sum_{\ell=1}^{k-1} \Big|\{(\ell, \ell) \leq (p, q) < (\ell + 1, \ell + 1)\}\Big| \, b_{\ell\ell}(Z - Z^*)^\sharp_{\ell\ell} + b_{kk}\frac{\sqrt{r_G^*}}{k} \\
&\leq \sum_{\ell=1}^{k-1}(2\ell + 1)b_{\ell\ell}\frac{\sqrt{r_G^*}}{\ell} + b_{kk}\frac{\sqrt{r_G^*}}{k} \leq 3\sqrt{r_G^*}\sum_{\ell=1}^{k-1} b_{\ell\ell} + b_{kk}\frac{\sqrt{r_G^*}}{k} \leq 3\sqrt{r_G^*}\sum_{\ell=1}^{k} b_{\ell\ell} \\
&= \sqrt{r_G^*}\Gamma_k.
\end{aligned}$$

Then, under the assumption that $\rho \geq 10\Gamma_k\sqrt{r_G^*(A, \rho, \delta)}$, we get from (6.100) that $\langle \Phi^*, Z - Z^* \rangle \geq (4/5)\rho$. and then:

$$\sup_{\Phi \in S^*_{SLOPE}} \langle \Phi, Z - Z^* \rangle \geq \langle \Phi^*, Z - Z^* \rangle \geq \frac{4}{5}\rho.$$

Since this is true for any $Z \in H(\rho, A)$, we conclude that:

$$\Delta(\rho, A) = \inf_{Z \in H(\rho,A)} \sup_{\Phi \in S^*_{SLOPE}} \langle \Phi^*, Z - Z^* \rangle \geq \frac{4}{5}\rho.$$

that is, ρ satisfies the A-sparsity equation from Definition 8.

6.2.5.8 Proof of Lemma 45

From Lemma 39, we get that Assumption 3.3 holds with $G : Z \in \mathbb{R}^{d\times d} \to \|Z\|_2^2$ and $A = 2/\theta$, for any $\rho > 0$ and $\delta \in (0, 1)$.

For r and $\rho > 0$, we define $\mathcal{C}_{r,\rho} := \{Z \in \mathcal{C} : \|Z - Z^*\|_{SLOPE} \leq \rho, \|Z - Z^*\|_2 \leq \sqrt{r}\}$. Let $A > 0$. For any ρ and $r > 0$. We have

$$\begin{aligned}
\sup_{Z\in\mathcal{C}_{r,\rho}} |\langle \Sigma - \hat{\Sigma}_N, Z - Z^* \rangle| &\leq \sup_{Z\in(\rho B_{SLOPE}\cap\sqrt{r}B_2)} |\langle \Sigma - \hat{\Sigma}_N, Z \rangle| \\
&= \sqrt{r} \sup_{Z\in(\frac{\rho}{\sqrt{r}}B_{SLOPE}\cap B_2)} |\langle \Sigma - \hat{\Sigma}_N, Z \rangle| \\
&= \sqrt{r}\|\Sigma - \hat{\Sigma}_N\|_{\frac{\rho}{\sqrt{r}}}
\end{aligned} \tag{6.101}$$

where $\|.\|_{\rho/\sqrt{r}}$ is the $SLOPE/\ell_2$ interpolation norm defined in (4.3). Assumption 4.2 is granted for $t = 2\log(ed^2/k^2)$. Let us now check that $k \leq d/(e^2\log(d))$: we have that $k^2\log(ek^2) \leq ed^2$, hence,

$$2\log(\lceil\log(k^2)\rceil) \leq 2\log(\log(k^2) + 1) = 2\log(\log(ek^2)) \leq 2\log\left(\frac{ed^2}{k^2}\right),$$

that is, $t \geq \max\left(2\log(\lceil\log(k^2)\rceil), 2\log(ed^2/k^2)\right)$. We are then in position to apply Theorem 26 with $\gamma = 2$ and $t = 2\log(ed^2/k^2)$: there exists a universal constant $c_0 > 0$ such that, provided that $N \geq \log\left(ed^2\right) + t$, one has with probability at least $1 - 2\exp(-t/2)$:

$$\|\Sigma - \hat{\Sigma}_N\|_{\frac{\rho}{\sqrt{r}}} \leq \frac{c_0 w^2}{\sqrt{N}} \min\left(\frac{\rho}{\sqrt{r}}, d\right).$$

Plugging this last result into (6.101), we get that:

$$\sup_{Z \in \mathcal{C}_{r,\rho}} |\langle \Sigma - \hat{\Sigma}_N, Z - Z^* \rangle| \leq \sqrt{r}\frac{c_0 w^2}{\sqrt{N}} \min\left(\frac{\rho}{\sqrt{r}}, d\right) \tag{6.102}$$

with probability at least $1 - 2\exp(-t/2)$. Next, let us define $b := 3c_0 w^2$ and for $\rho > 0$, consider

$$r^*(\rho) := \frac{bA}{\sqrt{N}} \min\left(bA\frac{d^2}{\sqrt{N}}; \rho\right).$$

One can check that for this choice of r^*, one has $(\sqrt{r^*(\rho)}c_0 w^2/\sqrt{N}) \min\left(\rho/\sqrt{r^*(\rho)}, d\right) \leq r^*(\rho)/3A$ whatever the value of ρ is. From (6.102) we then deduce that $r^*_{\mathrm{RERM,G}}(A, \rho, 2e^{-t/2}) \leq r^*(\rho)$. Let us now consider

$$\rho^* := 10\Gamma^*_k \frac{bA}{\sqrt{N}} \min\left(10\Gamma^*_k; d\right),$$

where $\Gamma^*_k := 3\sum_{\ell=1}^k b_{\ell\ell}(t)$. It is straighforward to verify that

$$\rho^* \geq 10\Gamma^*_k r^*(\rho^*)^{1/2} \geq 10\Gamma^*_k r^*_{\mathrm{RERM,G}}\left(A, \rho^*, 2e^{-t^*/2}\right)^{1/2}$$

which, according to Lemma 44, guarantees that ρ^* satisfies the A-sparsity equation from Definition 8. Finally, plugging the expression of ρ^* into the one of $r^*(\rho^*)$, we get that $r^*(\rho^*) = (b^2A^2/N)\min\left(d, 10\Gamma^*_k\right)^2$. The previous results hold provided that $N \geq \log\left(ed^2\right) + t$, which is granted by the assumption that $N \geq 3\log(ed^2)$. This concludes the proof, noting that $2\exp(-t/2) = 2k^2/(ed^2)$.

6.2.5.9 Proof of Theorem 46

From Lemmas 38 and 39, we get that Assumption 3.3 holds with $G : Z \to \|Z\|_2^2$ and $A = 2/\theta$. From Lemma 45, we get the existence of a constant $b > 0$ such that, provided that $N \geq 3\log(ed^2)$, defining $\rho^* := 10\Gamma^*_k(bA/\sqrt{N})\min\left(10\Gamma^*_k; d\right)$ and $r^* = (b^2A^2/N)\min\left(d, 10\Gamma^*_k\right)^2$, with $\Gamma^*_k = \Gamma_k(2\log(ed^2/k^2))$, one has $r^*_{\mathrm{RERM,G}}(A, \rho^*, 2k^2/ed^2) \leq r^*$ and ρ^* satsifies the A-sparsity equation from Definition 8. Let us now upper bound Γ^*_k:

$$\begin{aligned}
\Gamma^*_k &= 3\sum_{\ell=1}^k b_{\ell\ell}\left(2\log\left(\frac{ed^2}{k^2}\right)\right) \\
&\leq 3\left(\sum_{\ell=1}^k \sqrt{\log\left(\frac{ed^2}{\ell^2}\right)} + \sum_{\ell=1}^k \sqrt{2\log\left(\frac{ed^2}{k^2}\right)}\right) \\
&\leq 3\sum_{\ell=1}^k \sqrt{\log\left(\frac{ed^2}{\ell^2}\right)} + 3k\sqrt{2\log\left(\frac{ed^2}{k^2}\right)}.
\end{aligned} \tag{6.103}$$

Concerning the first term in this last inequality, we have:

$$\left(\sum_{\ell=1}^{k}\sqrt{\log\left(\frac{ed^2}{\ell^2}\right)}\right)^2 = 2\sum_{m<\ell}\sqrt{\log\left(\frac{ed^2}{\ell^2}\right)}\sqrt{\log\left(\frac{ed^2}{m^2}\right)} + \sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right)$$
$$\leq 2\sum_{m<\ell}\log\left(\frac{ed^2}{\ell^2}\right) + \sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right)$$
$$\leq 3k\sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right). \tag{6.104}$$

Moreover, we have:

$$\sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right) \leq \sum_{\ell=1}^{k}\int_{u=\ell-1}^{\ell}\log\left(\frac{ed^2}{u^2}\right)\mathrm{d}u$$
$$= \int_{u=0}^{k}\log\left(\frac{ed^2}{u^2}\right) = k\log(ed^2) - 2\left[u\log(u)-u\right]_0^k$$
$$= k\log\left(\frac{ed^2}{k^2}\right) + 2k \leq 3k\log\left(\frac{ed^2}{k^2}\right). \tag{6.105}$$

Combining (6.103), (6.104) and (6.105), we finally get that $\Gamma_k^* \leq (9+3\sqrt{2})k\sqrt{\log\left(\frac{ed^2}{k^2}\right)} \leq 14k\sqrt{\log\left(ed^2/k^2\right)}$. As a consequence, we have

$$10\Gamma_k^* \leq 140k\sqrt{\log\left(\frac{ed^2}{k^2}\right)} \leq 140k\sqrt{2\log\left(d\right)} \leq 140k\sqrt{\frac{2d}{e^2k}} \leq d,$$

since we assumed that $k \leq \min\left(d/(e^2\log(d)), (e/(140\sqrt{2}))^2 d\right)$. We conclude that $\min\left(10\Gamma_k^*, d\right) = 10\Gamma_k^* \leq 140k\sqrt{\log\left(ed^2/k^2\right)}$. Plugging this result into the expression of r^* and ρ^*, we finally get that:

$$r^* \leq 140^2 b^2 A^2 \frac{k^2}{N}\log\left(\frac{ed^2}{k^2}\right) \quad \text{and} \quad \rho^* \leq 140^2 bA\frac{k^2}{\sqrt{N}}\log\left(\frac{ed^2}{k^2}\right),$$

so that $r^*/\rho^* = bA/\sqrt{N}$. As a consequence, (3.7) is satisfied as soon as:

$$\frac{10b}{21\sqrt{N}} < \lambda < \frac{2b}{3\sqrt{N}}$$

which is (4.38). We are then in position to apply Theorem 9, which allows us to conclude that, with probability at least $1-2k^2/ed^2$:

$$\|\hat{Z}_\lambda^{RERM} - Z^*\|_{SLOPE} \leq \rho^* \quad , \quad G(\hat{Z}_\lambda^{RERM} - Z^*) \leq r^* \quad \textit{and} \quad P\mathcal{L}_{\hat{Z}_\lambda^{RERM}} \leq A^{-1}r^*.$$

This concludes the proof.

6.2.5.10 Proof of Corollary 47

The proof follows exactly the same lines as the one of Corollary 43, so we do not detail it here.

6.2.5.11 Proof of Lemma 48

Consider $A = 2/\theta$ and $\gamma > 0$. In the rest of the proof we write $r^*(\rho)$ for $r^*_{\mathrm{RMOM,G}}(A,\gamma,\rho)$. For any $J \subset [d]^2$, let P_J be the coordinate projection on J. Consider $\rho > 0$. Let $I := \mathrm{supp}(Z^*)$ be

the set of non-zero coefficients of Z^*. From Lemma 38, we have that $|I| \leq k^2$. Moreover, we know that for any $Z \neq 0$, $\partial\|.\|_1(Z) = \left\{\Phi \in S_\infty : \langle \Phi, Z\rangle = \|Z\|_1\right\}$, where S_∞ is the unit-sphere for $\|.\|_\infty$. Since $Z^* \in Z^* + \frac{\rho}{20}B_1$, we have that $\partial\|.\|_1(Z^*) \subset \Gamma_{Z^*}(\rho) = \underset{Z \in Z^* + \frac{\rho}{20}B_1}{\cup} \partial\|.\|_1(Z)$. Let then $\Phi^* \in \partial\|.\|_1(Z^*)$. Consider

$$Z \in \bar{H}_{\rho,A} := \left\{Z \in \mathcal{C} : \|Z - Z^*\|_1 = \rho \text{ and } \|Z - Z^*\|_2 \leq \sqrt{2/\theta}r^*(\rho)\right\}.$$

Since Z^* and $P_I^c(Z)$ have disjoint supports, we can choose Φ^* so that it is also norming for $P_I^c(Z)$. Then, we have:

$$\begin{aligned}
\langle \Phi^*, Z - Z^*\rangle &= \langle \Phi^*, P_I(Z - Z^*)\rangle + \langle \Phi^*, P_I^c(Z - Z^*)\rangle \\
&\geq -\|\Phi^*\|_\infty \|P_I(Z - Z^*)\|_1 + \langle \Phi^*, P_I^c(Z)\rangle \\
&= -(\|Z - Z^*\|_1 - \|P_I(Z - Z^*)\|_1) + \|P_I^c(Z - Z^*)\|_1 \\
&= 2\|P_I^c(Z - Z^*)\|_1 - \|Z - Z^*\|_1 \\
&= \|Z - Z^*\|_1 - 2\|P_I(Z - Z^*)\|_1 \\
&= \rho - 2\|P_I(Z - Z^*)\|_1
\end{aligned} \tag{6.106}$$

where we used the fact that $\|\Phi^*\|_\infty = 1$. Then, since $Z \in \bar{H}_{\rho,A}$, we have:

$$\|P_I(Z - Z^*)\|_1 \leq k\|P_I(Z - Z^*)\|_2 \leq k\|Z - Z^*\|_2 \leq k\sqrt{\frac{2}{\theta}}r^*(\rho) \tag{6.107}$$

Combining (6.106) and (6.107), we finally get that:

$$\langle \Phi^*, Z - Z^*\rangle \geq \rho - 2k\sqrt{\frac{2}{\theta}}r^*(\rho). \tag{6.108}$$

As a consequence, $\sup_{\Phi \in \Gamma_{Z^*}(\rho)}\langle \Phi, Z - Z^*\rangle \geq \langle \Phi^*, Z - Z^*\rangle \geq \rho - 2k\sqrt{2/\theta}r^*(\rho)$. This being true whatever $Z \in \bar{H}_{\rho,A}$, it follows that $\bar{\Delta}(\rho) \geq \rho - 2k\sqrt{2/\theta}r^*(\rho)$. We conclude that any ρ such that $\rho \geq 10k\sqrt{2/\theta}r^*(\rho)$ satisfies $\bar{\Delta}(\rho) \geq (4/5)\rho$.

6.2.5.12 Proof of Lemma 49

Consider $\gamma > 0$. From Lemma 39, we get that Assumption 3.8 holds with $G : Z \in \mathbb{R}^{d\times d} \to (\theta/2)\|Z\|_2^2$ and $A = 1$, for any $\gamma > 0$, in particular for the value of γ we have just set. Moreover, Assumption 4.1 is granted for $t = 1$ and $w \geq 0$. Let then $c_0 > 0$ be the constant provided by Theorem 25, and consider $B := 3c_0w^2$ and $D := 1600w^2$. Let us define the following function:

$$r : (\gamma, \rho) \to \max\left(\sqrt{\frac{B\rho}{\gamma}}\left(\frac{6}{N}\log\left(\frac{2B(ed)^2}{\gamma\theta\rho}\sqrt{\frac{6}{N}}\right)\right)^{1/4}; D\sqrt{\frac{K}{N\theta}}\right).$$

We also consider

$$\rho^* := \max\left(400\sqrt{3}B\frac{k^2}{\gamma}\sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)}; 10Dk\sqrt{\frac{2K}{N\theta^2}}\right),$$

as well as $r^*(\gamma) = r(\gamma, \rho^*)$. One can check that ρ^* such defined satisfies both of the two conditions below:

$$\begin{aligned}
&(1) \quad \rho \geq 10k\sqrt{\frac{2B\rho}{\theta\gamma}}\left(\frac{6}{N}\log\left(\frac{2B(ed)^2}{\gamma\theta\rho}\sqrt{\frac{6}{N}}\right)\right)^{1/4} \\
\text{and} \quad &(2) \quad \rho \geq 10kD\sqrt{\frac{2K}{N\theta^2}},
\end{aligned} \tag{6.109}$$

so that $\rho^* \geq 10k\sqrt{2/\theta}r^*$. Let us define $k^* = \sqrt{\theta/2}\rho^*/r^*$. We have $\log(ed/k^*) + 1 \leq \log(ed/10k) + 1$, so that Assumption 4.1 still holds with w, $t = 1$ and k^*. Then, since we assumed that $N \geq 2\log(ed/k) + 1 \geq 2\log(ed/k^*) + 1$, Theorem 25 applies and allows us to affirm that

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_i\widetilde{X}_i^\top - \mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top]\right\|_{k^*}\right] \leq c_0 w^2\sqrt{\frac{6(k^*)^2\log(ed/k^*)}{N}}, \tag{6.110}$$

where $\|\cdot\|_{k^*}$ is the ℓ_1/ℓ_2 interpolation norm defined in (4.2) for $k = k^*$.

