Journal of Machine Learning Research 25 (2024) 1-90 Submitted 10/23; Revised 12/24; Published 12/24

Learning with a linear loss function: excess risk and estimation bounds for ERM, minmax MOM and their regularized versions with applications to robustness in sparse PCA.

Guillaume Lecué LECUE@ESSEC.EDU
IDS Department
ESSEC, Business school
3 Av. Bernard Hirsch, 95000 Cergy, France.

Lucie Neirac LUCIE.NEIRAC@ENSAE.FR
Statistics Department
CREST, ENSAE, IPParis
5 Av. Henry Le Chatelier, Palaiseau, France.

Editor: Pradeep Ravikumar

Abstract

Motivated by several examples, we consider a general framework of learning with linear loss functions. In this context, we provide excess risk and estimation bounds that hold with large probability for four estimators: ERM, minmax MOM and their regularized versions. These general bounds are applied for the problem of robustness in sparse PCA. In particular, we improve the state of the art result for this this problems, obtain results under weak moment assumptions as well as for adversarial contaminated data.

Keywords: SDP relaxation, empirical processes, robustness, heavy-tailed, adversarial contamination, high-dimensional statistics.

1. Introduction

Community detection, phase recovery, signed clustering, angular group synchronization, Max-Cut, sparse PCA, and the sparse 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, they can all be written in such a way that a common analysis of various estimators introduced for these problems can be analyzed the same way. All these problems 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 for all the problems cited above as well as many other ones. 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. Indeed, this observation is the baseline of Chrétien et al. (2021): 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 (see, for instance, Vapnik (2000), Boucheron et al. (2013) or Koltchinskii (2011a)) developed during the last fifty years for ERM in this very specific framework of the linear loss function.

General excess risk and estimation bounds have therefore been obtained in Chrétien et al. (2021) for ERM using a linear loss function. State-of-the art techniques like localization, homogeneity argument, local curvature and complexity fixed points equation have been used in Chrétien et al. (2021) to obtain these general bounds that have then been applied in Community detection, phase recovery, signed clustering, angular group synchronization and Max-Cut. This new perspective allowed us to obtain new results or recover older one with a new proof technique but most importantly it showed that a common analysis of several problems that looks a priori very different can be performed.

The aim of the article is to push forward the analysis of statistical procedures based on linear loss functions and to show that this viewpoint allows to deal with the problem of structural risk minimization and of robustness[1] in all the problems cited above and in many other ones (some of them are given below). Indeed, Chrétien et al. (2021) only deals with ERM procedures. However, some problems rely on some structure such as sparsity and other are facing the problem of robustness. For these issues, ERM is not the right answer and these two problems call for other procedures such as regularized ERM (for structural risk minimization) or the recently introduced minmax MOM estimator Lecué and Lerasle (2020) (for robustness issues). It is therefore the first contribution of this paper to derive general statistical bounds for regularized ERM, minmax MOM and its regularized version in the framework of linear loss functions. As an illustration these bounds are applied to the problem of sparse PCA. Using our viewpoint, we improve state-of the art results for this problem (improvement on the rates and the deviation for less stringent assumptions) as well as getting robust (to heavy-tailed data and to adversarial contamination) versions of these results thanks to the minmax MOM approach. Another aim of this article is to show that the linear loss functions appears in many problems and so we provide a list of problems that can be recast in this framework. But first, we explain how linear loss function appear only recently, even though they are simpler than many other loss functions previously used in machine learning such as the quadratic or the logistic loss functions.

Statistics, machine learning and optimization got closer during the last twenty years and gave birth in part to *data sciences*. One consequence of these connections is that nowadays statistical estimators and machine learning procedures should be computable on a laptop in a reasonable amount of time and should not be purely theoretical objects. This viewpoint shed some lights on algorithms from the statistical perspective and may now be seen as statistical procedures that can receive a statistical analysis such as satisfying excess risk bounds. For instance, statistical properties of some gradient-descent based algorithms and SDP relaxation procedures have been obtained during the last twenty years. In particular, the SDP relaxation has proved to be very successful first in optimization and nowadays in statistics for many graph related issues such as community detection. From our perspective, SDP relaxation has been at the origin of many examples of ERMs based on a linear loss function.

Semidefinite programming (SDP) as a mathematical concept was introduced in the late 1980s and early 1990s. The foundations of SDP were laid down by researchers such as

[1] In all this article, robustness means robust to data contamination and to heavy-tailed data.

Yurii Nesterov, Arkadi Nemirovski, and others (Nemirovskii and Nesterov (1985), Boyd and Vandenberghe (1997), Nesterov (1998)), who extended the ideas of linear programming to semidefinite matrices, allowing for the optimization of linear functions subject to semidefinite constraints. The theoretical development and algorithms for solving SDP problems gained significant attention during this period, leading to its establishment as a fundamental optimization framework within the mathematical community.

The growing interest in SDPs in recent years is due to several compelling factors. One of the main factors is its broad applicability, as it can address a wide variety of complex problems arising in various mathematical contexts, including graph theory Gaar et al. (2022), Gualandi (2009), combinatorial optimization Gutekunst and Williamson (2019), signal processing Luo and Yu (2006), quantum information Wang et al. (2016b), for the Komlós conjecture Bansal et al. (2019) or in integer programming Rendl (2016). Its potency lies in its ability to efficiently handle non-convex and combinatorial optimization challenges by approximating them with convex semidefinite constraints. At the same time, the development of efficient algorithms for solving SDP problems, such as interior-point methods Helmberg et al. (1996) and first-order methods Monteiro (2003), has significantly improved the feasibility of tackling large-scale SDPs, thereby broadening the range of possibilities for applying SDP to real-world problems.

From our point-of-view SDP relaxations provide many examples of machine learning procedures such as ERM or RERM (regularized ERM) based on a linear loss function. We are now providing some of these examples and later we will dive deeper into the example of sparse PCA.

Notations. Throughout this paper, we use uppercase letters for matrix and lowercase letters for vectors. For a matrix $A \in \mathbb{R}^{N\times P}$, we note $A \geqslant 0$ to indicate that $A_{ij} \geqslant 0$ for any $(i,j) \in \{1,\dots,N\} \times \{1,\dots,P\}$ and $A \succeq 0$ to say that A is positive semidefinite. For A and $B \in \mathbb{R}^{N\times P}$, we define their Frobenius inner product as $\langle A, B \rangle := \text{Trace}(B^\top A)$, and we write $A \circ B$ for their element-wise product. If x is a vector in $\mathbb{C}^d$ then $|x|$ denotes the vector in $\mathbb{R}^d$ made of the modules of the coordinates of x. We denote $[N] = \{1,\dots,N\}$.

Community detection. SDPs have been used to handle the problem of community detection on graphs in Guédon and Vershynin (2016) or Fei and Chen (2019) under the Stochastic Block Model assumption, which is as follows. 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 $i \nsim j$ 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 $A = (A_{i,j})_{1\leqslant i,j,\leqslant d}$ the observed symmetric adjacency matrix, such that, for all $1 \leqslant i \leqslant j \leqslant d$, 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:

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

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^*:

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

This estimator is therefore obtained as the solution of 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.

Variable clustering. SDP estimators have been used 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 independant 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 need to define the membership matrix $Q \in \mathbb{R}^{p\times K}$ 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 membership matrix $Z^* \in \mathbb{R}^{d\times d}$ defined by $Z^*_{ij} = |G_k|^{-1}\mathbb{1}_{\{i \text{ and } j \text{ belong to the same group } G_k\}}$. There is a one-to-one correspondence between partitions G and their corresponding membership 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 (2007), the authors show that the best partition for the X_i's can be estimated with the one corresponding to the following membership matrix:

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

where $A := \frac{1}{N}\sum_{i=1}^{N} X_i X_i^\top$ is the empirical covariance of the X_i's. In the noiseless case, we would have $Z^* \in \operatorname{argmax}_{Z\in\mathcal{C}}\langle \mathbb{E}[A], Z\rangle$. The estimator $\hat{Z}$ can therefore 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.

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 Bandeira et al. (2016). The authors consider that they observe $d(d-1)/2$ measurements of the following form:

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

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 to the following maximization problem:

$$\underset{x\in\mathcal{E}}{\operatorname{argmax}} \left\{\bar{x}^\top \, \mathbb{E}A \, x\right\} \text{ where } \mathcal{E} := \Big\{x \in \mathbb{C}^d : |x_i| = 1 \text{ for all } i = 1, \ldots, d\Big\}.$$

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

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

They show that in this setting, x^* can be obtain from Z^* as its leading unit-length eigen vector. Since $\mathbb{E}[A]$ is not known and only observed through A, $\hat{Z} \in \operatorname{argmin}_{Z\in\mathcal{C}} \left(-\big\langle A, Z\big\rangle\right)$ is a natural estimator for Z^*. This is therefore another example of an ERM estimator based on the observation of the matrix A and the linear loss function $Z \to \ell_Z(A) = -\big\langle A, Z\big\rangle$.

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 $V := \{1, \ldots, d\}$ 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:

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

Then, using an SDP relaxation by removing the rank constraint, we recover the classical Max-Cut SDP relaxation procedure introduced by Goemans and Williamson. The ERM counterpart based on the data A is

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

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

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:

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

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

$$\min\big(\big\langle \mathbb{E}[A], Z\big\rangle : Z \succeq 0, Z_{ii} = 1 \forall i, \text{rank}(Z) = 1\big)$$

which may be relaxed by dropping the rank constraint:

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

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 $\hat{Z} \in \text{argmin}_{Z\in\mathcal{C}}\big\langle A, Z\big\rangle$ which is a SDP optimization problem that we see as an ERM with the linear loss function $Z \to \ell_Z(A) := \big\langle A, Z\big\rangle$.