For r and $\rho > 0$, define $\mathcal{C}_{r,\rho} := \left\{Z \in \mathcal{C} : \|Z - Z^*\|_1 = \rho \text{ and } \|Z - Z^*\|_2 \leq \sqrt{2/\theta}r^*(\rho)\right\}$. Let us now upper bound $E_G(r^*, \rho^*)$ and $V_{K,G}(r^*, \rho^*)$ from Definition 22.

Bounding the complexity term $E(r^*, \rho^*)$. Let $\sigma_1, \dots, \sigma_N$ be *i.i.d.* rademacher variables independent from the $\widetilde{X}_i$'s. We have $\mathcal{C}_{r^*,\rho^*} \subset \sqrt{2/\theta}r^*\,(k^*B_1 \cap B_2)$. As a consequence:

$$\begin{aligned}\sup_{Z\in\mathcal{C}_{r^*,\rho^*}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right| &\leq \sqrt{\frac{2}{\theta}}r^*\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right| \\ &= \sqrt{\frac{2}{\theta}}r^*\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\frac{1}{N}\sum_{i=1}^{N}\sigma_i\widetilde{X}_i\widetilde{X}_i^\top, Z\right\rangle\right|. \end{aligned}\tag{6.111}$$

Now, it follows from the desymmetrization inequality (see Theorem 2.1 in Koltchinskii (2011c)) that:

$$\begin{aligned}&\mathbb{E}\left[\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\frac{1}{N}\sum_{i=1}^{N}\sigma_i\widetilde{X}_i\widetilde{X}_i^\top, Z\right\rangle\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_i\widetilde{X}_i^\top - E[\widetilde{X}_i\widetilde{X}_i^\top], Z\right\rangle\right|\right] \\ &+ \frac{2}{\sqrt{N}}\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right| \\ \leq &2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_i\widetilde{X}_i^\top - E[\widetilde{X}_i\widetilde{X}_i^\top]\right\|_{k^*}\right] + \frac{2}{\sqrt{N}}\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right| \\ &\leq 2c_0w^2\sqrt{\frac{6(k^*)^2\log(ed/k^*)}{N}} + \frac{2}{\sqrt{N}}\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right|, \end{aligned}\tag{6.112}$$

where we used (6.110) in the last inequality.

Concerning the second term in (6.112), we have for any $Z \in (k^*B_1 \cap B_2) \cap (\mathcal{C} - Z^*)$:

$$\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top, Z\right]\right\rangle = \left\langle\theta\beta^*(\beta^*)^\top + Id, Z\right\rangle \overset{(i)}{=} \theta\left\langle\beta^*(\beta^*)^\top, Z\right\rangle \overset{(ii)}{\leq} \theta\|\beta^*\|_2^2\|Z\|_2 \leq \theta$$

where we used the fact that $\langle Id, Z\rangle = \mathrm{Tr}(Z) = 0$ in (i) and Cauchy-Schwarz in (ii). as a consequence:

$$\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right| \leq \theta. \tag{6.113}$$

Combining (6.111), (6.112) and (6.113), we finally get that:

$$\begin{aligned} E_G(r^*,\rho^*) &= \mathbb{E}\left[\sup_{\mathcal{C}_{r^*,\rho^*}}\left|\frac{1}{N}\sum_{i=1}^N \sigma_i \mathcal{L}_Z(\widetilde{X}_i)\right|\right] \\ &\leq \sqrt{\frac{2}{\theta}}r^*\left(2c_0w^2\sqrt{\frac{6(k^*)^2\log(ed/k^*)}{N}}+\frac{2\theta}{\sqrt{N}}\right) \\ &\leq \sqrt{2\theta}r^*\left(3c_0w^2\sqrt{\frac{6(k^*)^2\log(ed/k^*)}{\theta^2 N}}\right), \end{aligned} \tag{6.114}$$

where we used the assumption that $\theta \leq k \leq k^*$.

Bounding the variance term $V_{K,G}(r^*,\rho^*)$. Let us now upper bound the variance term

$$V_{K,G}(r^*,\rho^*) = \sqrt{\frac{K}{N}}\sup_{Z\in\mathcal{C}_{r^*,\rho^*}}\sqrt{\boldsymbol{V}ar(\mathcal{L}_Z(\widetilde{X}_i))}.$$

For $\widetilde{X}$ distributed as the $\widetilde{X}_i$'s and $Z \in \mathcal{C}_{r^*,\rho^*}$, one has:

$$\begin{aligned} &\boldsymbol{V}ar(\mathcal{L}_Z(\widetilde{X})) = \mathbb{E}[((\mathcal{L}_Z(\widetilde{X}) - P(\mathcal{L}_Z(\widetilde{X}))^2] = \mathbb{E}[\langle \widetilde{X}\widetilde{X}^\top - \mathbb{E}[\widetilde{X}\widetilde{X}^\top], Z-Z^*\rangle^2] \\ &= \sum_{p,q,s,t=1}^d \mathbb{E}\left[(\widetilde{X}^{(p)}\widetilde{X}^{(q)} - \mathbb{E}[\widetilde{X}^{(p)}\widetilde{X}^{(q)}])(\widetilde{X}^{(s)}\widetilde{X}^{(t)} - \mathbb{E}[\widetilde{X}^{(s)}\widetilde{X}^{(t)}])\right](Z-Z^*)_{pq}(Z-Z^*)_{st} \\ &= \sum_{p,q,s,t=1}^d T_{p,q,s,t}(Z-Z^*)_{pq}(Z-Z^*)_{st} \end{aligned}$$

where we defined $T_{p,q,s,t} := \mathbb{E}\left[(\widetilde{X}^{(p)}\widetilde{X}^{(q)} - \mathbb{E}[\widetilde{X}^{(p)}\widetilde{X}^{(q)}])(\widetilde{X}^{(s)}\widetilde{X}^{(t)} - \mathbb{E}[\widetilde{X}^{(s)}\widetilde{X}^{(t)}])\right]$ for all $1 \leq p,q,s,t \leq d$. Remembering that Assumption 4.1 is granted, we have:

$$T_{p,q,s,t} = \begin{cases} \|(\widetilde{X}^{(p)})^2 - \mathbb{E}[(\widetilde{X}^{(p)})^2]\|_{L_2}^2 \leq (2w^2)^2 & \text{if } p=q=s=t \\ \|\widetilde{X}^{(p)}\widetilde{X}^{(q)} - \mathbb{E}[\widetilde{X}^{(p)}\widetilde{X}^{(q)}]\|_{L_2}^2 \leq (2w^2)2 & \text{if } (p,q)=(s,t), p\neq q \\ 0 \text{ otherwise.} \end{cases}$$

Then:

$$\begin{aligned} \boldsymbol{V}ar(\mathcal{L}_Z(\widetilde{X})) &\leq \sum_{p=1}^d 4w^4(Z-Z^*)_{pp}^2 + \sum_{p\neq q} 4w^4(Z-Z^*)_{pq}^2 \\ &= \sum_{p,q=1}^q 4w^4(Z-Z^*)_{pq}^2 = 4w^4\|Z-Z^*\|_2^2 \leq (8w^4/\theta)(r^*)^2. \end{aligned}$$

This being true for any $Z \in \mathcal{C}_{\rho^*,r^*}$ and any $\widetilde{X}$ distributed as the $\widetilde{X}_i$'s, we conclude that:

$$V_{K,G}(r,\rho) \leq 2w^2\sqrt{\frac{2K}{N\theta}}r^*. \tag{6.115}$$

Combining (6.114) and (6.115), we finally get that:

$$\max\left(\frac{E_G(r^*,\rho^*)}{\gamma}, 400\sqrt{2}V_{K,G}(r^*,\rho^*)\right) \leq \max\left(\frac{B}{\gamma}\sqrt{\frac{6(\rho^*)^2}{N}\log\left(\sqrt{\frac{2}{\theta}}\frac{edr^*}{\rho^*}\right)}, D\sqrt{\frac{K}{N\theta}}r^*\right)$$

Now, one can check that r^* satisfies both of the two conditions below:

$$(3)\quad \frac{B}{\gamma}\sqrt{\frac{6(\rho^*)^2}{N}\log\left(\sqrt{\frac{2}{\theta}}\frac{edr^*}{\rho^*}\right)} \leq (r^*)^2 \quad \text{and} \quad (4)\quad D\sqrt{\frac{K}{N\theta}}r^* \leq (r^*)^2$$

Then, we have:

$$\max\left(\frac{E_G(r^*,\rho^*)}{\gamma}, 400\sqrt{2}V_{K,G}(r^*,\rho^*)\right) \leq (r^*)^2$$

which, according to Definition 22, allows us to conclude that $r^*_{\mathrm{RMOM,G}}(\gamma,\rho^*) \leq r^*$. Moreover, we have from (6.109) that $\rho^* \geq \sqrt{2/\theta}10kr^* \geq \sqrt{2/\theta}10kr^*_{\mathrm{RMOM,G}}(\gamma,\rho^*)$, that is, ρ^* satisfies the sparsity equation from Definition 23. This concludes the proof.

6.2.5.13 Proof of Theorem 50

The assumptions of Lemma 49 are met, which gives us the existence of two positive constants B and D such that, defining

$$\rho^* := \max\left(\frac{400\sqrt{3}Bk^2}{\gamma}\sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)}; 10Dk\sqrt{\frac{2K}{N\theta^2}}\right)$$

and

$$r^*(\gamma,\rho) := \max\left(\sqrt{v}\left(\frac{6}{N}\log\left(\frac{2B(ed)^2}{\gamma\theta\rho}\sqrt{\frac{6}{N}}\right)\right)^{1/4}; D\sqrt{\frac{K}{N\theta}}\right),$$

one has $r^*_{\mathrm{RMOM,G}}(\gamma,\rho^*) \leq r^*(\gamma,\rho^*)$ and ρ^* satisfies the sparsity equation from Definition 23. From Lemma 39, we get that Assumption 3.8 holds with $G : Z \in \mathbb{R}^{d\times d} \to (\theta/2)\|Z\|_2^2$ and $A = 1$ for any $\gamma > 0$, as a result of which the validity conditions of Theorem 24 are met. Then, fixing $\gamma = 1/32000$ and defining $\lambda = (11r^*(\gamma,2\rho^*))/(40\rho^*)$, it is true that with probability at least $1 - 2\exp(-72K/625)$,

$$\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_1 \leq 2\rho^*, \quad P\mathcal{L}_{\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}} \leq \frac{93}{100}(r^*(\gamma,2\rho^*))^2$$
$$\text{and} \quad \|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_2 \leq \sqrt{\frac{2}{\theta}}r^*(\gamma,2\rho^*). \tag{6.116}$$

Now, we can write:

$$\rho^* \leq D_1\frac{k}{\sqrt{N\theta^2}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right) \tag{6.117}$$

with $D_1 := \max(400\sqrt{3}B\gamma^{-1}, 10\sqrt{2}D)$. On the other hand, since $d \geq k$, we get that:

$$\rho^* \geq D_2k^2\sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)} \geq D_2\frac{k^2}{\sqrt{N\theta^2}},$$

where $D_2 := 400\sqrt{3}B\gamma^{-1}$. As a consequence, we have:

$$\log\left(\frac{B(ed)^2}{\gamma\theta\rho*}\sqrt{\frac{6}{N}}\right) \leq \log\left(\frac{B(ed)^2}{\gamma\theta}\sqrt{\frac{6}{N}}\frac{\sqrt{N\theta^2}}{D_2k^2}\right) = \log\left(\frac{1}{4\sqrt{5}}\left(\frac{ed}{k}\right)^2\right) \leq 2\log\left(\frac{ed}{k}\right),$$

so that:

$$
\begin{aligned}
r^*(\gamma, 2\rho^*) &\leq \max\left(\sqrt{\frac{2B\rho^*}{\gamma}}\left(\frac{12}{N}\log\left(\frac{ed}{k}\right)\right)^{1/4}; D\sqrt{\frac{K}{N\theta}}\right) \\
&\leq \max\left(\sqrt{\frac{2B}{\gamma}}\left(\frac{12}{N}\log\left(\frac{ed}{k}\right)\right)^{1/4}\frac{\sqrt{D_1 k}}{(N\theta^2)^{1/4}}\max\left(\sqrt{k}\log\left(\frac{ed}{k}\right)^{1/4}; K^{1/4}\right); D\sqrt{\frac{K}{N\theta}}\right) \\
&\leq \max\left(\sqrt{\frac{2BD_1}{\gamma N\theta}}12^{1/4}\max\left(\sqrt{k}\log\left(\frac{ed}{k}\right)^{1/4}; K^{1/4}\right)^2; D\sqrt{\frac{K}{N\theta}}\right) \\
&\leq \frac{C}{\sqrt{N\theta}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right), \qquad (6.118)
\end{aligned}
$$

where $C := \max\left(12^{1/4}\sqrt{2BD_1\gamma^{-1}}; D\right)$. Combining (6.116), (6.117) and (6.118), we finally get that, with probability at least $1 - 2\exp(-72K/625)$:

$$
\begin{aligned}
&\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_1 \leq 2D_1\frac{k}{\sqrt{N\theta^2}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right), \\
&\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_2 \leq \frac{\sqrt{2}C}{\sqrt{N\theta^2}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right) \\
&\text{and} \quad P\mathcal{L}_{\hat{Z}^{\mathrm{RMOM}}_{K,\lambda}} \leq \frac{93C^2}{100N\theta}\max\left(k^2\log\left(\frac{ed}{k}\right); K\right).
\end{aligned}
$$

T his concludes the proof.

6.2.5.14 Proof of Corollary 51

From Theorem 50, we get the existence of a universal constant $C_2 > 0$ such that with probability at least $1-\exp(-72K/625)$, $\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_2 \leq C_2(N\theta^2)^{-1/2}\max\left(k\sqrt{\log(ed/k)}; \sqrt{K}\right)$. Now, we can use Davis-Kahan sin-theta theorem (see Corollary 1 in Yu et al. (2014)) to get the existence of a universal constant $c_0 > 0$ such that $\sin(\Theta(\hat{\beta}, \beta^*)) = (1/\sqrt{2})\|\hat{\beta}\hat{\beta}^\top - \beta^*(\beta^*)^\top\|_2 \leq (c_0/g)\|\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*\|_2$ where $g := \lambda_1 - \lambda_2$ (λ_i being the i^{th} largest eigen value of Z^*) is the spectral gap of Z^*. Here, we know that $Z^* = \beta^*(\beta^*)^\top$ is rank one, with 1 as order one eigen value and 0 as order $d-1$ eigen value. Then we get $g = 1$, which leads us to the desired result, with $D = \sqrt{2}c_0 \times C_2$.

7.1 Distance metric learning: convexity of the constraint set

Here we show that the constraint set $\mathcal{C}$ of the ERM estimator of the distance metric learning problem presented in Section 2 is convex. We recall the definition of this set:

$$\mathcal{C} := \left\{ Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \sum_{i,j=1}^{N} \left\langle \left(Y_i - Y_j\right)\left(Y_i - Y_j\right)^\top, Z \right\rangle^{1/2} \geq 1 \right\}$$

where $(Y_i)_{i=1}^N$ are N given points in $\mathbb{R}^d$. Fot the sake of simplicity, we define, for $(i,j) \in [d]^2$, $V_{ij} = \left(Y_i - Y_j\right) \in \mathbb{R}^d$. Let Z_1 and Z_2 be two elements of $\mathcal{C}$, and consider $t \in [0,1]$. Let us show that $Z' = tZ_1 + (1-t)Z_2$ still belongs to $\mathcal{C}$. We have:

$$\begin{aligned}
\left(\sum_{i,j=1}^{N} \left\langle V_{ij}V_{ij}^\top, Z' \right\rangle^{1/2} \right)^2 &= \sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z' \right\rangle + \sum_{(i,j)\neq(p,q)} \left\langle V_{ij}V_{ij}^\top, Z' \right\rangle^{1/2} \left\langle V_{pq}V_{pq}^\top, Z' \right\rangle^{1/2} \\
&\geq \sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z' \right\rangle \\
&= t \sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z_1 \right\rangle + (1-t) \sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z_2 \right\rangle \\
&= t \left(\sqrt{\sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z_1 \right\rangle} \right)^2 + (1-t) \left(\sqrt{\sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z_2 \right\rangle} \right)^2 \\
&\geq t \left(\sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z_1 \right\rangle^{1/2} \right)^2 + (1-t) \left(\sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z_2 \right\rangle^{1/2} \right)^2 \\
&\geq t + (1-t) = 1
\end{aligned}$$

since each $\sum_{(i,j)\in[N]^2} \left\langle V_{ij}V_{ij}^\top, Z_\ell \right\rangle^{1/2}$, $\ell \in \{1,2\}$, is larger or equal to one, as $Z_\ell \in \mathcal{C}$. Then, $Z' \in \mathcal{C}$. We conclude that $\mathcal{C}$ is convex.

7.2 A property of local complexity fixed points

Let H be a Hilbert space and $\mathcal{C} \subset H$. We consider a linear loss function defined for all $Z \in \mathcal{C}$ by $\ell_Z : X \in H \to -\langle X, Z \rangle$ and its associated oracle over $\mathcal{C}$: $Z^* \in \operatorname{argmin}_{Z\in\mathcal{C}} P\ell_Z$. The excess loss function of $Z \in \mathcal{C}$ is defined as $\mathcal{L}_Z = \ell_Z - \ell_{Z^*}$. Let $\|\cdot\|$ be a norm defined (at least) over the span of $\mathcal{C}$. Let $G : H \to \mathbb{R}$ be a function. For all $\rho > 0$ and $r > 0$, we consider the localized model $\mathcal{C}_{\rho,r} = \{Z \in \mathcal{C} : \|Z - Z^*\| \leq \rho, G(Z - Z^*) \leq r\}$ with respect to a G localization and the associated Rademacher complexity

$$E(r,\rho) = \mathbb{E}\left[\sup_{Z\in\mathcal{C}_{\rho,r}} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathcal{L}_Z(X_i) \right| \right]$$

and variance term

$$V(r,\rho) = \sup_{Z\in\mathcal{C}_{\rho,r}} \sqrt{\mathrm{Var}(\mathcal{L}_Z)}.$$

Let θ and τ be two positive constants. We consider a local complexity fixed point: for all $\rho > 0$,

$$r^*(\rho) = \inf\left(r > 0 : \max\left(\theta E(r,\rho), \tau V(r,\rho)\right) \le r^2\right).$$

Proposition 7.1. We assume that $\mathcal{C}$ is star-shaped in Z^*. We assume that G is such that for all $\alpha \ge 1$ and all $Z \in \mathcal{C}, G(\alpha(Z - Z^*)) \ge \alpha G(Z - Z^*)$. Then, for all $\rho > 0$ and $b \ge 1$, we have $r^*(\rho) \le r^*(b\rho) \le \sqrt{b} r^*(\rho)$.

Proof. Let $\rho > 0$ and $b \ge 1$. For all $r > 0$, $\mathcal{C}_{\rho,r} \subset \mathcal{C}_{b\rho,r}$ and so $r^*(\rho) \le r^*(b\rho)$. Let us now prove the second inequality.

We start with some homogeneity property of the complexity and variance terms:

$$E(\sqrt{b}r, b\rho) \le bE(r,\rho) \text{ and } V(\sqrt{b}r, b\rho) \le bV(r,\rho). \tag{7.1}$$

We prove (7.1) for the complexity term, the proof for the variance term is identical. Let $Z \in \mathcal{C}_{b\rho,\sqrt{b}r}$ and define Z_0 such that $Z = Z^* + b(Z_0 - Z^*)$. Since $b \ge 1$ and $\mathcal{C}$ is star-shaped in Z^*, $Z_0 \in \mathcal{C}$. Moreover, $b\|Z_0 - Z^*\| = \|Z - Z^*\| \le b\rho$ and, by the property of G, $bG(Z_0 - Z^*) \le G(Z - Z^*) \le br^2$. We conclude that $Z_0 \in \mathcal{C}_{\rho,r}$. Moreover, by linearity of the loss function, we have $\mathcal{L}_Z = b\mathcal{L}_{Z_0}$. We deduce that

$$\sup_{Z\in\mathcal{C}_{b\rho,\sqrt{b}r}} \left|\frac{1}{N}\sum_{i=1}^N \sigma_i \mathcal{L}_Z(X_i)\right| \le b \sup_{Z_0\in\mathcal{C}_{\rho,r}} \left|\frac{1}{N}\sum_{i=1}^N \sigma_i \mathcal{L}_{Z_0}(X_i)\right| \tag{7.2}$$

and so (7.1) holds for the complexity term. It also holds for the variance using similar tools.

Next, it follows from (7.1) that

$$r^*(b\rho) = \inf\left(r > 0 : \max\left(\theta E(r,b\rho), \tau V(r,b\rho)\right) \le r^2\right) = \inf\left(r > 0 : \max\left(\theta E\left(\sqrt{b}\frac{r}{\sqrt{b}}, b\rho\right), \tau V\left(\sqrt{b}\frac{r}{\sqrt{b}}, b\rho\right)\right)\right.$$
$$\left. \le \inf\left(r > 0 : \max\left(\theta E\left(\frac{r}{\sqrt{b}}, \rho\right), \tau V\left(\frac{r}{\sqrt{b}}, \rho\right)\right) \le \left(\frac{r}{\sqrt{b}}\right)^2\right) \le \sqrt{b} r^*(\rho).\right.$$

■

7.3 A property of the sparsity equation

We consider the same setup as in Section 7.2 and define for all $\rho > 0$,

$$H_\rho = \left\{Z \in \mathcal{C} : \|Z - Z^*\| = \rho, G(Z - Z^*) \le (r^*(\rho))^2\right\}, \Gamma_{Z^*}(\rho) = \bigcup_{Z:\|Z-Z^*\|\le\rho/20} \partial\|\cdot\|(Z)$$

and $\Delta(\rho) = \inf_{Z\in H_\rho} \sup_{\Phi\in\Gamma_{Z^*}(\rho)} \langle \Phi, Z - Z^* \rangle$. In the previous section we said that ρ satisfies the sparsity equation when $\Delta(\rho) \ge c_0\rho$ where $0 < c_0 < 1$ is some absolute constant. In the following result we show that if ρ satisfies the sparsity equation then any number larger than ρ also satisfies this equation.

Proposition 7.2. We assume that $\mathcal{C}$ is star-shaped in Z^*. We assume that G is such that for all $\alpha \ge 1$ and all $Z \in \mathcal{C}, G(\alpha(Z - Z^*)) \ge \alpha G(Z - Z^*)$. Let $0 < c_0 < 1$. Then, for all $\rho > 0$ and $b \ge 1$, if ρ is such that $\Delta(\rho) \ge c_0\rho$ then $\Delta(b\rho) \ge c_0 b\rho$.

Proof. Let $\rho > 0$ be such that $\Delta(\rho) \geq c_0\rho$ and let $b \geq 1$. Let $Z \in H_{b\rho}$. Let us show that there exists $\Phi \in \Gamma_{Z^*}(b\rho)$ such that $\langle \Phi, Z - Z^* \rangle \geq c_0 b\rho$.

Let Z_0 be such that $Z = Z^* + b(Z_0 - Z^*)$. Since $b \geq 1$ and $\mathcal{C}$ is star-shaped in Z^*, $Z_0 \in \mathcal{C}$. Moreover, $b\|Z_0 - Z^*\| = \|Z - Z^*\| = b\rho$ and, using the property of G and Proposition 7.1, $bG(Z_0 - Z^*) \leq G(Z - Z^*) \leq (r^*(b\rho))^2 \leq b(r^*(\rho))^2$. Therefore, we have $Z_0 \in H_\rho$. But, since we assumed that $\Delta(\rho) \geq c_0\rho$, there exists $\Phi \in \Gamma_{Z^*}(\rho)$ such that $\langle \Phi, Z_0 - Z^* \rangle \geq c_0\rho$ and so $\langle \Phi, Z - Z^* \rangle \geq c_0 b\rho$. We conclude the proof by noting that $\Gamma_{Z^*}(\rho) \subset \Gamma_{Z^*}(b\rho)$ and so $\Phi \in \Gamma_{Z^*}(b\rho)$. ∎

7.4 Additional proofs for signed clustering

Proof of Proposition 6.3: control of $S_1(Z)$ adapted from Fei and Chen (2019b)

The noise matrix W is symmetric and has been decomposed as $W = \Psi + \Psi^\top$ where Ψ has been defined in (7.13). For all $Z \in \mathcal{C} \cap (Z^* + rB_1^{d\times d})$, we have

$$\begin{aligned}
S_1(Z) &= \langle \mathcal{P}(Z - Z^*), W \rangle = \langle \mathcal{P}(W), Z - Z^* \rangle \\
&= \langle UU^\top W, Z - Z^* \rangle + \langle WUU^\top, Z - Z^* \rangle - \langle UU^\top WUU^\top, Z - Z^* \rangle \\
&= 2\langle UU^\top W, Z - Z^* \rangle - \langle UU^\top WUU^\top, Z - Z^* \rangle \\
&= 2\langle UU^\top \Psi, Z - Z^* \rangle + 2\langle UU^\top \Psi^\top, Z - Z^* \rangle - \langle UU^\top(\Psi + \Psi^\top)UU^\top, Z - Z^* \rangle \\
&= 2\langle UU^\top \Psi, Z - Z^* \rangle + 2\langle UU^\top \Psi^\top, Z - Z^* \rangle - 2\langle UU^\top \Psi UU^\top, Z - Z^* \rangle \\
&= 2\langle UU^\top \Psi, Z - Z^* \rangle + 2\langle UU^\top \Psi^\top, Z - Z^* \rangle - 2\langle UU^\top \Psi, (Z - Z^*)UU^\top \rangle. \qquad (7.3)
\end{aligned}$$

An upper bound on $S_1(Z)$ follows from an upper bound on the three terms in the right side of (7.3). Let us show how to bound the first term. Similar arguments can be used to control the other two terms.

Consider $V := UU^T\Psi$. Let us find a high-probability upper bound on the term $\langle UU^\top \Psi, Z - Z^* \rangle = \langle V, Z - Z^* \rangle$ uniformly over $Z \in \mathcal{C} \cap (Z^* + rB_1^{d\times d})$. For all $k \in [K]$, $i \in \mathcal{C}_k$ and $j \in [d]$, we have

$$V_{ij} = \sum_{t=1}^{d} (UU^T)_{it}\Psi_{tj} = \sum_{t\in\mathcal{C}_k} \frac{1}{l_k}\Psi_{tj} = \frac{1}{l_k}\sum_{t\in\mathcal{C}_k}\Psi_{tj} = \frac{1}{l_k}\sum_{t\in\mathcal{C}_k}\Psi_{tj}.$$

Therefore, given $j \in [n]$ the V_{ij}'s are all equal for $i \in \mathcal{C}_k$. We can therefore fix some arbitrary index $i_k \in \mathcal{C}_k$ and have $V_{ij} = V_{i_k j}$ for all $i \in \mathcal{C}_k$. Moreover, $(V_{i_k j} : k \in [K], j \in [d])$ is a family of independent random variables. We now have

$$\begin{aligned}
\langle V, Z - Z^* \rangle &= \sum_{k\in[K]}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]} V_{ij}(Z - Z^*)_{ij} = \sum_{k\in[K]}\sum_{j\in[n]} l_k V_{i_k j}\sum_{i\in\mathcal{C}_k}\frac{(Z - Z^*)_{ij}}{l_k} \\
&= \sum_{k\in[K]}\sum_{j\in[n]} l_k V_{i_k j} w_{kj}
\end{aligned}$$

which is a weighted sum of dK independent centered random variables $X_{k,j} := l_k V_{i_k j}$ with weights $w_{k,j} = (1/l_k)\sum_{i\in\mathcal{C}_k}(Z - Z^*)_{ij}$ for $k \in [K], j \in [d]$. We now idenfity some properties on the weights w_{kj}.

The weights are such that

$$\sum_{k\in[K]}\sum_{j\in[d]} |w_{k,j}| \leq \frac{c_1 K}{d}\sum_{k\in[K]}\sum_{j\in[d]}\sum_{i\in\mathcal{C}_k}|(Z - Z^*)_{ij}| = \|Z - Z^*\|_1\frac{c_1 K}{d} \leq \frac{c_1 rK}{d}$$

which is equivalent to say that the weights vector $w = (w_{kj} : k \in [K], j \in [d])$ is in the ℓ_1^{Kd}-ball $(c_1 rK/d)B_1^{Kd}$. It is also in the unit ℓ_∞^{Kd}-ball since for all $k \in [K]$ and $j \in [d]$,

$$|w_{k,j}| \leq \sum_{i \in \mathcal{C}_k} \frac{|(Z - Z^*)_{ij}|}{l_k} \leq \|Z - Z^*\|_\infty \leq 1.$$

We therefore have $w \in B_\infty^{Kd} \cap (c_1 rK/d)B_1^{Kd}$ and so

$$\sup_{Z \in \mathcal{C} \cap (Z^* + rB_1^{d \times d})} \langle V, Z - Z^* \rangle \leq \sup_{w \in B_\infty^{Kd} \cap (c_1 rK/d)B_1^{Kd}} \sum_{k \in [K], j \in [d]} X_{k,j} w_{k,j}. \tag{7.4}$$

It remains to find a high-probability upper bound on the latter term. We can use the following lemma to that end.

Lemma 79. *Consider $X_{k,j} = \sum_{t \in \mathcal{C}_k} \Psi_{tj}$ for $(k,j) \in [K] \times [d]$. For all $0 \leq R \leq Kd$, if $\lceil R \rceil \geq 2eKd \exp(-(9/32)d\rho/K)$ then with probability at least $1 - \big(\lceil R \rceil/(2eKd)\big)^{\lceil R \rceil}$,*

$$\sup_{w \in B_\infty^{Kd} \cap RB_1^{Kd}} \sum_{(k,j) \in [K] \times [d]} X_{k,j} w_{k,j} \leq 4\sqrt{8c_0} R \sqrt{\frac{d\rho}{K} \log\left(\frac{2eKd}{\lceil R \rceil}\right)}.$$

Proof of Lemma 79. Consider $N = Kd$ and assume that $1 \leq R \leq N$. We denote by $X_1^* \geq X_2^* \geq \cdots, \geq X_N^*$ (resp. $w_1^* \geq \cdots \geq w_N^*$) the non-decreasing rearrangement of $|X_{k,j}|$ (resp. $|w_{k,j}|$) for $(k,j) \in [K] \times [d]$. We have

$$\begin{aligned}
&\sup_{w \in B_\infty^N \cap RB_1^N} \sum_{(k,j) \in [K] \times [d]} X_{k,j} w_{k,j} \leq \sup_{w \in B_\infty^N \cap RB_1^N} \sum_{i=1}^N X_i^* w_i^* \\
&\leq \sup_{w \in B_\infty^N} \sum_{i=1}^{\lceil R \rceil} X_i^* w_i^* + \sup_{w \in RB_1^N} \sum_{i=\lceil R \rceil + 1}^N X_i^* w_i^* \leq \sum_{i=1}^{\lceil R \rceil} X_i^* + R X_{\lceil R \rceil + 1}^* \leq 2 \sum_{i=1}^{\lceil R \rceil} X_i^*.
\end{aligned}$$

Moreover, for all $\tau > 0$, using a union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lceil R \rceil} X_i^* > \tau\right) = \mathbb{P}\left(\exists I \subset [K] \times [d] : |I| = \lceil R \rceil \text{ and } \sum_{(k,j) \in I} |X_{k,j}| > \tau\right) \\
&= \mathbb{P}\left(\max_{I \subset [K] \times [d] : |I| = \lceil R \rceil} \max_{u_{k,j} = \pm 1, (k,j) \in I} \sum_{(k,j) \in I} u_{k,j} X_{k,j} > \tau\right) \\
&\leq \sum_{I \subset [K] \times [d] : |I| = \lceil R \rceil} \sum_{u \in \{\pm 1\}^{\lceil R \rceil}} \mathbb{P}\left(\sum_{(k,j) \in I} X_{k,j} u_{k,j} > \tau\right) \\
&= \sum_{I \subset [K] \times [d] : |I| = \lceil R \rceil} \sum_{u \in \{\pm 1\}^{\lceil R \rceil}} \mathbb{P}\left(\sum_{(k,j) \in I} \sum_{t \in \mathcal{C}_k} \Psi_{t,j} u_{k,j} > \tau\right).
\end{aligned}$$

Let us now control each term of the latter sum thanks to Bernstein inequality. The random variables $(\Psi_{t,j} : t, j \in [d])$ are independent with variances at most $\rho = \delta \max(1 - \delta(2p-1)^2, 1 - \delta(2q-1)^2)$ since $\mathrm{Var}(\Psi_{ij}) = 0$ when $i > j$ and $\mathrm{Var}(\Psi_{ij}) = \mathrm{Var}(A_{ij} - \mathbb{E}[A_{ij}]) = \mathrm{Var}(A_{ij}) \leq \rho$ for $j \geq i$. Moreover, $|\Psi_{ij}| = 0$ when $j < i$ and $|\Psi_{ij}| = |W_{ij}| = |A_{ij} - \mathbb{E}A_{ij}| \leq 2$ for $j \geq i$ because $A_{ij} \in \{-1, 0, 1\}$. It follows from Bernstein's inequality that for all $I \subset [K] \times [d]$ satisfying $|I| = \lceil R \rceil$ and $u \in \{\pm 1\}^{\lceil R \rceil}$ that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{(k,j) \in I} \sum_{t \in \mathcal{C}_k} \Psi_{t,j} u_{k,j} > \tau\right) \leq \exp\left(\frac{-\tau^2}{2\lceil R \rceil l_k \rho + 4\tau/3}\right) \leq \exp\left(\frac{-\tau^2}{2\lceil R \rceil c_0 d\rho/K + 4\tau/3}\right) \\
&\leq \exp\left(\frac{-\tau^2}{4\lceil R \rceil c_0 d\rho/K}\right)
\end{aligned}$$

when $\tau \leq (3/2)\lceil R\rceil c_0 d\rho/K$. Therefore, $\sup_{w\in B_\infty^N \cap RB_1^N} \sum X_{k,j}w_{k,j} \leq 2\tau$ with probability at least

$$1-\binom{N}{\lceil R\rceil}2^{\lceil R\rceil}\exp\left(\frac{-\tau^2}{4\lceil R\rceil c_0 d\rho/K}\right) \geq 1-\exp\left(\frac{-\tau^2}{8\lceil R\rceil c_0 d\rho/K}\right)$$

when

$$(3/2)\lceil R\rceil c_0 d\rho/K \geq \tau \geq \sqrt{8c_0}\lceil R\rceil\sqrt{\frac{d\rho}{K}\log\left(\frac{2eN}{\lceil R\rceil}\right)}$$

which is a non vacuous condition since $\lceil R\rceil \geq 2eN\exp(-(9/32)d\rho/K)$. The result follows, in the case $1 \leq R \leq N$, by taking $\tau = \sqrt{8c_0}\lceil R\rceil\sqrt{(d\rho/K)\log\left(2eN/\lceil R\rceil\right)}$ and using that $2R \geq \lceil R\rceil$ when $R \geq 1$.