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

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

where $Z \succeq 0$ is positive semidefinite. We note that, since one has $\text{Tr}((\text{X} - \text{Y})(\text{X} - \text{Y})^\top\text{Z}) = \|Z^{1/2}(X - Y)\|_2^2$, 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 $\min_{Z\succeq 0}\sum_{i,j=1}^N d_Z(X_i, X_j)^2$. However, this last problem is trivially solved by $Z = 0$ hence, we may add some constraints: we suppose to know M points $(Y_i)_{i=1,\dots,M}$, distinct from the X_i's, for which we want $\sum_{i,j=1}^M d_Z(Y_i, Y_j) \geqslant 1$ to be satisfied. 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:

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

One can show that the set $\mathcal{C}$ is convex (see Appendix A.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^*: $\hat{Z} \in \text{argmin}_{Z\in\mathcal{C}}\big\langle A, Z\big\rangle$ which is again an ERM estimator with the linear loss function $Z \to \ell_Z(A) = \big\langle A, Z\big\rangle$, constructed from an observation of the random matrix A.

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

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

where $\mathfrak{S}_N$ is the set of all permutations of $[N]$. Finding a solution to (2) is a standard problem in optimal transport that can be lifted to the matrix problem

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

and $\mathcal{C}$ is the set of all $N \times N$ bi-stochastic matrices (i.e. of matrices with non-negative entries summing to one along rows and columns). Indeed, if τ^* denotes an optimal solution to (2) 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 $\sigma \geqslant 0$ and $(G_i, G'_i)_{i=1}^N$ are $2N$ i.i.d. standard mean zero random vectors in $\mathbb{R}^d$. The quadratic assignment problem for this two noisy cloud of points is a solution to the problem

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

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 OT 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, $\hat{Z} \neq Z^*$. Once again, one may looked at $\hat{Z}$ as an ERM for a linear loss function.

The sparse single index model. For this last example, 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:

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

where $\beta^* \in \mathbb{R}^d$ 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 p_0. The joint density of X is then $p = \otimes_{j=1}^d p_0$ 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 $x \in \mathbb{R}$ and the first and second score functions associated with p are defined for p-almost all $x = (x_j)_{j=1}^d$ by

$$S(x) = (s(x_j))_{1 \leqslant j \leqslant d} \in \mathbb{R}^d \text{ and } T(x) = S(x)S(x)^\top - \operatorname{diag}\left(\left(s'(x_j)\right)_{1 \leqslant j \leqslant d}\right).$$

Unlike the previous examples, the dimension d may be larger than N however, the target index β^* is assumed to be k-sparse with $k < N$. We therefore fall into the realm of structural learning. The work of Yang et al. (2018) focuses on this problem where it is proved that β^* can be obtained as the leading eigenvector of

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

and $A := YT(X)$. Regularized ERM promotes the sparsity structure via a ℓ_1-regularization. The oracle Z^* 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, the linear loss function $Z \to \ell_Z(A) = -\langle A, Z\rangle$ and a ℓ_1 regularization.

Goal of the paper. The list of examples provided above indicates that there is a real interest in the general study of linear loss functions in machine learning (we will provide later one more examples in structural learning for which we will provide a complete statistical analysis). Our aim is to propose such a unified methodology to obtain statistical properties of classical machine learning procedures based on linear loss functions such as the SDP procedures introduced above that we are now looking as ERM procedures constructed with a linear loss function. We continue the work begun in Chrétien et al. (2021) and go further here by presenting three other estimators that address the two problems of structural risk minimization and robustness. Our machine learning viewpoint allows to introduce new procedures (addressing the previously mentioned two issues) as well as study their statistical properties.

Framework. Our general framework is as follows. Let H be a Hilbert space. Let A be a random vector in H that we observe and $\mathcal{C} \subset H$ be a constraint set (most of the time it will be a convex set). We suppose to be interested in an object which is the solution to the 'oracle' optimization problem

$$Z^* \in \underset{Z\in\mathcal{C}}{\operatorname{argmax}} \langle \mathbb{E}[A], Z\rangle. \tag{3}$$

In some cases, Z^* is not our direct object of interest, but knowing about it enables us to achieve our objective (for instance, by retrieving one of its first eigen-vector). We then propose several estimators for the estimation of the oracle Z^*, among which we will choose depending on the presence or not of some particular structure and on the quality of the data (presence or not of corrupted and/or heavy-tailed data).

The first estimator we propose is the one studied in Chrétien et al. (2021) and is the standard ERM estimator built on the random matrix A but for the (non standard) linear loss function, that is $Z \to \ell_Z(A) = -\langle A, Z\rangle$:

$$\hat{Z} \in \underset{Z\in\mathcal{C}}{\operatorname{argmax}} \langle A, Z\rangle. \tag{4}$$

Then, we turn to two classical machine learning and statistics problems: structured learning and robustness. Leveraging on our view point (i.e. all the previous procedures are all ERMs), we attack the structural learning problem by proposing a regularized version of this ERM estimator by adding a regularization function to the objective function in (4). Afterwards, we turn to the robustness problem and introduce an estimator based on the median of means (MOM) principle, which has been introduced in Lecué and Lerasle (2020) and that is called the minmax MOM. This latter estimator addresses the problem of robustness and can be constructed whatever the loss function is and 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 the sparse PCA problem. Using our approach we are able to handle this classical statistical problem using our general excess risk and estimation bounds. As a result we improve the state-of-the art results in sparse PCA as well as introduce new procedures with statistical optimal guarantees that solve the problems of robust structural learning for this problem. Efficient robust gradient descent based algorithms may easily be derived from these procedures as in Lecué and Lerasle (2020). We provide such a construction in Remark 5 below. We will however not dive deeper into the algorithmic consequences of our approach.

2. General excess risk and estimation bounds for ERM, minmax MOM estimators and their regularized versions

In this section, 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 results for ERM are taken from Chrétien et al. (2021) and are recalled here for completeness and because it presents an 'easy' setup for the introduction of two key tools: local complexity fixed points and local curvature equations. The proofs of all the results are postponed to Section 5. They use state-of-the art machinery such as localization, homogeneity argument, 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 apply these results in the sparse PCA framework later to show how these general bounds can be applied in a concrete example.

2.1 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 provide for each example the random matrix A appearing in (3) and (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 $p \geqslant 1$, 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*, $\ell_Z : X \in H \to -\langle X, Z \rangle$ (it is an alignment measure, which 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 $\mathcal{C}$, i.e. we want to estimate/learn/infer/test

$$Z^* \in \operatorname*{argmin}_{Z \in \mathcal{C}} P\ell_Z. \tag{5}$$

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 $Z \to P\ell_Z$. However, we have access to a sample distributed according to P. This sample / 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, $A := \frac{1}{N}\sum_{i=1}^N X_i X_i^\top$ is the empirical covariance of the observed variables X_i's;
- in angular synchronization, $A = \left(e^{\iota\delta_{ij}} + \epsilon_{ij}\right)_{1\leqslant i<j\leqslant d}$ is made of the noisy measurements of the pairwise offsets;
- in the Max-Cut problem, A is the adjacency matrix of the observed graph;
- in phase recovery, $A := \left(XX^\top\right)\circ(I_N - BB^+)$, where X is the vector of the N observed measurements and B is the measurement matrix;
- in distance metric learning, $A := \sum_{i,j=1}^N (X_i - X_j)(X_i - X_j)^\top$ where the X_i's are the observed data from which we want to learn the metric;
- in noisy optimal transport, $A = (\|X_i - Y_j\|_2^2)_{1\leqslant i,j\leqslant N}$, where $\{X_1, \dots, X_N\}$ and $\{Y_1, \dots, Y_N\}$ are the two sets of observed vectors that we wish to transport one over the other;
- in the sparse single index model, $A = \frac{1}{N}\sum_{i=1}^N Y_i T(X_i)$, 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.

Remark 1 *Most of the problems introduced in Section 1 are presented as maximization problems, whereas ERM is a minimization problem. Given the linearity of the loss function,* *there are several ways to write the maximization problem into a minimization one: one may 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 $\ell_Z : A \to -\langle A, Z\rangle$, i.e. we take the opposite of the loss function, which is still a linear one.*

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 1, $\mathcal{C}$ is a convex set by construction. For our theoretical purpose, we will however need a weaker assumption given now: the star-shaped property.

Definition 2.1 *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 Section 1 because a convex set is star-shaped in any of its elements.

2.2 The ERM estimator and its regularized version: definition 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. The bounds for the ERM are taken from Chrétien et al. (2021). We reproduce them here because they introduce key quantities (localization, local curvature and complexity fixed points) in an 'easy' setup and they will appear in the study of the three other estimators in a more convoluted way.

2.2.1 ERM for the linear loss function

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

$$\hat{Z} \in \underset{Z\in\mathcal{C}}{\operatorname{argmin}}\, P_N\ell_Z \text{ where } P_N\ell_Z = \frac{1}{N}\sum_{i=1}^{N}\ell_Z(X_i) = \frac{1}{N}\sum_{i=1}^{N}\langle -X_i, Z\rangle.$$

The ERM is the natural empirical version of the oracle Z^* since $P\ell_Z$ appearing in the definition of Z^* in (5) has been replaced by its empirical counterpart $P_N\ell_Z$. When there is only one observation, ie $N = 1$, 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 (see, for instance, Koltchinskii (2011a) and Boucheron et al. (2005)). The results recalled below are for the special case of the linear loss function and are taken from Chrétien et al. (2021). They are however based on nowadays classical concepts in machine learning.

A key quantity driving the rate of convergence of the ERM 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 Chrétien et al. (2021) or Section 3 below), since it requires to control with large probability the supremum of an empirical processes indexed by a "localized classes", i.e. of 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 2.2 *[Complexity fixed point parameter] Let $0 < \Delta < 1$. The fixed point complexity parameter at deviation $1-\Delta$ is*

$$r^*(\Delta) = \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z\in\mathcal{C}:P\mathcal{L}_Z\leqslant r}(P_N - P)\mathcal{L}_Z \leqslant \frac{r}{2}\right] \geqslant 1-\Delta\right). \tag{6}$$

In what follows, we give some statistical properties of the ERM $\hat{Z}$ build from this complexity parameter. They are taken from Chrétien et al. (2021) up to the slight modification that the results from Chrétien et al. (2021) have been stated in the case $N = 1$ and $X_1 = A$. They can however be extended to the general sample size N just by replacing the empirical measure P_1 by P_N. Below we state these results in the general case.