For $0 \leq R \leq 1$, we have

$$\sup_{w\in B_\infty^{Kd}\cap RB_1^{Kd}} \sum_{(k,j)\in[K]\times[d]} X_{k,j}w_{k,j} = R\max_{(k,j)\in[K]\times[d]}|X_{k,j}|$$

and using Bernstein inequality as above we get that with probability at least $1-\exp(-K\tau^2/(8c_0 d\rho))$, $\max_{(k,j)\in[K]\times[d]}|X_{k,j}| \leq \tau$ when $3c_0 d\rho/(2K) \geq \tau \geq \sqrt{8c_0 d\rho\log(dK)/K}$ which is a non vacuous condition when $9c_0 d\rho \geq dK\log(dK)$. By taking $\tau = \sqrt{8c_0 d\rho\log(dK)/K}$, we obtain, that for all $0 \leq R \leq 1$, if $9c_0 d\rho \geq 4K\log(dK)$ then with probability at least $1-1/(dK)$,

$$\sup_{w\in B_\infty^{Kd}\cap RB_1^{Kd}} \sum_{(k,j)\in[K]\times[d]} X_{k,j}w_{k,j} \leq R\sqrt{\frac{8c_0 d\rho\log(dK)}{K}}.$$

■

We apply Lemma 79 for $R = c_1 rK/d$ to control (7.4):

$$\mathbb{P}\left[\sup_{Z\in\mathcal{C}\cap(Z^*+rB_1^{d\times d})}\left\langle V, Z-Z^*\right\rangle \leq c_2 r\sqrt{\frac{K\rho}{d}\log\left(\frac{2eKd}{\lceil\frac{c_1 rK}{d}\rceil}\right)}\right] \geq 1-\left(\frac{\lceil\frac{c_1 rK}{d}\rceil}{2eKd}\right)^{\lceil\frac{c_1 rK}{d}\rceil}$$

when $\lceil c_1 rK/d\rceil \geq 2eKd\exp(-(9/32)d\rho/K)$.

Using the same methodology, we can prove exactly the same result for the quantity

$$\sup_{Z\in\mathcal{C}\cap(Z^*+rB_1^{d\times d})}\left\langle UU^\top\Psi^\top, Z-Z^*\right\rangle.$$

We can also use the same method to upper bound $\sup_{Z\in\mathcal{C}\cap(Z^*+rB_1^{d\times d})}\langle UU^\top\Psi^\top, (Z-Z^*)UU^\top\rangle$, we simply have to check that the weights vector $w' = (w'_{kj} : k\in[K], j\in[d])$ where $w'_{kj} = (1/l_k)\sum_{i\in\mathcal{C}_k}[(Z-Z^*)UU^\top]_{ij}$ is also in $B_\infty^{Kd}\cap(c_1 rK/d)B_1^{Kd}$ for any $Z\in\mathcal{C}$ such that $\|Z-Z^*\|_1 \leq r$. This is indeed the case, since we have for all $i\in[d]$, $k'\in[K]$ and $j\in\mathcal{C}_{k'}$, $[(Z-Z^*)UU^\top]_{ij} = \sum_{p=1}^d (Z-Z^*)_{ip}(UU^\top)_{pj} = \sum_{p\in\mathcal{C}_{k'}}(Z-Z^*)_{ip}/l_{k'}$ which is therefore constant for all elements in $j\in\mathcal{C}_{k'}$. Therefore, we have

$$\sum_{k\in[K]}\sum_{j\in[d]}|w'_{kj}| = \sum_{k\in[K]}\sum_{k'\in[K]}\sum_{j\in\mathcal{C}_{k'}}\left|\frac{1}{l_k l_{k'}}\sum_{i\in\mathcal{C}_k}\sum_{p\in\mathcal{C}_{k'}}(Z-Z^*)_{ip}\right|$$
$$\leq \sum_{k\in[K]}\sum_{k'\in[K]}\sum_{j\in\mathcal{C}_{k'}}\frac{1}{l_k l_{k'}}\sum_{i\in\mathcal{C}_k}\sum_{p\in\mathcal{C}_{k'}}\left|(Z-Z^*)_{ip}\right| \leq \|Z-Z^*\|_1\frac{c_1 K}{d} \leq \frac{c_1 rK}{d}$$

and for all $k' \in [K]$ and $j \in \mathcal{C}_{k'}$,

$$|w'_{kj}| = \left| \frac{1}{l_k l_{k'}} \sum_{i \in \mathcal{C}_k} \sum_{p \in \mathcal{C}_{k'}} (Z - Z^*)_{ip} \right| \leq \|Z - Z^*\|_\infty \leq 1.$$

Therefore, $w' \in B_\infty^{Kd} \cap (c_1 rK/d) B_1^{Kd}$ and we obtain exactly the same upper bound for the three terms in (7.3). This concludes the proof of Proposition 6.3.

Proof of Proposition 6.4: control of the $S_2(Z)$ term from Fei and Chen (2019b)

In this section, we prove Proposition 6.4. We follow the proof from Fei and Chen (2019b) but we only consider the "dense case" which is when $d\delta\nu \geq \log d$ – we recall that $\nu = \max(2p - 1, 1 - 2q)$. For a similar uniform control of $S_2(Z)$ in the "sparse case ", when $c_0 \leq d\delta\nu \leq \log d$ for some absolute constant c_0, we refer the reader to Fei and Chen (2019b).

For all $Z \in \mathcal{C}$, we have $S_2(Z) = \langle \mathcal{P}^\perp(Z - Z^*), W \rangle = \langle \mathcal{P}^\perp(Z), W \rangle$ because, by construction of the projection operator, $\mathcal{P}^\perp(Z^*) = 0$. Therefore, $S_2(Z) \leq \|\mathcal{P}^\perp(Z)\|_* \|W\|_{\text{op}}$ where $\|\cdot\|_*$ denotes the nuclear norm (i.e. the sum of singular values) and $\|\cdot\|_{\text{op}}$ denotes the operator norm (i.e. maximum of the singular value). In the following Lemma 80, we prove an upper bound for $\|\mathcal{P}^\perp(Z)\|_*$ and then, we will obtain a high-probability upper bound onto $\|W\|_{\text{op}}$.

Lemma 80. *For all $Z \in \mathcal{C} \cap (Z^* + rB_1^{d\times d})$, we have*

$$\|\mathcal{P}^\perp(Z)\|_* = \text{Tr}(\mathcal{P}^\perp(Z)) \leq \frac{c_1 k}{d} \|Z - Z^*\|_1 \leq \frac{c_1 Kr}{d}.$$

Proof. Since $Z \succeq 0$ so it is for $(\mathbb{I}_d - UU^\top) Z (\mathbb{I}_d - UU^\top)$ and so $\mathcal{P}^\perp(Z) = (\mathbb{I}_d - UU^\top) Z (\mathbb{I}_d - UU^\top) \succeq 0$ therefore $\|\mathcal{P}^\perp(Z)\|_* = \text{Tr}(\mathcal{P}^\perp(Z))$. Next, we bound the trace of $\mathcal{P}^\perp(Z)$.

Since $\mathbb{I}_d - UU^\top$ is a projection operator, it is symmetric and $(\mathbb{I}_d - UU^\top)^2 = \mathbb{I}_d - UU^\top$, moreover, $\text{Tr}(Z) = d = \text{Tr}(Z^*)$ when $Z \in \mathcal{C}$ so

$$\begin{aligned}
&\text{Tr}(\mathcal{P}^\perp(Z)) = \text{Tr}(\mathcal{P}^\perp(Z - Z^*)) = \text{Tr}((\mathbb{I}_d - UU^\top)(Z - Z^*)(\mathbb{I}_d - UU^\top)) \\
&= \text{Tr}((\mathbb{I}_d - UU^T)^2 (Z - Z^*)) = \text{Tr}((\mathbb{I}_d - UU^T)(Z - Z^*)) \\
&= \text{Tr}(Z) - \text{Tr}(Z^*) + \text{Tr}(UU^T(Z^* - Z)) = \sum_{i,j} (UU^T)_{ij} (Z^* - Z)_{ij} \\
&= \sum_{k \in [K]} \sum_{i,j \in \mathcal{C}_k} \frac{1}{l_k} (Z^* - Z)_{ij} \overset{\text{(i)}}{=} \sum_{k \in [K]} \frac{1}{l_k} \sum_{i,j \in \mathcal{C}_k} |(Z^* - Z)_{ij}| \\
&\leq \frac{c_1 K}{d} \sum_{k \in [K]} \sum_{i,j \in \mathcal{C}_k} |(Z^* - Z)_{ij}| \leq \frac{c_1 K}{d} \|Z - Z^*\|_1
\end{aligned}$$

where we used in *(i)* that for i and j in a same community, we have $Z^*_{ij} = 1$ and $Z_{ij} \in [0, 1]$, thus $(Z^* - Z)_{ij} = |(Z^* - Z)_{ij}|$. Finally, when Z is in the localized set $\mathcal{C} \cap (Z^* + rB_1^{d\times d})$, we have $\|Z - Z^*\|_1 \leq r$ which concludes the proof. □

Now, we obtain a high-probability upper bound on $\|W\|_{\text{op}}$. In the following, we apply this result in the "dense case" (i.e. $n\delta\nu \geq \log d$) to get the uniform bound onto $S_2(Z)$ over $Z \in \mathcal{C} \cap (Z^* + rB_1^{d\times d})$.

Lemma 81 (Lemma 4 in Fei and Chen (2019b))**.** *With probability at least $1 - \exp(-\delta\nu d)$, $\|W\|_{op} \leq 16\sqrt{\delta\nu d} + 168\sqrt{\log(d)}$.*

Proof. Let A' be an independent copy of A and $R \in \mathbb{R}^{d\times d}$ be a random symmetric matrix independent from both A and A' whose sub-diagonal entries are independent Rademacher

random variables. Using a symmetrization argument (see Chapter 2 in Koltchinskii (2011b) or Chapter 2.3 in van der Vaart and Wellner (1996)), we obtain for $W = A - \mathbb{E}A$,

$$\mathbb{E}\|W\|_{\text{op}} = \mathbb{E}\|A - \mathbb{E}A'\|_{\text{op}} \overset{\text{(i)}}{\leq} \mathbb{E}\|A - A'\|_{\text{op}} \overset{\text{(ii)}}{=} \mathbb{E}\|(A - A') \circ R\|_{\text{op}} \overset{\text{(iii)}}{\leq} 2\mathbb{E}\|A \circ R\|_{\text{op}}$$

where $\circ$ is the entry-wise matrix multiplication operation, *(i)* comes from Jensen's inequality, *(ii)* occurs since $A - A'$ and $(A - A') \circ R$ are identically distributed and *(iii)* is the triangle inequality. Next, we obtain an upper bound onto $\mathbb{E}\|A \circ R\|_{\text{op}}$.

We define the family of independent random variables $(\xi_{ij} : 1 \leq i \leq j \leq d)$ where for all $1 \leq i \leq j \leq d$

$$\xi_{ij} = \begin{cases} \frac{1}{\sqrt{|\mathbb{E}A_{ij}|}} \text{ with probability } \frac{\mathbb{E}A_{ij}}{2} \\ -\frac{1}{\sqrt{|\mathbb{E}A_{ij}|}} \text{ with probability } \frac{\mathbb{E}A_{ij}}{2} \\ 0 \text{ with probability } 1 - \mathbb{E}A_{ij}. \end{cases} \tag{7.5}$$

We also put $b_{ij} := \sqrt{|\mathbb{E}A_{ij}|}$ for all $1 \leq i \leq j \leq d$. It is straightforward to see that $(\xi_{ij}b_{ij} : 1 \leq i \leq j \leq d)$ and $(A_{ij}R_{ij} : 1 \leq i \leq j \leq d)$ have the same distribution. As a consequence, $\|A \circ R\|_{\text{op}}$ and $\|X\|_{\text{op}}$ have the same distribution where $X \in \mathbb{R}^{d\times d}$ is a symmetric matrix with $X_{ij} = \xi_{ij}b_{ij}$ for $1 \leq i \leq j \leq d$. An upper bound on $\mathbb{E}\|X\|_{\text{op}}$ follow from the next result due to A. S. Bandeira and Van Handel (2016).

Theorem 82 (Corollary 3.6 in A. S. Bandeira and Van Handel (2016)). *Let $\xi_{ij}, 1 \leq i \leq j \leq d$ be independent symmetric random variables with unit variance and $(b_{ij}, 1 \leq i \leq j \leq d)$ be a family of real numbers. Let $X \in \mathbb{R}^{d\times d}$ be the random symmetric matrix defined by $X_{ij} = \xi_{ij}b_{ij}$ for all $1 \leq i \leq j \leq d$. Consider $\sigma := \max_{1\leq i\leq d}\left\{\sqrt{\sum_{j=1}^{d} b_{ij}^2}\right\}$. Then, for any $\alpha \geq 3$,*

$$\mathbb{E}\|X\|_{op} \leq e^{\frac{2}{\alpha}}\left[2\sigma + 14\alpha \max_{1\leq i\leq j\leq d}\left\{\|\xi_{ij}b_{ij}\|_{2\lceil\alpha\log(d)\rceil}\right\}\sqrt{\log(d)}\right]$$

where, for any $q > 0$, $\|\cdot\|_q$ denotes the L_q-norm.

Since $(\xi_{ij} : 1 \leq i \leq j \leq d)$ are independent symmetric such that $\text{Var}(\xi_{ij}) = \mathbb{E}[\xi_{ij}^2] = 1$ we can apply Lemma 82 to upper bound $\mathbb{E}\|X\|_{\text{op}} = \mathbb{E}\|A \circ R\|_{\text{op}}$. We have $\|\xi_{ij}b_{ij}\|_{2\lceil\alpha\log(d)\rceil} \leq 1$ for any $\alpha \geq 3$ and $b_{ij}^2 = |\mathbb{E}A_{ij}| \leq \delta\nu$. It therefore follows from Lemma 82 for $\alpha = 3$ that

$$\mathbb{E}\|W\|_{\text{op}} \leq 2e^{\frac{2}{3}}\left[2\sqrt{d\delta\nu} + 42\sqrt{\log(d)}\right] \leq 8\sqrt{d\delta\nu} + 168\sqrt{\log(d)}. \tag{7.6}$$

The final step to prove Lemma 81 is a concentration argument showing that $\|W\|_{\text{op}}$ is close to its expectation with high-probability. To that end we rely on a general result for Lipschitz and separately convex functions from Boucheron, Lugosi, and Massart (2013b). We first recall that a real-valued function f of N variables is said separately convex when for every $i = 1, \ldots, N$ it is a convex function of the i-th variable if the rest of the variables are fixed.