Theorem 2.3 (Theorem 1 in Chrétien et al. (2021)) *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 $P\mathcal{L}_{\hat{Z}} \leqslant r^*(\Delta)$.*

From Theorem 2.3, we get that one way to grab some information on the ERM is to get an upper bound for the complexity fixed point $r^*(\Delta)$. To that end, one needs to understand the shape of the sets $\mathcal{C} \cap \{Z : P\mathcal{L}_Z \leqslant r\}$ for $r > 0$. This task may however be hard because of the shape of the neighborhoods $\{Z : P\mathcal{L}_Z \leqslant r\}$ given by the excess risk. In that case, it has been shown Chinot et al. (2018) that one can leverage on a *local curvature* of the excess risk to introduce easier to compute fixed points. We are now introducing the complexity fixed point associated with this other localization and then the notion of local curvature. In what follows, G is some function from H to $\mathbb{R}$.

Definition 2.4 *[Complexity fixed point parameter with G-localization] Let $0 < \Delta < 1$. The fixed point complexity parameter with respect to the G-localization at deviation $1-\Delta$ is*

$$r_G^*(\Delta) = \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z\in\mathcal{C}:G(Z-Z^*)\leqslant r}(P_N - P)\mathcal{L}_Z \leqslant \frac{r}{2}\right] \geqslant 1-\Delta\right). \tag{7}$$

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 2.5 *For all $Z \in \mathcal{C}$, if $P\mathcal{L}_Z \leqslant r_G^*(\Delta)$ then $P\mathcal{L}_Z \geqslant G(Z^* - Z)$.*

There are examples where one can show a curvature of the excess risk over the entire set $\mathcal{C}$ - this is for instance the case in the sparse PCA example below (see Lemma 4.3 below). In that case, we speak about a *global* curvature. What shows the following result is that we only need a *local* curvature of the excess risk around Z^* to hold in order to get statistical bounds for the ERM $\hat{Z}$.

Theorem 2.6 (Corollary 1 in Chrétien et al. (2021)) *We assume that the constraint $\mathcal{C}$ is star-shaped in Z^* and that the "local curvature" Assumption 2.5 holds for some $0 < \Delta < 1$. With probability at least $1 - \Delta$, it holds true that*

$$r_G^*(\Delta) \geqslant P\mathcal{L}_{\hat{Z}} \geqslant G(Z^* - \hat{Z}).$$

Finally a third and final estimation bound is given in the following for cases where Assumption 2.5 is hard to verify. They are situations where the shape of the local subsets $\mathcal{C} \cap \{Z : P\mathcal{L}_Z \leqslant r\}$ is hard to understand. In that case, we can simplify this assumption by considering neighborhoods with respect to the G function.

Assumption 2.7 *For all $Z \in \mathcal{C}$, if $G(Z^* - Z) \leqslant r_G^*(\Delta)$, then $P\mathcal{L}_Z \geqslant G(Z^* - Z)$.*

The following result establishes that, under Assumption 2.7, $\hat{Z}$ is a good estimate of Z^* with respect to the G function, but no guarantee on the excess risk is obtained.

Theorem 2.8 (Theorem 2 in Chrétien et al. (2021)) *We assume that the constraint $\mathcal{C}$ is star-shaped in Z^* and that the "local curvature" Assumption 2.7 holds for some $0 < \Delta < 1$. We assume that the G function is continuous, $G(0) = 0$ and $G(\lambda(Z^* - Z)) \leqslant \lambda G(Z^* - Z)$ for any $\lambda \in [0, 1]$ and $Z \in \mathcal{C}$. Then, with probability at least $1 - \Delta$, it holds true that $G(Z^* - \hat{Z}) \leqslant r_G^*(\Delta)$.*

We refer the reader to Chrétien et al. (2020) for the application of these results in community detection, signed clustering, angular group synchronization (for both multiplicative and additive models) 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 to improve the rates of convergence. They are however situations such as in sparse PCA or in the sparse single index model where the target has a structure that can be beneficial in order to improve statistical performance. In such cases, one may consider some regularization procedures like in the following section.

2.2.2 Regularized ERM for the linear loss

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 2.1 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:

$$\hat{Z}^{\mathrm{RERM}} \in \underset{Z\in\mathcal{C}}{\operatorname{argmin}} \left(P_N \ell_Z + \lambda \|Z\|\right) \tag{8}$$

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 $\mathcal{C} \cap \{Z : \|Z - Z^*\| \leqslant r\}$. 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 2.9 *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*

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

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 a description of the local curvature of the excess risk. As we already mentioned above, the G function describes the curvature of the excess risk locally around the oracle.

Assumption 2.10 *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 $\|Z - Z^*\| \leqslant \rho^*$, then $AP\mathcal{L}_Z \geqslant G(Z - Z^*)$.*

We now leverage on the structure inducing property of the regularization norm and explain what features must the radius ρ^* appearing in Assumption 2.10 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:

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

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 2.11 *For $A > 0$, $\rho > 0$ and $\delta \in (0,1)$ we define:*

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

and

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

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

Note that it is always true that $\Delta(\rho, A) \leqslant \rho$ – 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 $\|\cdot\|$. More details and intuition on the objects introduced in Definition 2.11 may be found in Lecué and Mendelson (2017) and Chinot et al. (2018).

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 2.12 *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 2.10 holds and $\rho^* > 0$ satisfies the A-sparsity equation from Definition 2.11. Define the function $r^*(.) := r^*_{\mathrm{RERM,G}}(A, ., \delta)$ and assume that*

$$\frac{10}{21A} \frac{r^*(\rho^*)}{\rho^*} < \lambda < \frac{2}{3A} \frac{r^*(\rho^*)}{\rho^*}. \tag{9}$$

Then, with probability at least $1 - \delta$, the following bounds hold for the RERM estimator defined in (8):

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

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 2.10 is trivially verified with $A = 1$, and as a consequence Theorem 2.12 applies.

2.3 Median of Means estimators: 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 median-of-means (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 2.13 *[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 $\{X_1, \dots, X_N\}$ that the adversary gives to the statistician. Hence, the dataset at hands $\{X_1, \dots, X_N\}$ 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 $\mathcal{O} \cup \mathcal{I} = \{1, \dots, N\}$ is unknown to the statistician.*

Remark 2 *Since there are two types of data considered in Assumption 2.13 (the 'good' $\tilde{X}_i$s and the corrupted ones X_is), 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:*

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

where $\tilde{X}$ has the same probability distribution as $\tilde{X}_1, \dots, \tilde{X}_N$. It is also the same for the L_2-norm: for all $Z \in H$, $\|Z\|_{L_2} = \sqrt{\mathbb{E}\left\langle \tilde{X}, Z \right\rangle^2}$. 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 and Ronchetti (2009). There was an important renewal of this topic during the last ten years with Catoni (2012), in statistics and Diakonikolas et al. (2016) in computer science. The approach we use in this section is based on the median-of-means principle introduced in Alon et al. (1999), Nemirovsky and Yudin (1983) and Jerrum et al. (1986): $[N]$ is partitioned into K equal-size groups $B_1, \dots, B_K$ (w.l.o.g. 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:

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

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.

2.3.1 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) and Remark 5 below for more details). When the dataset has been split into K equal size blocks, it takes the following form:

$$\hat{Z}_K^{\mathrm{MOM}} \in \underset{Z \in \mathcal{C}}{\mathrm{argmin}} \sup_{Z' \in \mathcal{C}} \mathrm{MOM}_K(\ell_Z - \ell_{Z'}) \tag{10}$$

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 probability 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 2.14 *Let $\sigma_1, \ldots, \sigma_N$ be N independent Rademacher variables which are independent of the $\tilde{X}_i$'s. For $\gamma > 0$, we define:*

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

where, for all $r > 0$,

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

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

Theorem 2.15 *We consider the adversarial contamination setup of Assumption 2.13. 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 $K \geqslant 100|\mathcal{O}|$. Then, it holds true that with probability at least $1 - \exp(-72K/625)$, $P\mathcal{L}_{\hat{Z}_K^{\text{MOM}}} \leqslant r^*_{\text{MOM,ER}}(\gamma)^2$.*

Compared to the fixed point from Definition 2.2 describing the rate of convergence of the ERM, we note that the one from Definition 2.14 uses a local Rademacher complexity, denoted by $E(r)$, and a variance term, denoted by $V_K(r)$. In particular, there is no need to upper bound with high probability the supremum of an empirical process but only its expectation. For minmax MOM estimators, the task of computing fixed point complexity parameters is therefore easier. Moreover, as one can see in Theorem 2.15, 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 2.13. 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 $AP\mathcal{L}_Z \geqslant \|Z-Z^*\|_{L_2}^2, \forall Z \in \mathcal{C}$ holds, i.e. with a margin parameter equal to 2, as introduced in 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 2.16 *Let $\sigma_1, \dots, \sigma_N$ be independent Rademacher variables which are independent of the $\tilde{X}_i$'s. For $\gamma > 0$, we define*

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

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

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

Assumption 2.17 *There exists $A > 0$ such that for any $Z \in \mathcal{C}$, if $\|Z - Z^*\|_{L_2}^2 \leqslant C_{K,A}$, then $\|Z - Z^*\|_{L_2}^2 \leqslant AP\mathcal{L}_Z$, where $C_{K,A} := \max\left(r^*_{\text{MOM},L_2}(\gamma)^2, \gamma^{-1}A^2(K/N)\right)$ for $\gamma = 1/3200$.*

Looking at Assumption 2.17, 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 2.17 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 2.18 *We consider the adversarial contamination setup of Assumption 2.13. 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 2.17 holds. Let K be a divisor of N such that $K \geqslant 100|\mathcal{O}|$. Then, it holds true that with probability at least $1-\exp(-72K/625)$:*

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

Theorem 2.18 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 w.r.t. 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 2.19 *Let $\sigma_1, \dots, \sigma_N$ 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:*

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

where, for all $r > 0$,

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

and

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

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

The difference between $r^*_{\text{MOM,ER}}$ and $r^*_{\text{MOM,G}}$ 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 2.20 is that $r^*_{\text{MOM,G}}$ may be easier to compute than $r^*_{\text{MOM,ER}}$, 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 $r^*_{\text{MOM,ER}} \leqslant r^*_{\text{ERM,G}}$ 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 (10) under this local curvature assumption.