Theorem 83 (Theorem 6.10 in Boucheron et al. (2013b)). *Let $\mathcal{X}$ be a convex compact set in $\mathbb{R}$ with diameter B. Let $X_1, \cdots, X_N$ be independent random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^N \to \mathbb{R}$ be a separately convex and 1-Lipschitz function, with respect to the ℓ_2^N-norm (i.e. $|f(x) - f(y)| \leq \|x - y\|_2$ for all $x, y \in \mathcal{X}^N$). Then $Z = f(X_1, \ldots, X_N)$ satisfies, for all $t > 0$, with probability at least $1 - \exp(-t^2/B^2)$, $Z \leq \mathbb{E}[Z] + t$.*

We apply Theorem 83 to $Z := \|W\|_{\text{op}} = f(A_{ij} - \mathbb{E}A_{ij}, 1 \leq i \leq j \leq d) = \frac{1}{\sqrt{2}}\|A - \mathbb{E}A\|_{\text{op}}$ where f is a 1-Lipschitz with respect to ℓ_2^N-norm for $N = d(d-1)/2$ and separately convex function

and $(A_{ij} - \mathbb{E}A_{ij}, 1 \le i \le j \le d)$ is a family of N independent random variables. Moreover, for each $i \ge j$, $(A - \mathbb{E}A)_{ij} \in [-1 - \delta(2p-1), 1 - \delta(2q-1)]$, which is a convex compact set with diameter $B = 2(1 + \delta(p-q)) \le 4$. Therefore, it follows from Theorem 83 that for all $t > 0$, with probability at least $1 - \exp(-t^2/16)$, $\|W\|_{\mathrm{op}} \le \mathbb{E}\|W\|_{\mathrm{op}} + \sqrt{2}t$. In particular, we finish the proof of Lemma 81 for $t = 4\sqrt{\delta\nu d}$ and using the bound from (7.6). □

It follows from Lemma 81 that when $n\nu\delta \ge \log d$, $\|W\|_{\mathrm{op}} \le 184\sqrt{d\delta\nu}$ with probability at least $1 - \exp(-\delta\nu d)$. Using this later result together with Lemma 80 concludes the proof of Proposition 6.4.

7.5 Proof of Theorem 34 for Angular Synchronization with additive noise

Here we consider the following model: we observe $C = Z^* + \sigma W$, where $Z^* = x^*(\overline{x^*})^\top$, $x^* = (e^{\iota\theta_i})_{i=1}^d$ and $W \in \mathbb{C}^{d\times d}$ is a complex Wigner matrix (i.e. $W = \overline{W}^T$, its above-diagonal entries are complex numbers whose real and imaginary parts are independent normally distributed random variables with mean zero and variance $1/2$, and its diagonal entries are zero).

Let us first show that Z^* is the oracle in our approach. That is to show that

$$Z^* \in \operatorname*{argmax}_{Z\in\mathcal{C}} \langle \mathbb{E}C, Z\rangle \tag{7.7}$$

where $\mathcal{C} := \{Z \in \mathbb{H}_d : Z \succeq 0, \mathrm{diag}(Z) = \mathbf{1}_d\}$.

We recall that the offsets are $\delta_{ij} = (\theta_i - \theta_j)[2\pi]$ for $i, j = 1, \dots, d$. Consider $\gamma_1, \dots, \gamma_d \in [0, 2\pi[$ and define $x_i = e^{\iota\gamma_i}, i = 1, \dots, d$. We have for all $i, j = 1, \dots, d$:

$$\gamma_i - \gamma_j = \delta_{ij}[2\pi] \iff e^{\iota\gamma_i} - e^{\iota\delta_{ij}}e^{\iota\gamma_j} = 0 \iff Z^*_{ij}x_j - x_i \iff \mathbb{E}C_{ij}x_j - x_i = 0$$

It follows that

$$\operatorname*{argmin}_{x\in\mathbb{C}^d:|x_i|=1\forall i} \left\{ \sum_{i,j=1}^d |\mathbb{E}C_{ij}x_j - x_i|^2 \right\} = \left\{ (e^{\iota(\theta_i+\theta_0)})_{i=1}^d : \theta_0 \in [0, 2\pi) \right\}. \tag{7.8}$$

Let us now rewrite the latter optimization problem as a SDP problem.

Let $x \in \mathbb{C}^d$ be such that $|x_i| = 1$ for all $i = 1, \dots, d$. We have

$$\begin{aligned}
&\sum_{i,j=1}^d |\mathbb{E}C_{ij}x_j - x_i|^2 = \sum_{i,j=1}^d (\mathbb{E}C_{ij}x_j - x_i)\overline{(\mathbb{E}C_{ij}x_j - x_i)} \\
&= \sum_{i,j=1}^d \left[\underbrace{|\mathbb{E}C_{ij}x_j|^2}_{=1} + \underbrace{|x_i|^2}_{=1} - (\mathbb{E}C_{ij}x_j + \overline{\mathbb{E}C_{ij}x_j\bar{x}_i}) \right] = 2d^2 - 2\Re\left(\sum_{i,j=1}^d \mathbb{E}C_{ij}x_j\bar{x}_i \right) \\
&= 2d^2 - 2\Re\left(\overline{x}^\top \mathbb{E}C x \right) = 2d^2 - \overline{x}^\top \mathbb{E}C x,
\end{aligned}$$

where we used in the last inequality that $\mathbb{E}[C] = Z^* = x^*\overline{x^*}^\top$ so that $\overline{x}^\top\mathbb{E}Cx = |\langle x, x^*\rangle|^2 \in \mathbb{R}$. Next, we see that $\overline{x}^\top\mathbb{E}Cx = \langle \mathbb{E}C, X\rangle$ where $X = x\overline{x}^\top$, hence, minimizing $x \in \mathbb{C}^d \to \sum_{i,j=1}^d |\mathbb{E}[C]_{ij}x_j - x_i|^2$ over all $x \in \mathbb{C}^d$ such that $|x_i| = 1$ boils down to maximizing $X \in \mathbb{C}^{d\times d} \to \mathrm{tr}(\mathbb{E}[C]X) \in \mathbb{R}$ over the set $\{X = x\overline{x}^\top : x \in \mathbb{C}^d, |x_i| = 1\forall i\} = \{X \in \mathbb{H}_d : X \succeq 0, \mathrm{diag}(X) = \mathbf{1}_d, \mathrm{rank}(X) = 1\}$ where $\mathbb{H}_d$ is the set of hermitian matrices of size $d \times d$. Then, it follows from (7.8) that

$$\operatorname*{argmax}_{X\in\mathbb{H}_d:X\succeq 0,\mathrm{diag}(X)=\mathbf{1}_d,\mathrm{rank}(X)=1} \langle \mathbb{E}C, X\rangle = \{Z^*\}$$

because for all $\theta_0 \in [0, 2\pi)$, $(e^{\iota(\theta_i+\theta_0)})_{i=1}^d \overline{(e^{\iota(\theta_i+\theta_0)})_{i=1}^d}^\top = Z^*$. The latter inequality is almost the result that we want to prove in (7.7); it only remains to show that the rank one constraint may be dropped. We use the same approach as the one we used to show (4.15).

First, one can easily check that the following inclusion holds true

$$\mathcal{C} \subset \mathcal{C}' := \{Z \in \mathbb{C}^{d\times d} : |Z_{ij}| \leq 1, \forall i, j \in [d]\}.$$

Second, the objective function $Z \to \Re(\langle \mathbb{E}C, Z\rangle)$ is linear (with respect to real coefficient), hence, maximizing it over the convex set $\mathcal{C}'$ yields a solution at an extreme point of $\mathcal{C}'$, that is in the set of matrices $Z \in \mathbb{C}^{d\times d}$ such that $|Z_{ij}| = 1$ for all (i, j). Let $X = (e^{\iota\beta_{ij}})_{i,j\leq n}$ with $0 \leq \beta_{ij} < 2\pi$ be an extreme point of $\mathcal{C}'$. We have

$$\Re\Big(\langle \mathbb{E}C, X\rangle\Big) = \Re\left(\sum_{i,j=1}^d \mathbb{E}C_{ij}\overline{X_{ij}}\right) = \Re\left(\sum_{i,j=1}^d e^{\iota\delta_{ij}}e^{-\iota\beta_{ij}}\right) = \sum_{i,j=1}^d \cos(\delta_{ij} - \beta_{ij}) \leq d^2$$

and the maximum is achieved when $\beta_{ij} = \delta_{ij}[2\pi]$ for all (i, j), that is when $X = Z^*$. Since $Z^* \in \mathcal{C} \subset \mathcal{C}'$, then Z^* is the unique maximizer of $Z \to \langle \mathbb{E}C, Z\rangle$ over $\mathcal{C}$. Therefore (7.7) holds and so Z^* is also the oracle in this model.

7.5.1 Curvature of the objective function

Consider $Z \in \mathcal{C}$. We can write $Z = (x_{ij}e^{\iota\beta_{ij}})_{i,j=1,\dots,d}$, with $0 \leq x_{ij} \leq 1$ (we recall that $\mathcal{C} \subset \mathcal{C}'$ defined above) and $\beta_{ij} \in [0, 2\pi)$. On one side, we have

$$\begin{aligned}
\langle \mathbb{E}C, Z^* - Z\rangle = \Re\Big(\langle \mathbb{E}C, Z^* - Z\rangle\Big) &= \Re\left(\sum_{i,j=1}^d \mathbb{E}C_{ij}\overline{(Z^* - Z)_{ij}}\right) \\
&= \Re\left(\sum_{i,j=1}^d e^{\iota\delta_{ij}}(e^{-\iota\delta_{ij}} - x_{ij}e^{-\iota\beta_{ij}})\right) = \Re\left(\sum_{i,j=1}^d (1 - x_{ij}e^{\iota(\delta_{ij}-\beta_{ij})})\right) \\
&= \sum_{i,j=1}^d (1 - x_{ij}\cos(\delta_{ij} - \beta_{ij})).
\end{aligned}$$

On the other side, we have

$$\begin{aligned}
&\|Z^* - Z\|_2^2 = \sum_{i,j=1}^d |(Z^* - Z)_{ij}|^2 = \sum_{i,j=1}^d |e^{\iota\delta_{ij}} - x_{ij}e^{\iota\beta_{ij}}|^2 = \sum_{i,j=1}^d |1 - x_{ij}e^{\iota(-\delta_{ij}+\beta_{ij})}|^2 \\
&= \sum_{i,j=1}^d (1 - x_{ij}\cos(\beta_{ij} - \delta_{ij}))^2 + x_{ij}^2\sin^2(\beta_{ij} - \delta_{ij}) = \sum_{i,j=1}^d (1 - 2x_{ij}\cos(\beta_{ij} - \delta_{ij}) + x_{ij}^2) \\
&\leq 2\sum_{i,j=1}^d (1 - x_{ij}\cos(\beta_{ij} - \delta_{ij})) = 2\langle \mathbb{E}C, Z^* - Z\rangle
\end{aligned}$$

where we used in the last but one inequality that $0 \leq x_{ij} \leq 1$.

We conclude that the excess risk satisfies the following curvature: for all $Z \in \mathcal{C}$,

$$\langle \mathbb{E}C, Z^* - Z\rangle \geq \frac{1}{2}\|X^* - X\|_2^2. \tag{7.9}$$

That is Assumption 3.1 holds with $G : M \in \mathbb{C}^{d\times d} \to (1/2)\|M\|_2^2$.

7.5.2 Three upper bounds on the fixed point $r_G^*(\Delta)$ in the angular group synchronization model with additive noise

According to our methodology (associated with Theorem 5), we need to calculate the following fixed point: let $\Delta \in (0,1)$, and consider

$$r_G^*(\Delta) = \inf\left\{ r > 0 : \mathbb{P}\left(\sup_{Z\in\mathcal{C}: G(Z^*-Z)\le r} \langle C - \mathbb{E}C, Z - Z^*\rangle \le r/2 \right) \ge 1-\Delta \right\}$$
$$= \inf\left\{ r > 0 : \mathbb{P}\left(\sup_{Z\in\mathcal{C}_r} \langle \sigma W, Z - Z^*\rangle \le r/2 \right) \ge 1-\Delta \right\}$$

where $\mathcal{C}_r := \left\{ Z \in \mathcal{C} : \|Z - Z^*\|_2 \le \sqrt{2r} \right\}$.

For pedagogical purposes, we show how to perform this computation via three different means, yielding three different results. We obtain the following upper bounds

$$r_G^*(\exp(-d^2/2)) \le 32\sigma^2 d^2, r_G^*(\exp(-d/2)) \le 36 K_G^{\mathbb{C}} \sigma d^{3/2} \text{ and } r_G^*(5\exp(-d/2)) \le 2(20\sigma)^2 d.$$

Each of the three bounds follows from a different strategy. The first one is based on the inclusion $\mathcal{C}_r \subset Z^* + \sqrt{2r} B_2^{d\times d}$, the second one on $\mathcal{C}_r \subset \mathcal{C}$ and is therefore the approach that we called "global", and the last one follows from the strategy used in (Fei & Chen, 2019b) that we already used for the signed clustering problem.

7.5.2.1 First upper bound on the fixed point r_G^* using $\mathcal{C}_r \subset Z^* + \sqrt{2r} B_2^{d\times d}$

In this section, we use the following inclusion to obtain the result

$$\mathcal{C}_r \subset Z^* + \sqrt{2r} B_2^{d\times d}, \tag{7.10}$$

where $B_2^{d\times d}$ is the Euclidean ball in $\mathbb{R}^{d\times d}$. We have

$$\sup_{Z\in\mathcal{C}:\|Z-Z^*\|_2\le\sqrt{2r}} \langle \sigma W, Z - Z^*\rangle \le \sigma \sup_{Z\in\sqrt{2r}B_2^{d\times d}} \langle W, Z\rangle = \sigma\sqrt{2r}\|W\|_2.$$

Next, we use Borell's concentration inequality for Gaussian processes (Ledoux, 2001): for all $u > 0$, with probability at least $1 - \exp(-u^2/2)$, we have that

$$\|W\|_2 \le \mathbb{E}\|W\|_2 + \sup_{Z\in B_2^{d\times d}} \sqrt{\mathbb{E}\langle W, Z\rangle^2}\, u \le d + u.$$

As a consequence, for $\Delta = \exp(-d^2/2)$ and $r = 32\sigma^2 d^2$, we have, with probability at least $1-\Delta$,

$$\sup_{Z\in\mathcal{C}:\|Z-Z^*\|_2\le\sqrt{2r}} \langle \sigma W, Z - Z^*\rangle \le 2d\sigma\sqrt{2r} \le r/2.$$

Hence, one has $r_G^*(\Delta) \le 32\sigma^2 d^2$.

We apply Theorem 5 to get that with probability at least $1 - \exp(-d^2/2)$

$$\frac{1}{2}\|\tilde{Z} - Z^*\|_2^2 \le \langle \mathbb{E}C, Z^* - \tilde{Z}\rangle \le r_G^*(\exp(-d^2/2)) \le 32\sigma^2 d^2. \tag{7.11}$$

Next, we know that the oracle Z^* is the rank-one matrix $x^*\overline{x^*}^\top$ which has n as its largest eigenvalue and associated eigenspace $\{\lambda z^* : \lambda \in \mathbb{C}\}$. In particular, Z^* has a spectral gap $g = d$. Let $\bar{x} \in \mathbb{C}^d$ be a top eigenvector of $\tilde{Z}$ with norm $\|\bar{x}\|_2 = \sqrt{d}$. It follows from Davis-Kahan

Theorem (see, for example, Theorem 4.5.5 in Vershynin (2018) or Theorem 4 in Vu (2010)) that there exists an universal constant $c_0 > 0$ such that

$$\min_{z\in\mathbb{C}:|z|=1} \|\frac{\bar{x}}{\sqrt{d}} - z\frac{x^*}{\sqrt{d}}\|_2 \leq \frac{c_0}{g}\|\tilde{Z} - Z^*\|_2,$$

where $g = d$ is the spectral gap of Z^*. Using (7.12), we conclude that, with probability at least $1 - \exp(-d^2/2)$, it holds true that

$$\min_{z\in\mathbb{C}:|z|=1} \|\bar{x} - zx^*\|_2 \leq 8c_0\sigma\sqrt{d}.$$

7.5.2.2 Second upper bound on the fixed point r_G^*: the global approach

One may wonder what type of result we can get for the angular group synchronization problem with additive noise using the global approach. It is the aim of this last section to answer this question.

As for the community detection problem we will use Grothendieck inequality for the global approach. As in (7.10), the global approach is also using an inclusion of the localized set $\mathcal{C}_r$, but unlike (7.10), we just drop off the localization: we are simply using $\mathcal{C}_r \subset \mathcal{C}$. We have that

$$\sup_{Z\in\mathcal{C}:\|Z-Z^*\|_2\leq\sqrt{2r}} \langle \sigma W, Z - Z^*\rangle \leq \sigma \sup_{Z\in\mathcal{C}} \langle W, Z - Z^*\rangle \leq 2K_G^{\mathbb{C}}\sigma\|W\|_{cut},$$

where we used Grothendieck's inequality as in (6.91) but in the complex case ($K_G^{\mathbb{C}}$ denoting Grothendieck's constant in the complex case). Here the cut norm in the complex case is defined as

$$\|W\|_{cut} = \sup_{s_i,t_j\in\mathbb{C}:|s_i|=|t_j|=1} \Big|\sum_{i,j} W_{ij}s_i t_j\Big|.$$

We therefore end up with the computation of the cut norm of the noise as in Guédon and Vershynin (2016) for the community detection problem.

Here the noise being Gaussian, the cut norm $\|W\|_{cut}$ is the supremum of a Gaussian process for which we can use Borell's concentration inequality (see Ledoux (2001)) to get for all $u > 0$, with probability at least $1 - \exp(-u^2/2)$,

$$\|W\|_{cut} \leq \mathbb{E}\|W\|_{cut} + u \sup_{s_i,t_j\in\mathbb{C}:|s_i|=|t_j|=1} \sqrt{\mathbb{E}\Big|\sum_{i,j} W_{ij}s_i t_j\Big|^2} \leq \|W\|_{cut} + ud.$$

Next, we use Slepian's lemma (see Chapter 3 from Ledoux and Talagrand (1991)) to handle the complexity term $\mathbb{E}\|W\|_{cut}$. First, we need to upper bound the canonical metric associated with the Gaussian process: for every $s_i, s_i', t_j, t_j' \in \mathbb{C} : |s_i| = |t_j| = |s_i'| = |t_j'| = 1$ we have

$$\begin{aligned}
&\mathbb{E}\Big|\sum_{i,j} W_{ij}s_i t_j - \sum_{i,j} W_{ij}s_i' t_j'\Big|^2 = \sum_{i,j} |s_i t_j - s_i' t_j'|^2 \\
&= \sum_{i,j} |s_i(t_j - t_j') - (s_i' - s_i)t_j'|^2 \leq 2d\left(\sum_j |t_j - t_j'|^2 + \sum_i |s_i - s_i'|^2\right) \\
&= 2d\mathbb{E}\Big|\sum_i g_i(s_i - s_i') + \sum_j \eta_j(t_j - t_j')\Big|^2,
\end{aligned}$$

where $(g_i)_i, (\eta_j)_j$ are i.i.d. $\mathcal{N}(0,1)$. It follows from Slepian's lemma that

$$\mathbb{E}\|W\|_{cut} \leq \sqrt{2d}\mathbb{E} \sup_{s_i,t_j\in\mathbb{C}:|s_i|=|t_j|=1} \Big|\sum_i g_i(s_i - s_i') + \sum_j \eta_j(t_j - t_j')\Big| \leq 4\sqrt{2d}d.$$

Together with Borell's inequality above for $u = \sqrt{d}$, we obtain that with probability at least $1 - \exp(-n/2)$, $\|W\|_{cut} \leq 9n^{3/2}$.

As a consequence, for $\Delta' = \exp(-d/2)$, we have $r_G^*(\Delta') \leq 36K_G^{\mathbb{C}}\sigma d^{3/2}$. It follows from Theorem 5 that with probability at least $1 - \exp(-n/2)$

$$\frac{1}{2}\|\tilde{Z} - Z^*\|_2^2 \leq \left\langle \mathbb{E}C, Z^* - \tilde{Z} \right\rangle \leq r_G^*(\exp(-d/2)) \leq 36K_G^{\mathbb{C}}\sigma d^{3/2}. \tag{7.12}$$

and so

$$\min_{z\in\mathbb{C}:|z|=1} \|\bar{x} - zx^*\|_2 \leq c_0\sqrt{36K_G^{\mathbb{C}}\sigma}d^{1/4}.$$

We conclude that the global approach is better than the local approach using the inclusion in (7.10).

7.5.2.3 Third upper bound on the fixed point r_G^*: end of the proof of Theorem 34

The final approach is based on a decomposition from Fei and Chen (2019b) that we already used for the signed clustering problem. Here, for the angular group synchronization problem, the projection operator is simpler since Z^* is the rank-one matrix $x^*\overline{x^*}^\top$ and all the processes are Gaussian processes.

In order to work with independent random variables, we consider the following matrix $\Psi \in \mathbb{C}^{d\times d}$

$$\Psi_{ij} = \begin{cases} W_{ij} \text{ if } i \leq j \\ 0 \text{ otherwise,} \end{cases} \tag{7.13}$$

where 0 entries are considered as independent Gaussian variables with 0 variance and therefore, Ψ has independent Gaussian entries, and satisfies the relation $W = \Psi + \overline{\Psi}^\top$.

Like we did for the signed clustering problem, we decompose the inner product $\langle W, Z - Z^*\rangle$ into two parts according to the SVD of Z^*. We know that Z^* is the rank-one matrix $Z^* = x\overline{x^*}^\top$, then $v := x^*/\sqrt{d}$ is a unit singular vector of Z^* and we define the following projection operator

$$\mathcal{P} : M \in \mathbb{C}^{d\times d} \to v\bar{v}^\top M + Mv\bar{v}^\top - v\bar{v}^\top Mv\bar{v}^\top,$$

and its associated orthogonal projection

$$\mathcal{P}^\perp : M \in \mathbb{C}^{d\times d} \to M - \mathcal{P}(M) = (I_d - v\bar{v}^\top)M(I_d - v\bar{v}^\top).$$

For any $Z \in \mathcal{C}_r := \mathcal{C} \cap \{Z^* + \sqrt{2r}B_2^{d\times d}\}$, we consider the following decomposition as in (Fei & Chen, 2019b)

$$\langle W, Z - Z^*\rangle = \underbrace{\langle \mathcal{P}(Z - Z^*), W\rangle}_{S_1(Z)} + \underbrace{\langle \mathcal{P}^\perp(Z - Z^*), W\rangle}_{S_2(Z)}.$$

Next, we upper bound with large probability each of the two last terms uniformly over all $Z \in \mathcal{C}_r$. We start with the $S_1(Z)$ term: for any $Z \in \mathcal{C}_r$, we have

$$\begin{aligned}
S_1(Z) &= \langle W, \mathcal{P}(Z - Z^*)\rangle = \langle \mathcal{P}(W), Z - Z^*\rangle \\
&= \langle v\bar{v}^\top W, Z - Z^*\rangle + \langle Wv\bar{v}^\top, Z - Z^*\rangle - \langle v\bar{v}^\top Wv\bar{v}^\top, Z - Z^*\rangle \\
&= 2\langle v\bar{v}^\top W, Z - Z^*\rangle - \langle v\bar{v}^\top Wv\bar{v}^\top, Z - Z^*\rangle \\
&= 2\langle v\bar{v}^\top \Psi, Z - Z^*\rangle + 2\langle v\bar{v}^\top\overline{\Psi}^\top, Z - Z^*\rangle - \langle v\bar{v}^\top(\Psi + \overline{\Psi}^\top)v\bar{v}^\top, Z - Z^*\rangle \\
&= 2\langle v\bar{v}^\top \Psi, Z - Z^*\rangle + 2\langle v\bar{v}^\top\bar{\Psi}^\top, Z - Z^*\rangle - 2\langle v\bar{v}^\top\Psi v\bar{v}^\top, Z - Z^*\rangle \\
&= 2\langle v\bar{v}^\top \Psi, Z - Z^*\rangle + 2\langle v\bar{v}^\top\bar{\Psi}^\top, Z - Z^*\rangle - 2\langle v\bar{v}^\top\Psi, (Z - Z^*)v\bar{v}^\top\rangle.
\end{aligned}$$

Then, bounding separately each of those three terms will lead us to a bound for $S_1(Z)$. Let us show how to bound the first term. Similar arguments can be used to control the other two terms.

We define $V := v\bar{v}^\top \Psi$, so that $V_{ij} = \sum_k v_i \bar{v}_k \Psi_{kj} = \sum_{k \le j} v_i \bar{v}_k W_{kj}$. We want to find a high-probability upper bound on $\langle V, Z - Z^* \rangle$. To that end we simply use the inclusion $\mathcal{C}_r \subset Z^* + \sqrt{2r} B_2^{d \times d}$ to get

$$\sup_{Z \in \mathcal{C}_r} \langle V, Z - Z^* \rangle \le \sup_{Z \in \sqrt{2r} B_2^{d \times d}} \langle V, Z \rangle = \sqrt{2r} \|V\|_2$$

so that

$$\begin{aligned}\mathbb{E}\left[\sup_{Z \in \mathcal{C}_r} \langle V, Z - Z^* \rangle\right] &\le \sqrt{2r}\mathbb{E}\|V\|_2 \le \sqrt{2r}\left(\sum_{i,j}\sum_{k \le j} |v_i|^2 |v_j|^2 \mathbb{E}|\Psi_{kj}|^2\right)^{\frac{1}{2}} \\ &= \sqrt{2r}\left(\frac{1}{n}\sum_j j\right)^{1/2} \le \sqrt{2rd}.\end{aligned}$$

Moreover, we have

$$\mathbb{E}[|\langle V, Z - Z^* \rangle|^2] = \mathbb{E}[|\langle \Psi, v\bar{v}^T (Z - Z^*) \rangle|^2] = \sum_{i,j} |(v\bar{v}^T (Z - Z^*))_{ij}|^2$$

$$= \sum_{i,j} \left| \sum_k v_i \bar{v}_k (Z - Z^*)_{kj} \right|^2 \le \frac{1}{d^2} \sum_{i,j} \sum_k |(Z - Z^*)_{kj}|^2 \le \frac{\|Z - Z^*\|_2^2}{d} \le \frac{2r}{d}.$$

Now, we apply Borell's inequality (see, for instance, Theorem 4.1 in Ledoux (2001) or page 56-57 in Ledoux and Talagrand (1991)): for $u = d$, with probability at least $1 - \exp(-u^2/2)$,

$$\begin{aligned}\sup_{Z \in \mathcal{C}_r} \langle V, Z - Z^* \rangle &\le \mathbb{E}\left[\sup_{Z \in \mathcal{C}_r} \langle V, Z - Z^* \rangle\right] + \sup_{Z \in \mathcal{C}_r} \sqrt{\mathbb{E}\left[|\langle V, Z - Z^* \rangle|^2\right]} u \\ &\le \sqrt{2rd} + \sqrt{\frac{2r}{d}} u = 2\sqrt{2rd}.\end{aligned}$$

Similar calculus yield to the same upper bounds for the two other terms. Therefore, we obtain, with probability at least $1 - 3\exp(-d^2/2)$, it holds true that

$$\sup_{Z \in \mathcal{C}_r} S_1(Z) \le 6\sqrt{2rd}. \tag{7.14}$$

Now, it remains to control the second $S_2(Z)$ term. For any $Z \in \mathcal{C}_r$, we have

$$S_2(Z) = \langle W, \mathcal{P}^\perp(Z - Z^*) \rangle = \langle W, \mathcal{P}^\perp(Z) \rangle \le \|\mathcal{P}^\perp(Z)\|_* \|W\|_{\text{op}}.$$

Since $Z \succeq 0$, we have $(I_d - v\bar{v}^\top) Z (I_d - v\bar{v}^\top) \succeq 0$, that is $\mathcal{P}^\perp(Z) \succeq 0$ and so

$$\begin{aligned}\|\mathcal{P}^\perp(Z)\|_* = \text{Tr}(\mathcal{P}^\perp(Z)) &\overset{(i)}{=} \text{Tr}(\mathcal{P}^\perp(Z - Z^*)) = \text{Tr}((I_d - v\bar{v}^\top)(Z - Z^*)(I_d - v\bar{v}^\top)) \\ &\overset{(ii)}{=} \text{Tr}((I_d - v\bar{v}^\top)(Z - Z^*)) = \text{Tr}(Z - Z^*) - \text{Tr}(v\bar{v}^\top (Z - Z^*)) \\ &\overset{(iii)}{=} \text{Tr}(v\bar{v}^\top (Z^* - Z)) = \sum_{i,j} v_i \bar{v}_j (Z^* - Z)_{ij} \le \|v\|_2^2 \|Z - Z^*\|_2 \\ &= \|Z - Z^*\|_2 \le \sqrt{2r}.\end{aligned}$$

where (i) is due to the fact that $\mathcal{P}^{\perp}(Z^*) = 0$ by construction, (ii) holds because $(I_d - v\bar{v}^\top)$ is Hermitian and $(I_d - v\bar{v}^\top)^2 = (I_d - v\bar{v}^\top)$ and (iii) holds since $\mathrm{Tr}(Z) = \mathrm{Tr}(Z^*) = 1$ for any $Z \in \mathcal{C}_r$.

Moreover, it follows from Davidson and Szarek (2001) that, for all $u > 0$, with probability at least $1 - 2\exp(-u^2/2)$, $\|W\|_{\mathrm{op}} \leq 2\sqrt{d} + u$. We conclude that with probability at least $1 - 2\exp(-2d)$, we have

$$\sup_{Z \in \mathcal{C}_r} S_2(Z) \leq 4\sqrt{2dr}. \tag{7.15}$$

It follows from (7.14) and (7.15) that with probability at least $1 - 5\exp(-d/2)$

$$\sup_{Z \in \mathcal{C}_r} \langle W, Z - Z^* \rangle \leq 10\sqrt{2rd}.$$

Then, for $\Delta = 5\exp(-d/2)$ and $r = 2(20\sigma)^2 d$ we have, with probability at least $1 - \Delta$, $\sup_{Z \in \mathcal{C}_r} \langle \sigma W, Z - Z^* \rangle \leq r/2$. Hence, we conclude that $r^*_G(\Delta) \leq 2(20\sigma)^2 d$.

Now, we apply Corollary 5 to get that, with probability at least $1 - 5\exp(-d/2)$

$$\frac{1}{2}\|\tilde{Z} - Z^*\|_2^2 \leq 2(20\sigma)^2 d.$$

Next, we know that the oracle Z^* is the rank-one matrix $x^*\overline{x^*}^\top$ which has d for largest eigenvalue and associated eigenspace $\{\lambda z^* : \lambda \in \mathbb{C}\}$. In particular, Z^* has a spectral gap $g = d$. Let $\bar{x} \in \mathbb{C}^d$ be a top eigenvector of $\tilde{Z}$ with norm $\|\bar{x}\|_2 = \sqrt{d}$.

It follows from Davis-Kahan Theorem (see, for example, Theorem 4.5.5 in Vershynin (2018) or Theorem 4 in Vu (2010)) that there exists an universal constant $c_0 > 0$ such that

$$\min_{z \in \mathbb{C}: |z| = 1} \|\frac{\bar{x}}{\sqrt{d}} - z\frac{x^*}{\sqrt{d}}\|_2 \leq \frac{c_0}{g}\|\hat{Z} - Z^*\|_2,$$

where $g = n$ is the spectral gap of Z^*. Using the previous inequality, we conclude that, with probability at least $1 - 5\exp(-d/2)$

$$\min_{z \in \mathbb{C}: |z| = 1} \|\bar{x} - zx^*\|_2 \leq 40c_0\sigma.$$

7.6 Solving SDPs in practice

The practical implementation of our approach to the various problems we have studied here resorts to solving a convex optimization problem. In the present section, we describe the various algorithms we used for solving these SDPs.

7.6.1 Pierra's method

For SDPs with constraints on the entries, we propose a simple modification of the method initially proposed by Pierra (1984). Let $f\colon \mathbb{R}^{d\times d} \mapsto \mathbb{R}$ be a convex function. Let $\mathcal{C}$ denote a convex set which can be written as the intersection of convex sets $\mathcal{C} = \mathcal{S}_1 \cap \cdots \cap \mathcal{S}_J$. Let us define $\mathbb{H} = \mathbb{R}^{d\times d} \times \cdots \times \mathbb{R}^{d\times d}$ (J times) and let $\mathbb{D}$ denote the (diagonal) subspace of $\mathbb{H}$ of vectors of the form $(Z, \ldots, Z)$. In this new formalism, the problem can now be formulated as a minimization problem over the intersection of two sets only, i.e.

$$\min_{\mathsf{Z} \in \mathbb{H}} \left(\frac{1}{J} \sum_{j=1}^{J} f(\mathsf{Z}_j) : \mathsf{Z} = (\mathsf{Z}_j)_{j=1}^J \in (\mathcal{S}_1 \times \cdots \times \mathcal{S}_J) \cap \mathbb{D} \right).$$

Define $F(\mathsf{Z}) = \frac{1}{\mathsf{J}} \sum_{\mathsf{j}=1}^{\mathsf{J}} \mathsf{f}(\mathsf{Z}_\mathsf{j})$. The algorithm proposed by Pierra (1984) consists of performing the following iterations

$$\mathsf{Z}^{p+1} = \mathrm{Prox}_{I_{\mathcal{S}_1 \times \cdots \times \mathcal{S}_J} + \frac{1}{2}\varepsilon F}(\mathsf{B}^{\mathsf{p}}) \text{ and } \mathsf{B}^{\mathsf{p+1}} = \mathrm{Proj}_{\mathbb{D}}(\mathsf{Z}^{\mathsf{p+1}}).$$

Pierra's method can be shown to converge in the setting of our finite dimensional experiments using Martinet (1972, Chapter V).

7.6.1.1 Application to community detection

Let us now present the computational details of Pierra's method for the community detection problem. We will estimate its membership matrix $\bar{Z}$ via the following SDP estimator

$$\hat{Z} \in \mathrm{argmax}_{Z \in \mathcal{C}} \langle A, Z \rangle,$$

where $\mathcal{C} = \{Z \in \mathbb{R}^{d \times d}, Z \succeq 0, Z \geq 0, \mathrm{diag}(Z) \preceq I_d, \sum Z_{ij} \leq \lambda\}$ and $\lambda = \sum_{i,j=1}^{d} \bar{Z}_{ij} = \sum_{k=1}^{K} |\mathcal{C}_k|^2$ denotes the number of nonzero elements in the membership matrix $\bar{Z}$. The motivation for this approach stems from the fact that the membership matrix $\bar{Z}$ is actually the oracle, i.e., $Z^* = \bar{Z}$, where $Z^* \in \mathrm{argmax}_{Z \in \mathcal{C}} \langle \mathbb{E}[A], Z \rangle$. The function f to minimize in the Pierra algorithm is defined as $f(Z) = -\langle A, Z \rangle$.