Theorem 2.21 *We consider the adversarial contamination setup of Assumption 2.13. We assume that the constraint set $\mathcal{C}$ is star-shaped in Z^*. We consider a continuous function $G : H \to \mathbb{R}$. Let $\gamma = 1/6400$. We assume the existence of $0 < A < 2$ such that the local curvature Assumption 2.20 holds for those values of γ and G. Then, with probability at least $1 - \exp(-72K/625)$ it holds true that:*

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

Theorem 2.21 may be applied in the examples introduced from Section 1 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 as in the next section.

2.3.2 Regularized minmax MOM estimators for the linear loss function

We are now considering the setup of robust 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 2.3.1:

$$\hat{Z}_{K,\lambda}^{\text{RMOM}} \in \operatorname*{argmin}_{Z \in \mathcal{C}} \sup_{Z' \in \mathcal{C}} \left(\text{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\| - \|Z'\|) \right) \tag{11}$$

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

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

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

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

The sparsity equation introduced for the study of the RERM in Definition 2.11 has to be slightly modified according to this new definition of the complexity parameter.

Definition 2.23 *For $\gamma > 0$ and $\rho > 0$, let*

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

*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 $\bar{\Delta}(\rho) \geqslant 4\rho/5$.*

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 2.24 *We consider the adversarial contamination setup of Assumption 2.13. Let $K \in [N]$ be such that $K \geqslant 100|\mathcal{O}|$. Let $\rho^* > 0$ satisfying the sparsity equation from Definition 2.23. Let $\gamma = 1/3200$ and take $\lambda = (11/(40\rho^*)) r^*_{\text{RMOM,ER}}(\gamma, 2\rho^*)$ as regularization parameter. Then, with probability at least $1 - 2\exp(-72K/625)$,*

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

Note that one may replace $r^*_{\text{RMOM,ER}}(\gamma, 2\rho^*)$ by any real number r^* larger than $r^*_{\text{RMOM,ER}}(\gamma, 2\rho^*)$. This observation is particularly useful since we usually only know how to upper bound local complexity fixed points such as $r^*_{\text{RMOM,ER}}(\gamma, 2\rho^*)$ 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 2.25 *Let $(\sigma_i)_{i \leqslant N}$ be independent Rademacher variables independent of the $\tilde{X}_i$'s. For $\rho > 0$ and $\gamma > 0$, we define:*

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

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 $\hat{Z}_{K,\lambda}^{RMOM}$ lie (with high probability).

Definition 2.26 *For γ, ρ and $A > 0$, let:*

$$C_K(\gamma, \rho, A) := \max\left(320000 A^2 \frac{K}{N}, r^*_{\mathrm{RMOM},L_2}(\gamma,\rho)^2\right),$$

$$\widetilde{H}_{\rho,A} := \left\{ Z \in \mathcal{C} : \|Z - Z^*\| = \rho \text{ and } \|Z - Z^*\|_{L_2} \leqslant \sqrt{C_K(\gamma,\rho,A)} \right\}$$

and

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

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

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 2.27 *There exists A, γ and $\rho^* > 0$ such that ρ^* satisfies the A-sparsity equation from Definition 2.26 and for both $b \in \{1,2\}$ and all $Z \in \mathcal{C}$, if $\|Z - Z^*\|_{L_2}^2 = C_K(\gamma, b\rho^*, A)$ and $\|Z - Z^*\| \leqslant b\rho^*$, then $\|Z - Z^*\|_{L_2}^2 \leqslant AP\mathcal{L}_Z$.*

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 (w.r.t. to both L_2 and the regularization norm) bounds.

Theorem 2.28 *We consider the adversarial contamination setup of Assumption 2.13. Let K be a divisor of N and assume that $K \geqslant 100|\mathcal{O}|$. Grant Assumption 2.27 for some $A \in (0,1]$, $\gamma = 1/32000$ and ρ^* that satisfies the A-sparsity equation from Definition 2.26. Define $\lambda = (11/(40\rho^*))C_K(\gamma, 2\rho^*, A)$. Then it holds true that with probability at least $1 - 2\exp(-72K/625)$:*

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

and

$$\|\hat{Z}_{K,\lambda}^{\mathrm{RMOM}} - Z^*\|_{L_2}^2 \leqslant r^*_{\mathrm{RMOM,L_2}}(\gamma, 2\rho^*)^2.$$

Again the same result as the one of Theorem 2.28 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 localizations. An example of such a G function is given in the sparse PCA case studied later. Indeed, in Lemma 4.3 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 2.29 *Let $\sigma_1, \dots, \sigma_N$ 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:*

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

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

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 4.13 below. The final ingredient to derive the rate of convergence is the radius ρ that needs to satisfy a sparsity equation.

Definition 2.30 *For all γ and $\rho > 0$, consider*

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

*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 $\bar{\Delta}(\rho) \geqslant 4\rho/5$.*

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

Assumption 2.31 *There exists $A > 0$, $\gamma > 0$ and $\rho^* > 0$ such that ρ^* satisfies the spartsity equation from Definition 2.30 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 $\|Z - Z^*\| \leqslant b\rho^*$, then $AP\mathcal{L}_Z \geqslant G(Z - Z^*)$.*

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 2.32 *We consider the adversarial contamination setup of Assumption 2.13. Let $G : H \to \mathbb{R}$ be a continuous function such that $G(0) = 0$ and for all $\alpha \geqslant 1$ and $Z \in \mathcal{C}, G(\alpha(Z - Z^*)) \geqslant \alpha G(Z - Z^*)$. Let $K \in [N]$ be such that $K \geqslant 100|\mathcal{O}|$. Grant Assumption 2.31 for some $A \in (0, 1]$, $\gamma = 1/32000$ and ρ^* that satisfies the sparsity equation from*

Definition 2.30. 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:*

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

and

$$G(\hat{Z}^{\mathrm{RMOM}}_{K,\lambda} - Z^*) \leqslant r^*_{\mathrm{RMOM,G}}(\gamma, 2\rho^*)^2.$$

In the sparse PCA example, Theorem 2.32 will be applied for the study of a ℓ_1-regularized minmax MOM estimator. However, applying Theorem 2.32 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 2.30 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 $\mathcal{C}_{\rho,r} = \{Z \in \mathcal{C} : \|Z - Z^*\| \leqslant \rho, G(Z - Z^*) \leqslant r^2\}$. This step is usually the hardest one.

Remark 3 *The implementation of the methodology presented in this article depends on our ability to efficiently calculate the local curvature of the excess risk close to the oracle, some complexity fixed points as well as solving the sparsity equation in the case of structural learning. The most demanding step in this scheme is the computation of the local complexity fixed points which may require technical skills in empirical process theory, random matrices theory as well as in the geometry of Banach spaces. This is particularly true under weak moments assumptions. In our previous work Chrétien et al. (2020), we provide several examples of such computations of complexity fixed points. In the next section to come, we provide two other examples proving that performing such computations is possible. However, this step may be identified as the main limitation of our approach to the study of structural and robust learning with a linear loss function.*

3. Two examples of 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 use to compute local complexity fixed points like the ones used in the previous section. Indeed, in order to use any of the general statistical bounds presented in Section 2, we have to compute local complexity fixed points. We provide two such examples in this section that will be useful for the next section on the sparse PCA problem. Note that the bounds presented here hold under weak moment assumptions (i.e. roughly speaking $\log(d)$ moments are enough) and may be of independent interest.

In this section, we use the following notations: $X_1, \ldots, X_N$ are i.i.d. centered random vectors in $\mathbb{R}^d$ and we denote by Σ their covariance matrix, i.e. $\mathbb{E}X_1X_1^\top = \Sigma$. The entries of Σ are denoted by Σ_{pq} i.e. $\mathbb{E}X_{1p}X_{1q} = \Sigma_{pq}$ for all $p, q \in [d]$ where $X_1 = (X_{1j})_{j=1}^d$. We denote the empirical covariance matrix by $\hat{\Sigma}_N = (1/N)\sum_{i=1}^N X_iX_i^\top$ 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 $\Sigma - \hat\Sigma_N$ for two norms defined by interpolation bodies. The proofs of the two Theorems 3.2 and 3.4 below are postponed to Section 5.2.

3.1 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 w.r.t. some norm we need to assume some concentration properties on the X_i's. We therefore consider such an assumption now.

Assumption 3.1 *There exists $w \geqslant 0$ and $t \geqslant 1$ such that the following holds: for all $p, q \in [d]$ and all $2 \leqslant r \leqslant 2\log(ed/k) + t$ we have $\|X_{1p}X_{1q} - \mathbb{E}(X_{1p}X_{1q})\|_{L_r} \leqslant w^2 r$.*

In other words, Assumption 3.1 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 $\|UV\|_{\psi_1} \leqslant \|U\|_{\psi_2}\|V\|_{\psi_2}$ and the r-th moment of a ψ_α variable growths like $r^{1/\alpha}$ (see Chapter 1 in Chafaï et al. (2012) for more details). Assumption 3.1 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 3.1 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 first k-sparse principal component in deviation with an improved rate of convergence of order

$$\sqrt{\frac{k^2\log(ed/k)}{N}}. \tag{12}$$

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

$$\|A\| = \sup\left(\langle A, Z\rangle : Z \in kB_1 \cap B_2\right). \tag{13}$$

Theorem 3.2 *There exists an absolute constant c_0 such that the following holds. Grant Assumption 3.1 for some w and $t \geqslant 1$ and assume that $N \geqslant 2\log(ed/k)+t$. With probability at least $1-\exp(-t)$,*

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

Moreover, if $N \geqslant 2\log(ed/k)+1$, it holds true that $\mathbb{E}\left[\|\hat\Sigma_N - \Sigma\|\right] \leqslant c_0 w^2\sqrt{6k^2\log(ed/k)/N}$.

Remark 4 *Classical estimation result require the number of observations to be larger than $s\log(ed/s)$ where s is the sparsity of signal to be reconstructed. Here, we observe in Theorem 3.2 that N is only asked to be larger than $\log(ed/k)$ so it is a much weaker assumption than in the classical high-dimensional setup. The rational behind this phenomenon is 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 processes that require N to be larger than the sparsity up to a log factor. We don't have such a quadratic process to lower bound in our 'linear loss function' framework.*

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

As in the last section, we need some assumption on the existence of moments on the coordinates of X_1. We consider such an assumption now.

Assumption 3.3 *There exists $w \geqslant 0$ and $t \geqslant 3$ such that the following holds. For all $p, q \in [d]$ and all $2 \leqslant r \leqslant \log(ed^2) + t$ we have $\|X_{1p}X_{1q} - \mathbb{E}(X_{1p}X_{1q})\|_{L_r} \leqslant w^2 r$.*

Our aim is to analyze the statistical properties of a SLOPE regularization for the sparse PCA problem and to show that the optimal rate (12) can be achieved by a unique regularization method which does not require the a priori knowledge of 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 $q \geqslant p$. 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.

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

$$\|A\|_\rho = \sup\left(\left\langle A, Z\right\rangle : Z \in \rho B_{SLOPE} \cap B_2\right). \tag{14}$$

Theorem 3.4 *There exists an absolute constant c_0 such that the following holds. Let $k \in [d]$ and $\gamma \geqslant 1$. Grant Assumption 3.3 for some w and $t \geqslant \max\left(2\log(\lceil\log(k^2)\rceil), \gamma \log(ed^2/k^2)\right)$ and assume that $N \geqslant \log(ed^2) + t$. With probability at least $1 - 2\exp(-t/2)$,*

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

4. 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. 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 (see Halko et al. (2011) and Golub and Van Loan (2013)).

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 et al. (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,a). It can be stated as the following optimization problem:

$$\hat{v}_1 \in \underset{\|v\|_2=1, \|v\|_0 \leqslant k}{\operatorname{argmax}} \|\hat{\Sigma}_N v\|_2 \tag{15}$$

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 (15): 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 '$\|v\|_0 \leqslant k$', the constraint set is not convex. If the sparsity constraint was not there, then (15) 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 (15) 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 et al. (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.

4.1 SDP relaxation in sparse PCA

Let $X \in \mathbb{R}^d$ be a centered random vector with distribution P. Let $X_1, \dots, X_N \in \mathbb{R}^d$ 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 $k \leqslant d$, and such that the variance of the X_i's when projected onto v^* is maximal. This can be written as follows:

$$v^* \in \underset{v\in\mathcal{E}}{\operatorname{argmax}}\, \mathbb{E}[\langle X, v\rangle^2] \text{ where } \mathcal{E} := \left\{v \in \mathbb{R}^d : \|v\|_2 = 1, \|v\|_0 \leqslant k\right\}. \tag{16}$$

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 $\|v\|_0\leqslant k$, we have $\mathrm{Tr}(Z)=\|v\|_2^2=1$ and $\|Z\|_0\leqslant k^2$. Finding a solution of (16) is then equivalent (see d'Aspremont et al. (2007) and Lemaréchal and Oustry (2018)) to finding a top singular vector of $Z^\star$, where $Z^\star$ is solution of the optimization problem

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

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 (16). 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, $\mathcal{C}_0:=\left\{Z\in\mathbb{R}^{d\times d}: Z\succeq 0, \mathrm{Tr}(Z)=1, \|Z\|_0\leqslant k^2, \mathrm{rank}(Z)=1\right\}$ is not convex, because of two non-convex constraints: *the cardinality constraint " $\|Z\|_0\leqslant k^2$ "* 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:

$$Z^*\in \underset{Z\in\mathcal{C}}{\mathrm{argmax}}\langle \mathbb{E}[A], Z\rangle \text{ where } \mathcal{C}:=\{Z\in\mathbb{R}^{d\times d}: Z\succeq 0, \mathrm{Tr}(Z)=1\}. \tag{17}$$

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 2 to derive estimators for Z^* and provide statistical guarantees thanks to the results from Section 3.

This configuration allows us to refer to the work of Wang et al. (2016a). The authors study the sparse PCA problem where the distribution of the data $X_1,\dots,X_N$ 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. (2016a) for a definition). In particular, they propose the following ℓ_1-regularized ERM estimator

$$\hat{Z}\in\underset{Z\in\mathcal{C}}{\mathrm{argmin}}\left(\langle\frac{-1}{N}\sum_{i=1}^N X_iX_i^\top, Z\rangle+\lambda\|Z\|_1\right) \text{ where } \mathcal{C}:=\{Z: Z\succeq 0, \mathrm{Tr}(Z)=1\} \tag{18}$$

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

Theorem 4.1 *[Theorem 5 in Wang et al. (2016a)] Let $X_1,\dots,X_N\in\mathbb{R}^d$ be i.i.d random vectors with distribution in $\mathcal{P}$ and a covariance matrix satisfying the spiked covariance model: $\mathbb{E}[X_iX_i^\top]=I_d+\theta\beta^*(\beta^*)^\top$, 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 $Z\to\langle\frac{1}{N}\sum_{i=1}^N X_iX_i^\top, Z\rangle-\lambda\|Z\|_1$ over the model $\mathcal{C}$ defined in (18). 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 $4\log(d)\leqslant N\leqslant k^2d^2\theta^{-2}$ and $0<\theta\leqslant k$, then it holds true that:*

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

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 4.1 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.

4.2 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 5). The first one concerns the exactness in the spiked covariance model. That is, the oracle Z^* as defined by equation (17), 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 4.2 *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 (17).*

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 define localized subsets of our model. Then, later, based on the results from Section 3 we will compute the Rademacher complexities of these localized subsets and then the local complexity fixed points as introduced in Section 2. 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 4.3).

Lemma 4.3 *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 $P\mathcal{L}_Z = \langle \Sigma, Z^* - Z\rangle \geqslant (\theta/2)\|Z^* - Z\|_2^2$.*

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.

4.3 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 2.2.2, where the ℓ_1-norm is used as regularization norm:

$$\hat{Z}_\lambda^{\mathrm{RERM}} \in \underset{Z\in\mathcal{C}}{\operatorname{argmin}} \left(P_N \ell_Z + \lambda\|Z\|_1\right), \quad \text{where} \quad \mathcal{C} := \left\{Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \mathrm{Tr}(Z) = 1\right\} \tag{19}$$

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 2.2.2 to provide statistical guarantees on $\hat{Z}_\lambda^{\mathrm{RERM}}$.

Lemma 4.3 shows that, for any value of $\rho > 0$ and $\delta \in (0,1)$, Assumption 2.10 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 2.11. This is the purpose of the following Lemma (the proof is given in section 5.3.3).

Lemma 4.4 *Let $A > 0$, $\delta \in (0,1)$, and define $r^*(.) := r^*_{\mathrm{RERM,G}}(A, ., \delta)$. If $\rho \geqslant 10k\sqrt{r^*(\rho)}$, then ρ satisfies the A-sparsity equation from Definition 2.11.*

The last step is to compute the local complexity fixed point of Definition 2.9, which is what we are working on below.

Lemma 4.5 *Grant Assumption 3.1 with $t = \log(ed/10k)$. Suppose that β^* is k-sparse, with $k \leqslant ed/200$. Let $A = 2/\theta$ and assume that $N \geqslant 3\log(ed/10k)$. Then there exists an absolute constant $b > 0$ such that, defining:*

$$\rho^* := 200bAk^2\sqrt{\frac{1}{N}\log\left(\frac{ed}{k}\right)} \text{ and } r^*(\rho) := bA\sqrt{\frac{\rho^2}{N}\log\left(\frac{b^2A^2(ed)^4}{N\rho^2}\right)}, \tag{20}$$

one has $r^*_{\mathrm{RERM,G}}(A, \rho^*, 10k/ed) \leqslant r^*(\rho^*)$ and ρ^* satisfies the A-sparsity equation from Definition 2.11.

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

Theorem 4.6 *Grant Assumption 3.1 with $t = \log(ed/10k)$. Suppose that β^* is k-sparse, with $k \leqslant ed/200$. Assume that $N \geqslant 3\log(ed/10k)$ and that λ satisfies the following inequalities:*

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

where b is the absolute constant introduced in Lemma 4.5 above. Let $C = 40b$. Then, with probability at least $1 - 10k/ed$, it holds true that:

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

and

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

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 4.7 *Let $\hat{\beta} \in \mathbb{R}^d$ be a leading unit length eigenvector of $\hat{Z}_\lambda^{\text{RERM}}$. Under the conditions of Theorem 4.6, there exists an absolute constant $D > 0$ such that with probability at least $1 - 10k/ed$:*

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

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 3.1, 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. (2016a) 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. (2016a); this improvement is of the same order as the one obtained for the LASSO in Bellec et al. (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. (2016a) and the $\log(ed/k)$ obtained in Theorem 4.6. Our result also extend the scope of Theorem 4.1 to heavy-tailed data since we only require the existence of $\log d$ moments. However, to get this improvement for the Lasso type estimator (22), one needs to choose λ depending on k in (21), 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 4.6 but for 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.

4.4 $SLOPE$ regularized ERM estimator

In this section, we study a regularized ERM estimator of Z^* with the $SLOPE$ norm (introduced in Section 3.2, 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 $p \leqslant q$, $b_{pq}(t) =: \sqrt{\log(ed^2/pq) + t}$, 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:

$$\hat{Z}_{SLOPE}^{\text{RERM}} \in \underset{Z\in\mathcal{C}}{\text{argmin}} \left(P_N \ell_Z + \lambda\|Z\|_{SLOPE}\right) \text{ for } \mathcal{C} := \{Z \in \mathbb{R}^{d\times d} : Z \succeq 0, \text{Tr}(Z) = 1\} \qquad (22)$$

and a regularization parameter $\lambda > 0$ to be chosen later. This puts us in condition to use the results of Section 2.2.2 to provide statistical guarantees on $\hat{Z}_{SLOPE}^{\text{RERM}}$.

As before, the essence of Lemma 4.3 in this context is that, for any value of $\rho > 0$ and $\delta \in (0, 1)$, Assumption 2.10 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 4.8 *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 $\rho \geqslant 10\Gamma_k(t)\sqrt{r^*_{\mathrm{RERM,G}}(A, \rho, \delta)}$, then ρ satisfies the A-sparsity equation from Definition 2.11.*

Following the path traced by our methodology, all that remains is to calculate the complexity fixed-point parameter $r^*_{\mathrm{RERM,G}}(A, \rho, \delta)$ defined as

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

The next Lemma gives us an upper bound for $r^*_{\mathrm{RERM,G}}(A, \rho, \delta)$, when ρ satisfies the sparsity equation of Definition 2.11.