Let us denote by $\mathbb{S}_+$ the set of symmetric positive semi-definite matrices in $\mathbb{R}^{d \times d}$. The set $\mathcal{C}$ is the intersection of the sets

$$S_1 = \mathbb{S}_+; \qquad S_2 = \left\{ Z \in \mathbb{R}^{d \times d} \mid Z \geq 0 \right\}; \qquad S_3 = \left\{ Z \in \mathbb{R}^{d \times d} \mid \mathrm{diag}(Z) \preceq I \right\};$$
$$\text{and } S_4 = \left\{ Z \in \mathbb{R}^{d \times d} \mid \sum_{i,j=1}^{d} Z_{ij} \leq \lambda \right\}.$$

We now compute for all $\mathsf{B} = (\mathsf{B}_\mathsf{j})_{\mathsf{j}=1}^{\mathsf{4}} \in (\mathbb{R}^{\mathsf{d} \times \mathsf{d}})^{\mathsf{4}}$ and $j = 1, \ldots, 4$ ($J = 4$ here)

$$\mathrm{Prox}_{I_{S_1 \times \cdots \times S_4} + \frac{1}{2}\varepsilon F}(\mathsf{B})_j = \mathrm{Prox}_{I_{S_j} + \frac{1}{2J}\varepsilon f}(\mathsf{B}_\mathsf{j}).$$

We have for $J = 4$

$$\mathrm{Prox}_{I_{S_j} + \frac{1}{2J}\varepsilon f}(\mathsf{B}_\mathsf{j}) = \mathrm{argmin}_{Z \in S_j} \; -\frac{\epsilon}{2J} \langle A, Z \rangle + \frac{1}{2} \|Z - \mathsf{B}_j\|_F^2 = P_{S_j}\left(\mathsf{B}_j + \frac{\epsilon}{2J} A\right)$$

On the other hand, the projections operators $P_{S_j}, j = 1, 2, 3, 4$ are given by

$$P_{S_1}(\mathsf{Z}_1) = U \max\{\Sigma, 0\} U^\top, \quad \text{where } \mathsf{Z}_1 \text{ has eigenvalue decomposition } \mathsf{Z}_1 = U\Sigma U^\top,$$
$$P_{S_2}(\mathsf{Z}_2) = \max\{\mathsf{Z}_2, 0\}, \qquad P_{S_3}(\mathsf{Z}_3) = \mathsf{Z}_3 - \mathrm{diag}(\mathsf{Z}_3) + \min\{1, \mathrm{diag}(\mathsf{Z}_3)\},$$
$$P_{S_4}(\mathsf{Z}_4) = \frac{\lambda}{\sum_{ij} (\mathsf{Z}_4)_{ij}} \mathsf{Z}_4.$$

To sum up, Pierra's method can be formulated as follows.

For all iterations k in $\mathbb{N}$, compute the SVD of $\mathsf{B}_1^k + \frac{\epsilon}{2 \cdot 4} A = U^k \Sigma^k (U^k)^\top$. Then compute for all $j = 1, \ldots, 4$

$$\mathsf{B}_\mathsf{j}^{\mathsf{k+1}} = \frac{1}{4}\Bigg(U^k \max\left\{\Sigma^k, 0\right\} (U^k)^\top + \max\left\{\mathsf{B}_2^k + \frac{\epsilon}{2 \cdot 4}\mathsf{A}, 0\right\} + \mathsf{B}_3^k + \frac{\epsilon}{2 \cdot 4} A - \mathrm{diag}(\mathsf{B}_3^k + \frac{\epsilon}{2 \cdot 4} A)$$
$$+ \min\left\{1, \mathrm{diag}(\mathsf{B}_3^k + \frac{\epsilon}{2 \cdot 4} A)\right\} + \frac{\lambda}{\sum_{ij} (\mathsf{B}_4^k + \frac{\epsilon}{2 \cdot 4} A)_{ij}} \mathsf{B}_4^k + \frac{\epsilon}{2 \cdot 4} A \Bigg).$$

7.6.1.2 Application to signed clustering

Let us now turn to the signed clustering problem. We will estimate its membership matrix $\bar{Z}$ via the following SDP estimator $\hat{Z} \in \operatorname{argmax}_{Z \in \mathcal{C}} \langle A, Z \rangle$, where $\mathcal{C} = \{Z \in \mathbb{R}^{d \times d} : Z \succeq 0, Z_{ij} \in [0,1], Z_{ii} = 1, i = 1, \ldots, d\}$. As in the community detection case, the function f to minimize in the Pierra algorithm is defined as $f(Z) = -\langle A, Z \rangle$.

Let us denote by $\mathbb{S}_+$ the set of symmetric positive semi-definite matrices in $\mathbb{R}^{d \times d}$. The set $\mathcal{C}$ is the intersection of the sets $S_1 = \mathbb{S}_+$, $S_2 = \left\{Z \in \mathbb{R}^{d \times d} \mid Z \in [0,1]^{d \times d}\right\}$ and $S_3 = \left\{Z \in \mathbb{R}^{d \times d} \mid Z_{ii} = 1,\ i = 1, \ldots, d\right\}$.

As before, for $j = 1, \ldots, 3$

$$\operatorname{Prox}_{I_{S_j} + \frac{1}{2 \cdot 3} \varepsilon f}\left(\mathsf{B_j}\right) = P_{S_j}\left(\mathsf{B}_j + \frac{\epsilon}{2 \cdot 3} A\right)$$

and the projection operators P_{S_j}, $j = 1, 2, 3$ are given by

$$P_{S_1}(\mathsf{Z}_1) = \mathsf{U} \max\left\{\Sigma, 0\right\} \mathsf{U}^\top,\ \mathsf{P}_{\mathsf{S}_2}(\mathsf{Z}_2) = \min\left\{\max\left\{\mathsf{Z}_2, 0\right\}, 1\right\} \text{ and } \mathsf{P}_{\mathsf{S}_3}(\mathsf{Z}_3) = \mathsf{Z}_3 - \operatorname{diag}(\mathsf{Z}_3) + \mathsf{I}$$

To sum up, Pierra's method can be formulated as follows.

At each iteration k, compute the SVD of $\mathsf{B}_1^k + \frac{\epsilon}{2 \cdot 3} A = U^k \Sigma^k (U^k)^\top$. Then compute for all $j = 1, \ldots, 3$

$$\mathsf{B}_j^{k+1} = \frac{1}{3}\left(U^k \max\left\{\Sigma^k, 0\right\} U^{k^t} + \min\left\{\max\left\{\mathsf{B}_2^k + \frac{\epsilon}{2 \cdot 3} A, 0\right\}, 1\right\}\right.$$
$$\left. + \mathsf{B}_3^k + \frac{\epsilon}{2 \cdot 3} A - \operatorname{diag}\left(\mathsf{B}_3^k + \frac{\epsilon}{2 \cdot 3} A\right) + I\right).$$

7.6.2 The Burer-Monteiro approach and the Manopt Solver

To solve the Max-Cut and Angular Synchronization problems, we rely on Manopt, a freely available Matlab toolbox for optimization on manifolds Boumal, Mishra, Absil, and Sepulchre (2014). Manopt runs the Riemannian Trust-Region method on corresponding Burer-Monteiro non-convex problem with rank bounded by p as follows. The Burer-Monteiro approach consists of replacing the optimization of a linear function $\langle A, Z \rangle$ over the convex set $\mathcal{Z} = \{Z \succeq 0 : \mathcal{A}(Z) = b\}$ with the optimization of the quadratic function $\langle AY, Y \rangle$ over the non-convex set $\mathcal{Y} = \{Y \in \mathbb{R}^{d \times p} : \mathcal{A}(YY^T) = b\}$.

In the context of the Max-Cut problem, the Burer-Monteiro approach amounts to the following steps. Denoting by Z the positive semidefinite matrix $Z = zz^T$, note that both the cost function and the constraints lend themselves to be expressed linearly in terms of Z. Dropping the NP-hard rank-1 constraint on Z, we arrive at the well-known convex relaxation of Max-Cut from Goemans and Williamson (1995)

$$\hat{Z} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \langle A, Z \rangle,$$

where $\mathcal{C} := \{Z \in \mathbb{R}^{d \times d} : Z \succeq 0, Z_{ii} = 1, \forall i = 1, \ldots, d\}$.

If a solution $\hat{Z}$ of this SDP has rank 1, then $\hat{Z} = z^* z^{*^T}$ for some z^*, which then gives the optimal cut. Recall that in the general case of higher rank $\hat{Z}$,

$$\hat{Y} \in \underset{X \in \mathcal{B}}{\operatorname{argmin}} \langle AY, Y \rangle, \tag{7.16}$$

where $\mathcal{B} := \{Y \in \mathbb{R}^{d \times p} : \operatorname{diag}(YY^T) = 1\}$. Note that the constraint $\operatorname{diag}(YY^T) = 1$ requires each row of Y to have unit ℓ_2^p norm, rendering Y to be a point on the Cartesian product of d

unit spheres $\mathcal{S}_2^{p-1}$ in $\mathbb{R}^p$, which is a smooth manifold. Also note that the search space of the SDP is compact, since all Z feasible for the SDP have identical trace equal to d.

If the convex set $\mathcal{Z}$ is compact, and m denotes the number of constraints, it holds true that whenever p satisfies $\frac{p(p+1)}{2} \geq m$, the two problems share the same global optimum Barvinok (1995), Burer and Monteiro (2005). Building on pioneering work of Burer and Monteiro (2005), Boumal, Voroninski, and Bandeira (2016) showed that if the set $\mathcal{Z}$ is compact and the set $\mathcal{Y}$ is a smooth manifold, then $\frac{p(p+1)}{2} \geq m$ implies that, for almost all cost matrices A, global optimality is achieved by any Y satisfying a second-order necessary optimality conditions. Following Boumal et al. (2016), *for $p = \lceil \sqrt{2d} \rceil$, for almost all matrices A, even though (7.16) is non-convex, any local optimum Y is a global optimum (and so is $Z = YY^T$), and all saddle points have an escape (the Hessian has a negative eigenvalues).* Note that for $p > d/2$ the same statement holds true for *all* A, and was previously established by Boumal (2015).

Bibliography

Abbe, E., Bandeira, A. S., & Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, *62*(1), 471–487.

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, *53*, 16–38.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Ames, B. P. (2014). Guaranteed clustering and biclustering via semidefinite programming. *Mathematical Programming*, *147*(1-2), 429–465.

Anjos, M. F., & Lasserre, J. B. (2011). *Handbook on semidefinite, conic and polynomial optimization*. Springer Science & Business Media.

Arun, K., Huang, T., & Bolstein, S. (1987). Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *9*(5), 698–700.

Baik, J., Ben Arous, G., & Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* *33*(5), 1643–1697. doi:10.1214/009117905000000233

Bandeira, A., Boumal, S., & Singer, A. (2016). Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*.

Bandeira, A. S., Charikar, M., Singer, A., & Zhu, A. (2014). Multireference alignment using semidefinite programming. In *Proceedings of the 5th conference on innovations in theoretical computer science* (pp. 459–470). ACM.

Bandeira, A. S., & Van Handel, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, *44*(4), 2479–2506.

Banerjee, S., Sarkar, K., Gokalp, S., Sen, A., & Davulcu, H. (2012). Partitioning signed bipartite graphs for classification of individuals and organizations. In *International conference on social computing, behavioral-cultural modeling, and prediction* (pp. 196–204). Springer.

Bartlett, P. L., & Mendelson, S. (2006). Empirical minimization. *Probab. Theory Related Fields*, *135*(3), 311–334.

Barvinok, A. I. (1995). Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, *13*(2), 189–202.

Bellec, P. C., Lecué, G., & Tsybakov, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* *46*(6B), 3603–3642. doi:10.1214/17-AOS1670

Birgé, L., & Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, *97*(1-2), 113–150. doi:10.1007/BF01199316

Blekherman, G., Parrilo, P. A., & Thomas, R. R. (2012). *Semidefinite optimization and convex algebraic geometry*. SIAM.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Boucheron, S., Lugosi, G., & Massart, P. (2013a). *Concentration inequalities : a non asymptotic theory of independence*. Oxford University Press. Retrieved from https://hal.inria.fr/hal-00942704

Boucheron, S., Lugosi, G., & Massart, P. (2013b). *Concentration inequalities: A nonasymptotic theory of independence*. ISBN 978-0-19-953525-5. Oxford University Press.

Boumal, N., Mishra, B., Absil, P.-A., & Sepulchre, R. (2014). Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, *15*, 1455–1459. Retrieved from http://www.manopt.org

Boumal, N., Voroninski, V., & Bandeira, A. (2016). The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in neural information processing systems 29* (pp. 2757–2765).

Boumal, N. (2015). A Riemannian low-rank method for optimization over semidefinite matrices with block-diagonal constraints. *arXiv preprint arXiv:1506.00575*.

Boyd, S., El Ghaoui, L., Feron, E., & Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*. Siam.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bunea, F., Giraud, C., Luo, X., Royer, M., & Verzelen, N. (2018). Model assisted variable clustering: Minimax-optimal recovery and algorithms. arXiv: 1508.01939 [stat.ME]

Burer, S., & Monteiro, R. (2005). Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, *103*(3), 427–444.

Candès, E. J., & Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, *56*(5), 2053–2080. doi:10.1109/TIT.2010.2044061

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *48*(4), 1148–1185. Publisher: Institut Henri Poincaré. doi:10.1214/11-AIHP454

Chafaï, D., Guedon, O., Lecue, G., & Pajor, A. (2012). *Interactions between compressed sensing random matrices and high dimensional geometry*. Panoramas et Synthesess. Societe Mathematique de France, Paris.

Chiang, K.-Y., Whang, J., & Dhillon, I. S. (2012). Scalable Clustering of Signed Networks using Balance Normalized Cut. In *Acm conference on information and knowledge management (cikm)*.

Chinot, G., Lecué, G., & Lerasle, M. (2018). Statistical learning with lipschitz and convex loss functions. *arXiv preprint arXiv:1810.01090*.

Chrétien, S., & Corset, F. (2009). Using the eigenvalue relaxation for binary least-squares estimation problems. *Signal Processing*, *89*(11), 2079–2091.

Chrétien, S., Dombry, C., & Faivre, A. (2016). A semi-definite programming approach to low dimensional embedding for unsupervised clustering. *Frontiers in Applied Mathematics and Statistics*.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

Cucuringu, M. (2015). Synchronization over Z_2 and community detection in multiplex networks with constraints. *Journal of Complex Networks*, *3*, 469–506.

Cucuringu, M. (2016). Sync-Rank: Robust Ranking, Constrained Ranking and Rank Aggregation via Eigenvector and Semidefinite Programming Synchronization. *IEEE Transactions on Network Science and Engineering*, *3*(1), 58–79.

Cucuringu, M., Davies, P., Glielmo, A., & Tyagi, H. (2019). Sponge: A generalized eigenproblem for clustering signed networks. *AISTATS 2019*.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I., & Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev. 49*(3), 434–448. doi:10.1137/050645506

Davenport, M. A., & Romberg, J. (2016). An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, *10*(4), 608–622.

Davidson, K. R., & Szarek, S. J. (2001). Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, *1*(317-366), 131.

Davis, C., & Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, *7*(1), 1–46.

Davis, J. A. (1967). Clustering and structural balance in graphs. *Human Relations*, *20*(2), 181–187.

De Castro, Y., Gamboa, F., Henrion, D., Hess, R., Lasserre, J.-B., et al. (2019). Approximate optimal designs for multivariate polynomial regression. *The Annals of Statistics*, *47*(1), 127–155.

De Castro, Y., Gamboa, F., Henrion, D., & Lasserre, J.-B. (2017). Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Trans. Information Theory*, *63*(1), 621–630. doi:10.1109/TIT.2016.2619368

Demanet, L., & Hand, P. (2014). Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, *20*(1), 199–221.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., & Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016* (pp. 655–664). IEEE Computer Soc., Los Alamitos, CA.

Fei, Y., & Chen, Y. (2019a). Achieving the bayes error rate in stochastic block model by sdp, robustly. In A. Beygelzimer & D. Hsu (Eds.), *Conference on learning theory, COLT 2019, 25-28 june 2019, phoenix, az, USA* (Vol. 99, pp. 1235–1269). Proceedings of Machine Learning Research. PMLR. Retrieved from http://proceedings.mlr.press/v99/fei19a.html

Fei, Y., & Chen, Y. (2019b). Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Trans. Inform. Theory*, *65*(1), 551–571.