Lemma 4.9 *Grant Assumption 3.3 for $t = 2\log(ed^2/k^2)$. Suppose that β^* is k-sparse, with $k \leqslant d/(e^2\log(d))$. Let $A > 0$, and assume that $N \geqslant 3\log(ed^2)$. Then, there exists an absolute constant $b > 0$ such that, defining:*

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

one has $r^*_{\mathrm{RERM,G}}(A, \rho^*, 2k^2/(ed^2)) \leqslant r^*$ and ρ^* satisfies the A-sparsity equation rom Definition 2.11, where $\Gamma_k^* = \Gamma_k(2\log(ed^2/k^2))$ is the quantity introduced in Lemma 4.8.

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

Theorem 4.10 *Grant Assumption 3.3 for $t = 2\log(ed^2/k^2)$. Suppose that β^* is k-sparse, with $k \leqslant \min\left(d/(e^2\log(d)), (e/140\sqrt{2})^2 d\right)$. Assume that $N \geqslant 3\log(ed^2)$ and that λ satisfies the following inequalities:*

$$\frac{10b}{21\sqrt{N}} < \lambda < \frac{2b}{3\sqrt{N}}, \tag{23}$$

where b is the constant previously defined in Lemma 4.9. Then there exist an absolute constants $C_1 > 0$ such that one has with probability at least $1 - 2k^2/(ed^2)$:

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

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

and

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

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 4.11 *Let $\hat{\beta} \in \mathbb{R}^d$ be a leading unit-eigen vector of $\hat{Z}_\lambda^{\text{RSLOPE}}$. Under the conditions of Theorem 4.10, there exists an absolute constant $C > 0$ such that with probability at least $1 - 2k^2/ed^2$:*

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

Here again, we obtain a rate of convergence of magnitude $\sqrt{(1/N\theta^2)\log(ed^2/k^2)}$, 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^*.

4.5 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 2.13: the dataset $\{X_1, \dots, X_N\}$ 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

$$\hat{Z}_{K,\lambda}^{RMOM} \in \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \sup_{Z' \in \mathcal{C}} \left(\text{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\|_1 - \|Z'\|_1)\right) \tag{24}$$

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

In what follows, we provide some statistical guarantees on $\hat{Z}_{K,\lambda}^{RMOM}$ based on Theorem 2.32 which is our general result for regularized minmax MOM estimators for a general G function used for localization. Here, following Lemma 4.3, we will use $G : Z \to \|Z\|_2^2$ (and $A = 2/\theta$) 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 $r^*_{\text{RMOM},G}(\gamma, \rho)$ from Definition 2.29. But before that we find a sufficient condition on a radius ρ so that it satisfies the sparsity equation from Definition 2.23.

Lemma 4.12 *Consider $\gamma > 0$. If $\rho > 0$ is such that $\rho \geqslant 10k\sqrt{2/\theta}r^*_{\text{RMOM},G}(\gamma, \rho)$, then ρ satisfies the sparsity equation from Definition 2.23.*

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 $r^*_{\text{RMOM},G}(\gamma, \rho)$ as introduced in Definition 2.29, after which, thanks to Theorem 2.24, we will be able to provide some statistical bounds on $\hat{Z}_{K,\lambda}^{RMOM}$.

Lemma 4.13 *Grant assumption 3.1 for $t = 1$. Suppose that β^* is k-sparse, for some $k \in [d]$. Assume that $N \geqslant 2\log(ed/k)+1$ and that $\theta \leqslant k$. 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:*

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

and

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

one has $r^*_{\mathrm{RMOM,G}}(\gamma,\rho^*(\gamma)) \leqslant r^*(\gamma,\rho^*(\gamma))$ and $\rho^*(\gamma)$ satisfies the sparsity equation from Definition 2.23. The values of B and D are explicited in Section 5.3.12.

We are now ready to state our main result about the ℓ_1-regularized MOM estimator (24) for the sparse PCA problem.

Theorem 4.14 *Grant assumption 3.1 for $t = 1$. Suppose that β^* is k-sparse, for some $k \in [d]$. Assume that $N \geqslant 2\log(ed/k)+1$ and let K be a divisor of N such that $K \geqslant 100|\mathcal{O}|$. Let $\gamma = 1/32000$ and $\lambda := 11r^*(\gamma, 2\rho^*(\gamma))/(40\rho^*(\gamma))$, where $r^*(.,.)$ and $\rho^*(.)$ are defined in Lemma 4.13 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:*

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

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

and

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

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

Corollary 4.15 *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 4.14, there exists a universal constant $D > 0$ such that with probability at least $1-\exp(-72K/625)$:*

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

When $K \leqslant k^2\log(ed/k)$, we get a rate of convergence of magnitude $\sqrt{k^2/(N\theta^2)\log(ed/k)}$, 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 similar bound with a choice of λ independent of k.

Remark 5 *Our final remark deals with the implementation of an algorithm for an approximate construction of $\hat{\beta}$ from Corollary 4.15. The aim of this algorithm is to construct a sparse first principal component given a set of heavy-tailed data that may have been corrupted by an adversary.*

Using a similar approach to the MOM version of the projected sub-gradient descent algorithm introduced in Lecué and Lerasle (2020), we first derive a pseudo-algorithm for (24). We will then take a top eigenvector to the solution provided by such algorithm to get a (robust) sparse first principal component. But first, let us focus on the construction of an approximate solution to $\hat{Z}^{RMOM}_{K,\lambda}$. We slightly modify the original version of the MOM version of the projected gradient descent algorithm from Lecué and Lerasle (2020) with respect to the following two points:

a) *instead of designing an alternating ascent/descent minmax MOM algorithm for (24), we will describe a MOM version of the projected (sub)gradient descent algorithm for the minimization problem $\min_{Z\in\mathcal{C}}\left(\mathrm{MOM}_K(\ell_Z) + \lambda\|Z\|_1\right)$,*

b) *instead of selecting the median block of data and making a gradient descent over it at every iterations, we consider a larger set of data by taking all the 'inter-quartile blocks' of data and perform a gradient descent over it as in Depersin and Lecué (2022).*

Before writting the pseudo-code, we recall and introduce several notation. The dataset $\{X_1, \dots, X_N\}$ is split into K equal size blocks of data indexed by $(B_k)_k$ forming an equipartition of $[N]$. On each block, an empirical covariance matrix is constructed $\overline{(XX^\top)}_k = |B_k|^{-1}\sum_{i\in B_k} X_i X_i^\top$. The next function is using the K bucketed empirical covariance matrices $(\overline{(XX^\top)}_k)_k$ as input data: for all $Z \in \mathbb{R}^{d\times d}$,

$$f(Z) = \frac{-1}{|I_K|}\sum_{k\in I_K}\left\langle \overline{(XX^\top)}_k, Z\right\rangle^*_{(k)}$$

where if $(a_k)_k = (\langle\overline{(XX^\top)}_k, Z\rangle)_k$ then $(\langle\overline{(XX^\top)}_k, Z\rangle^*_{(k)})_k$ are the rearrangement of $(a_k)_k$ such that $a^*_{(1)} \leqslant \dots \leqslant a^*_{(K)}$ (this is the rearrangement of the values a_k's themselves and not of their absolute values) and

$$I_K = \left[\frac{K+1}{4}, \frac{3(K+1)}{4}\right] = \left\{\frac{K+1}{2} \pm k : k = 0, 1, \cdots, \frac{K+1}{4}\right\}$$

is the inter-quartiles interval - without loss of generality we assume that $K+1$ can be divided by 4. In other words, $f(Z)$ is the average sum over all inter-quartile values of the vector $(\langle\overline{(XX^\top)}_k, Z\rangle)_{k\in[K]}$. Note that we have taken quartiles but we could also have considered other quantiles; for instance a 95% coverage of the data in $\cup_{k\in I_K} B_k$ may also be considered. We denote by $\mathrm{proj}_{\mathcal{C}}$ the projection over $\mathcal{C} = \left\{Z \in \mathbb{R}^{d\times d} : 0 \preceq Z \preceq I_d, \mathrm{Tr}(Z) = 1\right\}$. The *next algorithm is a MOM version of the projected sub-gradient descent algorithm for the minimization problem*

$$\min_{Z \in \mathcal{C}} \left(f(Z) + \lambda \|Z\|_1\right).$$

```
input : the data $X_1, \ldots, X_N$, a number $K$ of blocks, a regularization parameter $\lambda$,
        a decreasing steps size sequences $(\eta_t)_t \subset \mathbb{R}_+^*$ and $\epsilon > 0$ a stopping
        parameter
output: A robust and sparse first principal component
Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$ at random
Construct the $K$ empirical covariance matrices $\overline{(XX^\top)}_k = (K/N) \sum_{i \in B_k} X_i X_i^\top$
Compute the coordinate-wise median-of-means
    $$Z^{(0)} = \left(\mathrm{Med}\left((\overline{(XX^\top)}_k)_{pq} : k \in [K]\right)\right)_{1 \leqslant p,q \leqslant d}$$
while $\|Z^{(t)} - Z^{(t+1)}\|_1 \geqslant \epsilon$ do
    Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$ at random
    Construct the $K$ empirical covariance matrices $\overline{(XX^\top)}_k = (K/N) \sum_{i \in B_k} X_i X_i^\top$
    Find the inter-quartile block numbers $k_1, \ldots, k_{(K+1)/2} \in [K]$ such that
        $$f(Z^{(t)}) = \frac{-1}{|I_K|} \sum_{j=1}^{(K+1)/2} \langle \overline{(XX^\top)}_{k_j}, Z^{(t)} \rangle.$$
    Construct $G^{(t)}$ a subgradient of $\|\cdot\|_1$ at $Z^{(t)}$ and the descent direction
        $$\nabla^{(t+1)} = \frac{-1}{|I_K|} \sum_{j=1}^{(K+1)/2} \overline{(XX^\top)}_{k_j} + \lambda G^{(t)}.$$
    Update $Z^{(t+1)} \leftarrow \mathrm{proj}_{\mathcal{C}}(Z^{(t)} - \eta_t \nabla^{(t+1)})$.
end
Return a top singular vector of $Z^{(t+1)}$.
```

Algorithm 1: A MOM projected sub-gradient descent algorithm for robust and sparse PCA.

Several variations of Algorithm1 may be considered. In particular, one can use proximal operators in place of sub-gradients. We refer the interested reader to Lecué and Lerasle (2020) for more examples of MOM versions of classical regularized algorithms.

5. Proofs

All the proofs from the previous sections – general excess risk and estimation bounds as well as applications – are gathered in this section.

5.1 Proofs of section 2

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 $Z \in \mathcal{C}$.

5.1.1 Proof of Theorem 2.12

Let $\delta \in (0,1)$. Let $A > 0$ and $\rho^* > 0$ be such that Assumption 2.10 holds, and assume that $\rho^* > 0$ satisfies the A-sparsity equation from Definition 2.11. Let $\gamma := 1/(3A)$. In the rest of the proof, we write $r^*(.)$ for $r^*_{\mathrm{RERM,G}}(A, ., \delta)$. Let us define

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

Consider the following event:

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

By definition of $r^*(.)$, Ω holds with probability at least $1-\delta$. Let us now prove the statistical bounds announced in Theorem 2.12 on the event Ω.

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

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

where (i) holds since $P_N\mathcal{L}_{\hat{Z}}^\lambda \leqslant 0$ by definition of $\hat{Z}$ and (ii) holds because of the choice of λ given in (9).

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

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 $G(Z - Z^*) \leqslant r^*(\rho^*)$ or $\|Z - Z^*\| \leqslant \rho^*$ 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

$$\|Z\| - \|Z^*\| = \|\alpha Z_0 - (\alpha-1)Z^*\| - \|Z^*\| \geqslant \alpha\|Z_0\| - (\alpha-1)\|Z^*\| - \|Z^*\| \geqslant \alpha(\|Z_0\| - \|Z^*\|)$$

and so

$$P_N\mathcal{L}_Z^\lambda = P_N\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \geqslant \alpha P_N\mathcal{L}_{Z_0} + \lambda\alpha(\|Z_0\| - \|Z^*\|) = \alpha P_N\mathcal{L}_{Z_0}^\lambda. \tag{25}$$

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 $P_N\mathcal{L}_Z^\lambda > \alpha P_N\mathcal{L}_{Z_0}^\lambda$. Hence, we only have to show that $Z \to P_N\mathcal{L}_Z^\lambda$ 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 $G(Z - Z^*) \leqslant r^*(\rho^*)$, or $\|Z_0 - Z^*\| \leqslant \rho^*$ and $G(Z - Z^*) = r^*(\rho^*)$.

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

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

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:

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

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

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

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

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

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

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

where the last inequality is due to the choice of λ given in (9).

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

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

But we know from Assuption 2.10 that $P\mathcal{L}_{Z_0} \geqslant A^{-1}G(Z_0 - Z^*)$, and on Ω we have $(P - P_N)\mathcal{L}_{Z_0} \leqslant \gamma r^*(\rho^*)$. Then we get:

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

where the last inequality comes from the choice of λ given in (9).

Then, we proved that $P_N\mathcal{L}_{Z_0}^{\lambda} > 0$ for any $Z_0 \in \partial(\mathcal{B})$ and as we said before, this implies that $P_N\mathcal{L}_{Z}^{\lambda}$ is positive over $\mathcal{C}\backslash\mathcal{B}$. Since $P_N\mathcal{L}_{\hat{Z}}^{\lambda} < 0$ we conclude that on Ω, $\hat{Z}$ necessarily belongs to $\mathcal{B}$, which proves the bounds announced in Theorem 2.12.

■

5.1.2 Proof of Theorem 2.15

The proof of this theorem is broken down into two steps. First, we identify an event Ω on which the estimator $\hat{Z}_{\mathrm{K}}^{\mathrm{MOM}}$ 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 $\hat{Z}_{K}^{\mathrm{MOM}}$. Let $\gamma = 1/6400$, and consider the set $\mathcal{C}_\gamma := \left\{Z \in \mathcal{C} : P\mathcal{L}_Z \leqslant (r^*)^2\right\}$. Define the event Ω_K as follows:

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

We start with showing that on Ω_K, the estimator $\hat{Z}$ satisfies the excess risk bound announced in Theorem 2.15.

Lemma 5.1 *On the event Ω_K, $P\mathcal{L}_{\hat{Z}} \leqslant r^*$.*

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 $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leqslant (r^*)^2/4$, that is $P_{B_k}\mathcal{L}_{Z_0} \geqslant P\mathcal{L}_{Z_0} - (r^*)^2/4 = (3/4)(r^*)^2$ and so $P_{B_k}\mathcal{L}_Z = \alpha P_{B_k}\mathcal{L}_{Z_0} \geqslant \alpha(3/4)(r^*)^2$ because $\alpha > 1$. This holds on strictly more than half of the blocks B_k, therefore $\mathrm{Med}(-P_{B_k}\mathcal{L}_Z : k \in [K]) \geqslant -(3/4)(r^*)^2$ and this holds for all $Z \in \mathcal{C}\backslash\mathcal{C}_\gamma$, hence, we have

$$\sup_{Z \in \mathcal{C}\backslash\mathcal{C}_\gamma} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leqslant -(3/4)(r^*)^2. \tag{26}$$

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

$$\sup_{Z \in \mathcal{C}_\gamma} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leqslant (r^*)^2/4. \tag{27}$$

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

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

that is, $\mathrm{MOM}_K(\ell_{Z^*} - \ell_{\hat{Z}}) \geqslant -(1/4)(r^*)^2 > -(3/4)(r^*)^2$. From (26) we conclude that, necessarily, $\hat{Z} \in \mathcal{C}_\gamma$, that is, $P\mathcal{L}_{\hat{Z}} \leqslant (r^*)^2$. ■

At this point, we proved that on the event Ω_K, the estimator $\hat{Z}$ satisfies the statistical bounds announced in Theorem 2.15. Now it remains to prove that Ω_K holds with high probability.

Lemma 5.2 *Assume that* $|\mathcal{O}| \leqslant K/100$. *Then* Ω_K *holds with probability at least* $1 - \exp(-72K/625)$.

Proof Let $\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t \geqslant 1\}} + 2(t - (1/2))\mathbb{1}_{\{1/2 \leqslant t \leqslant 1\}}$, so that for any $t \in \mathbb{R}$, $\mathbb{1}_{\{t \geqslant 1\}} \leqslant \phi(t) \leqslant \mathbb{1}_{\{t \geqslant 1/2\}}$. 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 $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - P)\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := (K/N)\sum_{i \in B_k} \mathcal{L}_Z(\widetilde{X}_i)$. Let $\psi(Z) = \sum_{k \in [K]} \mathbb{1}_{\{|F_Z(W_k)| \leqslant (r^*)^2/4\}}$. 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 $Z \in \mathcal{C}_\gamma$,

$$\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} \leqslant \frac{49K}{100}. \tag{28}$$

Indeed, consider $Z \in \mathcal{C}_\gamma$ such that (28) holds. Then, there exist at least $(1 - 49/100)K = (51/100)K$ blocks B_k on which $|F_Z(\widetilde{W}_k)| \leqslant (r^*)^2/4$. On the other hand, we know that $|\mathcal{O}| \leqslant K/100$, 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 $\sum_{k \in [K]} \mathbb{1}_{\{|F_Z(W_k)| \leqslant (r^*)^2/4\}} > K/2$, that is $\psi(Z) > K/2$, if (28) holds.

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

$$\begin{aligned}
&\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} \qquad (29)\\
&= \sum_{k \in [K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right)\right]\\
&= \sum_{k \in [K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{4}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right\}}\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right)\\
&\leqslant \sum_{k \in [K]} \left(\Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\left[\Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right)\\
&\leqslant \sup_{Z \in \mathcal{C}_\gamma} \left(\sum_{k \in [K]} \Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\left[\Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r^*)^2}{8}\right).
\end{aligned} \tag{30}$$

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

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

Plugging that into (29), we get:

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r^*)^2}{4}\right\}} \leqslant \frac{K}{200} + \sup_{Z\in\mathcal{C}_\gamma}\left(\sum_{k\in[K]}\Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\left[\Phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right). \tag{31}$$

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

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

Let now $\epsilon_1, \dots, \epsilon_K$ be Rademacher variables independent from the $\widetilde{X}_i$'s. By the symmetrization Lemma, we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\right] \leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]. \tag{33}$$

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:

$$\begin{aligned}\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right] &\leqslant 8\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k\frac{F_Z(\widetilde{W}_k)}{(r^*)^2}\right]\\ = \frac{8}{(r^*)^2}\mathbb{E}\left[\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}} - P)\mathcal{L}_Z\right].\end{aligned} \tag{34}$$

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:

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

Combining this with (32), (33) and (34), we finally get that, with probability at least $1-\exp(-72K/625)$

$$\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{4|F_Z(\widetilde{W}_k)|}{(r^*)^2}\right)\right]\leqslant\left(\frac{12}{25}+32\gamma\right)K. \tag{35}$$

Plugging that into (31), we conclude that with probability at least $1-\exp(-72K/625)$, one has

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

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.

■

5.1.3 Proof of Theorem 2.18.

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 $\hat{Z}_K^{\text{MOM}}$ and r^* for $r^*_{\text{MOM},L_2}(\gamma)$ with $\gamma=1/3200$. Let $0<A<1$ be such that Assumption 2.17 holds. We define $\nu=A^2/\gamma$, $\tau=(2A)^{-1}$, $C_{K,A}=\max\big((r^*)^2,\nu K/N\big)$ and $\mathcal{B}_{K,A}:=\left\{Z\in\mathcal{C}:\|Z-Z^*\|_{L_2}\leqslant\sqrt{C_{K,A}}\right\}$ - where the L_2-norm is defined as $Z\to\|Z\|_{L_2}=\mathbb{E}[\big\langle\tilde{X},Z\big\rangle^2]^{1/2}$. We consider the following event:

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

We show in the next three lemmas that, on Ω_K, $\hat{Z}$ satisfies the statistical bounds announced in Theorem .2.18. Then the fourth lemma will prove that Ω_K holds with large probability, the one announced in Theorem .2.18.