Fei, Y., & Chen, Y. (2020). Achieving the Bayes error rate in synchronization and block models by SDP, robustly. *IEEE Trans. Inform. Theory*, *66*(6), 3929–3953. doi:10.1109/TIT.2020.2966438

Fletcher, R. (1981). A nonlinear programming problem in statistics (educational testing). *SIAM Journal on Scientific and Statistical Computing*, *2*(3), 257–267.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3-5), 75–174.

Foucart, S., & Rauhut, H. (2013). *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. doi:10.1007/978-0-8176-4948-7

Gärtner, B., & Matoušek, J. (2012). Semidefinite programming. In *Approximation algorithms and semidefinite programming* (pp. 15–25). Springer.

Gilbert, J. C., & Josz, C. (2017). Plea for a semidefinite optimization solver in complex numbers.

Giraud, C., & Verzelen, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. *Mathematical Statistics and Learning*, *1*(3), 317–374.

Goemans, M. X. (1997). Semidefinite programming in combinatorial optimization. *Mathematical Programming*, *79*(1-3), 143–161.

Goemans, M. X., & Williamson, D. P. (1994). New 34-approximation algorithms for the maximum satisfiability problem. *SIAM Journal on Discrete Mathematics*, *7*(4), 656–666.

Goemans, M. X., & Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, *42*(6), 1115–1145.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations* (Fourth). Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, *57*(3), 1548–1566. doi:10.1109/TIT.2011.2104999

Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S., & Eisert, J. (2010). Quantum state tomography via compressed sensing. *Physical review letters*, *105*(15), 150401.

Grothendieck, A. (1953). Résumé de la théorie métrique des produits tensoriels topologiques. *Bol. Soc. Mat. São Paulo*, *8*, 1–79.

Grothendieck, A. (1956). *Résumé de la théorie métrique des produits tensoriels topologiques*. Soc. de Matemática de São Paulo.

Guédon, O., & Vershynin, R. (2016). Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields*, *165*(3-4), 1025–1049. Retrieved from https://doi.org/10.1007/s00440-015-0659-z

Guédon, O., & Vershynin, R. (2016). Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, *165*(3-4), 1025–1049.

Hajek, B., Wu, Y., & Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, *62*(5), 2788–2797.

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* *53*(2), 217–288. doi:10.1137/090771806

He, S., Luo, Z.-Q., Nie, J., & Zhang, S. (2008). Semidefinite relaxation bounds for indefinite homogeneous quadratic optimization. *SIAM Journal on Optimization*, *19*(2), 503–523.

Hong, D., Lee, H., & Wei, A. (2021). Optimal solutions and ranks in the max-cut sdp. arXiv: 2109.02238 [math.OC]

Hopkins, S. B. (2018). Sub-gaussian mean estimation in polynomial time. *CoRR*, *abs/1809.07425*. arXiv: 1809.07425. Retrieved from http://arxiv.org/abs/1809.07425

Hou, J. P. (2005). Bounds for the least Laplacian eigenvalue of a signed graph. *Acta Mathematica Sinica*, *21*(4), 955–960.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (Second). Wiley Series in Probability and Statistics. doi:10.1002/9780470434697

Jerrum, M. R., Valiant, L. G., & Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci. 43*(2-3), 169–188. doi:10.1016/0304-3975(86)90174-X

Johnstone, I. M., & Lu, A. Y. (2009a). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc. 104*(486), 682–693. doi:10.1198/jasa.2009.0121

Johnstone, I. M., & Lu, A. Y. (2009b). Sparse principal components analysis. *arXiv preprint arXiv:0901.4392*.

Journée, M., Nesterov, Y., Richtárik, P., & Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res. 11*, 517–553.

Karger, D., Motwani, R., & Sudan, M. (1998). Approximate graph coloring by semidefinite programming. *Journal of the ACM (JACM)*, *45*(2), 246–265.

Khot, S., & Naor, A. (2009). Approximate kernel clustering. *Mathematika*, *55*(1-2), 129–165.

Koltchinskii, V. (2011a). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Berlin: Springer.

Koltchinskii, V. (2011b). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Lecture Notes in Mathematics. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. doi:10.1007/978-3-642-22147-7

Koltchinskii, V. (2011c). *Oracle inequalities in empirical risk minimization and sparse recovery problems: École d'été de probabilités de saint-flour xxxviii-2008*. doi:10.1007/978-3-642-22147-7

Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., De Luca, E. W., & Albayrak, S. (2010). Spectral analysis of signed graphs for clustering, prediction and visualization. *SDM*, *10*, 559–570.

Lasserre, J. B. (2015). *An introduction to polynomial and semi-algebraic optimization*. Cambridge University Press.

Latała, R. et al. (1997). Estimation of moments of sums of independent real random variables. *The Annals of Probability*, *25*(3), 1502–1513.

Lavaei, J., & Low, S. H. (2011). Zero duality gap in optimal power flow problem. *IEEE Transactions on Power Systems*, *27*(1), 92–107.

Le, C. M., Levina, E., & Vershynin, R. (2016). Optimization via low-rank approximation for community detection in networks. *Ann. Statist. 44*(1), 373–400. Retrieved from https://doi.org/10.1214/15-AOS1360

Lecué, G., & Mendelson, S. (2014). Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.* ArXiv:1401.2188.

Lecué, G., & Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *Ann. Statist. 48*(2), 906–931. doi:10.1214/19-AOS1828

Lecué, G., & Mendelson, S. (2013). *Learning subgaussian classes: Upper and minimax bounds*. CNRS, Ecole polytechnique and Technion.

Lecué, G., & Mendelson, S. (2017). Regularization and the small-ball method i: Sparse recovery. arXiv: 1601.05584 [math.ST]

Lecué, G., & Mendelson, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist. 46*(2), 611–641. doi:10.1214/17-AOS1562

Ledoux, M. (2001). *The concentration of measure phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI.

Ledoux, M., & Talagrand, M. (1991). *Probability in Banach spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Isoperimetry and processes. Berlin: Springer-Verlag.

Ledoux, M., & Talagrand, M. (2013). *Probability in banach spaces*. A Series of Modern Surveys in Mathematics. Springer-Verlag, Berlin Heidelberg GmbH.

Lemaréchal, C., Nemirovskii, A., & Nesterov, Y. (1995). New variants of bundle methods. *Mathematical programming*, *69*(1-3), 111–147.

Lemaréchal, C., & Oustry, F. (2018). Semidefinite relaxations and lagrangian duality with application to combinatorial optimization. *INRIA, rapport de recherche*, (3710).

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010a). Predicting positive and negative links in online social networks. In *WWW* (pp. 641–650).

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010b). Signed Networks in Social Media. In *CHI* (pp. 1361–1370).

Lovász, L. (1979). On the shannon capacity of a graph. *IEEE Transactions on Information theory*, *25*(1), 1–7.

Ma, W.-K. K. (2010). Semidefinite relaxation of quadratic optimization problems and applications. *IEEE Signal Processing Magazine*, *1053*(5888/10).

Magdon-Ismail, M. (2015). Np-hardness and inapproximability of sparse pca. arXiv: 1502.05675 [cs.LG]

Mammen, E., & Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist. 27*(6), 1808–1829. doi:10.1214/aos/1017939240

Martinet, B. (1972). *Algorithmes pour la résolution de problèmes d'optimisation et de minimax* (Doctoral dissertation, Université Joseph-Fourier-Grenoble I).

Massart, P. (2007). *Concentration inequalities and model selection*. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Berlin: Springer.

Mazziotti, D. A. (2011). Large-scale semidefinite programming for many-electron quantum mechanics. *Physical review letters*, *106*(8), 083001.

Mendelson, S., Pajor, A., & Tomczak-Jaegermann, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal. 17*(4), 1248–1282. doi:10.1007/s00039-007-0618-7

Mossel, E., Neeman, J., & Sly, A. (2015). Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields*, *162*(3-4), 431–461. doi:10.1007/s00440-014-0576-6

Nemirovsky, A. S., & Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York.

Nesterov, Y. (1997). Semidefinite relaxation and non-convex quadratic optimization. *Optimization Methods and Software*, *12*, 1–20.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, *74*(3), 036104.

Olsson, C., Eriksson, A. P., & Kahl, F. (2007). Solving large scale binary quadratic problems: Spectral methods vs. semidefinite programming. In *2007 ieee conference on computer vision and pattern recognition* (pp. 1–8). IEEE.

Peng, J., & Wei, Y. (2007a). Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, *18*(1), 186–205.

Peng, J., & Wei, Y. (2007b). Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, *18*(1), 186–205. doi:10.1137/050641983. eprint: https://doi.org/10.1137/050641983

Peng, J., & Xia, Y. (2005). A new theoretical framework for k-means-type clustering. In *Foundations and advances in data mining* (pp. 79–96). Springer.

Pierra, G. (1984). Decomposition through formalization in a product space. *Mathematical Programming*, *28*(1), 96–115.

Pisier, G. (2012). Grothendieck's theorem, past and present. *Bulletin of the American Mathematical Society*, *49*(2), 237–323.

Porter, M. A., Onnela, J.-P., & Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, *56*(9), 1082–1097.

Royer, M. (2017). Adaptive clustering through semidefinite programming. In *Advances in neural information processing systems* (pp. 1795–1803).

Scobey, P., & Kabe, D. (1978). Vector quadratic programming problems and inequality constrained least squares estimation. *J. Indust. Math. Soc. 28*, 37–49.

Shapiro, A. (1982). Weighted minimum trace factor analysis. *Psychometrika*, *47*(3), 243–264.

Singer, A., & Shkolnisky, Y. (2011). Three-dimensional structure determination from common lines in Cryo-EM by eigenvectors and semidefinite programming. *SIAM Journal on Imaging Sciences*, *4*(2), 543–572.

Singer, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, *30*(1), 20–36.

Sun, J., Boyd, S., Xiao, L., & Diaconis, P. (2006). The fastest mixing markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM review*, *48*(4), 681–699.

Todd, M. J. (2001). Semidefinite optimization. *Acta Numerica*, *10*, 515–560.

Tsybakov, A. B. (2003). Optimal rate of aggregation. In *Computational learning theory and kernel machines (colt-2003)* (Vol. 2777, pp. 303–313). Lecture Notes in Artificial Intelligence. Springer, Heidelberg.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist. 32*(1), 135–166. doi:10.1214/aos/1079120131

van de Geer, S. A. (2000). *Applications of empirical process theory*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. With applications to statistics. New York: Springer-Verlag.

Vapnik, V. N., & Chervonenkis, A. Y. (1974). *Teoriya raspoznavaniya obrazov. statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow.

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York.

Vapnik, V. N. (2000). *The nature of statistical learning theory* (Second). Statistics for Engineering and Information Science. doi:10.1007/978-1-4757-3264-1

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Vu, V. (2010). Singular vectors under random perturbation. *Preprint available in arXiv:104.2000*.

Waldspurger, I., d'Aspremont, A., & Mallat, S. (2013). Phase recovery, maxcut and complex semidefinite programming. arXiv: 1206.0102 [math.OC]

Waldspurger, I., d'Aspremont, A., & Mallat, S. (2015). Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, *149*(1-2), 47–81.

Wang, T., Berthet, Q., & Samworth, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist. 44*(5), 1896–1930. doi:10.1214/15-AOS1369

Wolkowicz, H. (1999). Semidefinite and Lagrangian relaxations for hard combinatorial problems. In *Ifip conference on system modeling and optimization* (pp. 269–309). Springer.

Wolkowicz, H., Saigal, R., & Vandenberghe, L. (2012). *Handbook of semidefinite programming: Theory, algorithms, and applications*. Springer Science & Business Media.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2002). Distance metric learning, with application to clustering with side-information. In *Proceedings of the 15th international conference on neural information processing systems* (pp. 521–528). NIPS'02. Cambridge, MA, USA: MIT Press.

Xu, M., Jog, V., Sun, H., & Loh, P. (2020). Optimal rates for community estimation in the weighted stochastic block model. *Annals of statistics*.

Yang, Z., Balasubramanian, K., & Liu, H. (2018). On stein's identity and near-optimal estimation in high-dimensional index models. arXiv: 1709.08795 [math.ST]

Yu, Y., Wang, T., & Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, *102*(2), 315–323. doi:10.1093/biomet/asv008. eprint: /oup/backfile/content_public/journal/biomet/102/2/10.1093_biomet_asv008/1/asv008.pdf

Yu, Y., Wang, T., & Samworth, R. J. (2014). A useful variant of the davis–kahan theorem for statisticians. arXiv: 1405.0680 [math.ST]

Zhang, S., & Huang, Y. (2006). Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, *16*(3), 871–890.

Zhang, S. (2000). Quadratic maximization and semidefinite relaxation. *Mathematical Programming*, *87*(3), 453–465.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Apprentissage avec une fonction de perte linéaire : bornes générales pour les estimateurs ERM et minimax MOM, avec applications.

Mots clés : Optimisation Semi-Définie Positive, Minimisation du Risque Empirique, Sparsité, Statistique robuste

Résumé : La détection de communautés sur des graphes, la récupération de phase, le clustering signé, la synchronisation angulaire, le problème de la coupe maximale, la sparse PCA, ou encore le single index model, sont des problèmes classiques dans le domaine de l'apprentissage statistique. Au premier abord, ces problèmes semblent très dissemblables, impliquant différents types de données et poursuivant des objectifs distincts. Cependant, la littérature récente révèle qu'ils partagent un point commun : ils peuvent tous être formulés sous la forme de problèmes d'optimisation semi-définie positive (SDP). En utilisant cette modélisation, il devient possible de les aborder du point de vue classique du machine learning, en se basant sur la minimisation du risque empirique (ERM) et en utilisant la fonction de perte la plus élémentaire: la fonction de perte linéaire. Cela ouvre la voie à l'exploitation de la vaste littérature liée à la minimisation du risque, permettant ainsi d'obtenir des bornes d'estimation et de développer des algorithmes pour résoudre ces problèmes. L'objectif de cette thèse est de présenter une méthodologie unifiée pour obtenir les propriétés statistiques de procédures classiques en machine learning basées sur la fonction de perte linéaire. Cela s'applique notamment aux procédures SDP, que nous considérons comme des procédures ERM. L'adoption d'un "point de vue machine learning" nous permet d'aller plus loin en introduisant d'autres estimateurs performants pour relever deux défis majeurs en apprentissage statistique : la parcimonie et la robustesse face à la contamination adversaire et aux données à distribution à queue lourde. Nous abordons le problème des données parcimonieuses en proposant une version régularisée de l'estimateur ERM. Ensuite, nous nous attaquons au problème de la robustesse en introduisant un estimateur basé sur le principe de la "Médiane des Moyennes" (MOM), que nous nommons l'estimateur minmax MOM. Cet estimateur permet de faire face au problème de la robustesse et peut être utilisé avec n'importe quelle fonction de perte, y compris la fonction de perte linéaire. Nous présentons également une version régularisée de l'estimateur minmax MOM. Pour chacun de ces estimateurs, nous sommes en mesure de fournir un "excès de risque" ainsi que des bornes d'estimation, en utilisant deux outils clés : les points fixes de complexité locale et les équations de courbure de la fonction d'excès de risque. Afin d'illustrer la pertinence de notre approche, nous appliquons notre méthodologie à cinq problèmes classiques en machine learning, pour lesquels nous améliorons l'état de l'art.

Title : Learning with a linear loss function: excess risk and estimation bounds for ERM and minimax MOM estimators, with applications.

Keywords : Semi-Definite Programming, Empirical Risk Minimization, Sparsity, Robust Statistics

Abstract : Community detection, phase recovery, signed clustering, angular group synchronization, Max-cut, sparse PCA, the single index model, and the list goes on, are all classical topics within the field of machine learning and statistics. At first glance, they are pretty different problems with different types of data and different goals. However, the literature of recent years shows that they do have one thing in common: they all are amenable to Semi-Definite Programming (SDP). And because they are amenable to SDP, we can go further and recast them in the classical machine learning framework of risk minimization, and this with the simplest possible loss function: the linear loss function. This, in turn, opens up the opportunity to leverage the vast literature related to risk minimization to derive excess risk and estimation bounds as well as algorithms to unravel these problems. The aim of this work is to propose a unified methodology to obtain statistical properties of classical machine learning procedures based on the linear loss function, which corresponds, for example, to the case of SDP procedures that we look as ERM procedures. Embracing a machine learning view point allows us to go into greater depth and introduce other estimators which are effective in handling two key challenges within statistical learning: sparsity, and robustness to adversarial contamination and heavy-tailed data. We attack the structural learning problem by proposing a regularized version of the ERM estimator. We then turn to the robustness problem and introduce an estimator based on the median of means (MOM) principle, which we call the minmax MOM estimator. This latter estimator addresses the problem of robustness and can be constructed whatever the loss function, including the linear loss function. We also present a regularized version of the minmax MOM estimator. For each of those estimators we are able to provide excess risk and estimation bounds, which are derived from two key tools: local complexity fixed points and curvature equations of the excess risk function. To illustrate the relevance of our approach, we apply our methodology to five classical problems within the frame of statistical learning, for which we improve the state-of-the-art results.

Institut Polytechnique de Paris
91120 Palaiseau, France