Lemma 5.3 *If there exists $\eta>0$ such that:*

$$\sup_{Z\in\mathcal{C}\backslash\mathcal{B}_{K,A}}\text{MOM}_K(\ell_{Z^*}-\ell_Z)<-\eta\quad\textit{and}\quad\sup_{Z\in\mathcal{B}_{K,A}}\text{MOM}_K(\ell_{Z^*}-\ell_Z)\leqslant\eta \tag{36}$$

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

Proof Assume that (36) holds. Then:

$$\inf_{Z\in\mathcal{C}\backslash\mathcal{B}_{K,A}}\text{MOM}_K(\ell_Z-\ell_{Z^*})>\eta. \tag{37}$$

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

$$T_K(Z^*)=\max\left(\sup_{Z\in\mathcal{C}\backslash\mathcal{B}_{K,A}}\text{MOM}_K(\ell_{Z^*}-\ell_Z),\sup_{Z\in\mathcal{B}_{K,A}}\text{MOM}_K(\ell_{Z^*}-\ell_Z)\right)\leqslant\eta. \tag{38}$$

By definition of $\hat{Z}$, we have $T_K(\hat{Z}) = \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_Z) \leqslant T_K(Z^*) \leqslant \eta$. But by (37), any $Z \in \mathcal{C} \backslash \mathcal{B}_{K,A}$ satisfies:

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

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

Lemma 5.4 *Assume that $K \geqslant 100|\mathcal{O}|$. Then on Ω_K, (36) 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:

$$P_{B_k}\mathcal{L}_Z = \alpha P_{B_k}\mathcal{L}_{Z_0}. \tag{39}$$

As $Z_0 \in \mathcal{B}_{K,A}$, on Ω_K there exist strictly more than $K/2$ blocks on which $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leqslant \tau C_{K,A}$. Moreover, since $\|Z_0 - Z^*\|_{L_2} = \sqrt{C_{K,A}}$, we get from Assumption 2.17 that $P\mathcal{L}_{Z_0} \geqslant A^{-1}\|Z_0 - Z^*\|_{L_2}^2 = A^{-1}C_{K,A}$. Then, on these blocks, $P_{B_k}(\ell_{Z_0} - \ell_{Z^*}) \geqslant P\mathcal{L}_{Z_0} - \tau C_{K,A} \geqslant (A^{-1} - \tau)C_{K,A}$, which implies that $P_{B_k}(\ell_{Z^*} - \ell_{Z_0}) \leqslant -(A^{-1} - \tau)C_{K,A} \leqslant -\tau C_{K,A}$, since we have $\tau = (2A)^{-1}$. From (39) we conclude that, on Ω_K, there exist srictly more than $K/2$ blocks B_k on which $P_{B_k}(\ell_{Z^*} - \ell_Z) \leqslant -\alpha\tau C_{K,A} \leqslant -\tau C_{K,A}$, since $\alpha \geqslant 1$. This is true for all $Z \in \mathcal{C} \backslash \mathcal{B}_{K,A}$; in other words, we have

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

Moreover, on Ω_K, for any $Z \in \mathcal{B}_{K,A}$, there exist stricly more than $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leqslant \tau C_{K,A}$, so that $P_{B_k}(\ell_Z - \ell_{Z^*}) \geqslant -\tau C_{K,A} + P(\ell_Z - \ell_{Z^*}) \geqslant -\tau C_{K,A}$, since $P(\ell_Z - \ell_{Z^*}) \geqslant 0$ by definition of Z^*. Then, we have $P_{B_k}(\ell_{Z^*} - \ell_Z) \leqslant \tau C_{K,A}$ on stricly more than $K/2$ blocks, which implies that $\mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leqslant \tau C_{K,A}$. This being true for any $Z \in \mathcal{B}_{K,A}$, we conclude that (36) holds with $\eta = \tau C_{K,A}$. ■

Lemma 5.5 *Grant Assumption 2.17 and assume that $K \geqslant 100|\mathcal{O}|$. On Ω_K, $P\mathcal{L}_{\hat{Z}} \leqslant 2\tau C_{K,A}$.*

Proof Assume that Ω_K holds. From Lemmas 5.3 and 5.4 , $\|\hat{Z} - Z^*\|_{L_2}^2 \leqslant C_{K,A}$, that is $\hat{Z} \in \mathcal{B}_{K,A}$. Therefore, on strictly more than $K/2$ blocks B_k, we have $\left|(P_{B_k} - P)\mathcal{L}_{\hat{Z}}\right| \leqslant \tau C_{K,A}$, and then on these blocks:

$$P\mathcal{L}_{\hat{Z}} \leqslant P_{B_k}\mathcal{L}_{\hat{Z}} + \tau C_{K,A}. \tag{40}$$

In addition, by definition of $\hat{Z}$ and (38) (for $\eta = \tau C_{K,A}$):

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

which implies the existence of $K/2$ blocks (at least) on which:

$$P_{B_k}\mathcal{L}_{\hat{Z}} \leqslant \tau C_{K,A} \tag{41}$$

As a consequence, there exist at least one block B_k on which (40) and (41) holds simultaneously. On this block, we have: $P\mathcal{L}_{\hat{Z}} \leqslant \tau C_{K,A} + \tau C_{K,A} = 2\tau C_{K,A}$, 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 2.18. In the final lemma, we show that Ω_K holds with high probability.

Lemma 5.6 *Assume that $|\mathcal{O}| \leqslant K/100$. Then Ω_K holds with probability at least* $1 - \exp(-72K/625)$.

Proof Let $\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t \geqslant 1\}} + 2(t - 1/2)\mathbb{1}_{\{1/2 \leqslant t \leqslant 1\}}$, so that for any $t \in \mathbb{R}$, $\mathbb{1}_{\{t \geqslant 1\}} \leqslant \phi(t) \leqslant \mathbb{1}_{\{t \geqslant 1/2\}}$. 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 $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - P)\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := (K/N)\sum_{i \in B_k} \mathcal{L}_Z(\widetilde{X}_i)$. Let $\psi(Z) = \sum_{k \in [K]} \mathbb{1}_{\{|F_Z(W_k)| \leqslant \tau C_{K,A}\}}$. 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 $Z \in \mathcal{B}_{K,A}$,

$$\sum_{k \in [K]} \mathbb{1}_{\{|F_Z(\widetilde{W}_k)| > \tau C_{K,A}\}} \leqslant \frac{49K}{100}. \tag{42}$$

Indeed, consider $Z \in \mathcal{C}$ such that (42) holds. Then, there exist at least $(1 - 49/100)K = (51/100)K$ blocks B_k on which $|F_Z(\widetilde{W}_k)| \leqslant \tau C_{K,A}$. On the other hand, we know that $|\mathcal{O}| \leqslant K/100$, 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 $\sum_{k \in [K]} \mathbb{1}_{\{|F_Z(W_k)| \leqslant \tau C_{K,A}\}} > K/2$, that is $\psi(Z) > K/2$, if (42) holds.

Then, we only have to show that (42) 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:

$$\begin{aligned}
&\sum_{k \in [K]} \mathbb{1}_{\{|F_Z(\widetilde{W}_k)| > \tau C_{K,A}\}} \\
&= \sum_{k \in [K]} \left[\mathbb{1}_{\{|F_Z(\widetilde{W}_k)| > \tau C_{K,A}\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right)\right] \\
&= \sum_{k \in [K]} \left(\mathbb{1}_{\{|F_Z(\widetilde{W}_k)| > \tau C_{K,A}\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right\}}\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) \\
&\leqslant \sum_{k \in [K]} \left(\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\left[\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) \\
&\leqslant \sup_{Z \in \mathcal{B}_{K,A}} \left(\sum_{k \in [K]} \Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\left[\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right).
\end{aligned} \tag{43}$$

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

$$\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{\tau C_{K,A}}{2}\right) \leqslant \frac{4}{(\tau C_{K,A})^2}\mathbb{E}\left[F_Z(\widetilde{W}_k)^2\right] = \frac{4}{(\tau C_{K,A})^2}\left(\frac{K}{N}\right)\mathrm{Var}(\mathcal{L}_Z(\widetilde{X}))$$
$$\leqslant \frac{4}{(\tau C_{K,A})^2}\left(\frac{K}{N}\right)\mathbb{E}[\mathcal{L}_Z(\widetilde{X})^2] = \frac{4}{(\tau C_{K,A})^2}\frac{K}{N}\|Z - Z^*\|_{L_2}^2 \leqslant \frac{4}{(\tau C_{K,A})^2}\frac{K}{N}C_{K,A} \leqslant \frac{4}{\tau^2 \nu} = \frac{1}{200}.$$

Plugging the latter result into (43), we get:

$$\sum_{k\in[K]} \mathbb{1}_{\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\}} \leqslant \frac{K}{200} + \sup_{Z\in\mathcal{B}_{K,A}}\left(\sum_{k\in[K]}\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\left[\Phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right). \tag{44}$$

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

$$\sum_{k\in[K]}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)$$
$$\leqslant \frac{12}{25}K + \mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right].$$

Let now $\epsilon_1, \dots, \epsilon_K$ be Rademacher variables independent from the $\widetilde{X}_i$'s. By the symmetrization Lemma, we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right) - \mathbb{E}\left[\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]\right] \leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right]$$

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:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{|F_Z(\widetilde{W}_k)|}{\tau C_{K,A}}\right)\right] \leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\frac{F_Z(\widetilde{W}_k)}{\tau C_{K,A}}\right]$$
$$\leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\frac{(\widetilde{P_{B_k}} - P)\mathcal{L}_Z}{\tau C_{K,A}}\right].$$

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

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{k\in[K]}\epsilon_k\frac{(\widetilde{P_{B_k}} - P)\mathcal{L}_Z}{C_{K,A}}\right] \leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\frac{K}{N}\sum_{i=1}^{N}\sigma_i\frac{\mathcal{L}_Z(\widetilde{X}_i)}{C_{K,A}}\right].$$

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:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^N \sigma_i \frac{\mathcal{L}_Z(\widetilde{X}_i)}{C_{K,A}}\right] \leqslant \frac{1}{C_{K,A}}\gamma(r^*)^2 N = \gamma N.$$

In the second case, we decompose the supremum into two parts:

$$\begin{aligned}&\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^N \sigma_i \mathcal{L}_Z(\widetilde{X}_i)\\ &= \max\left(\sup_{Z\in\mathcal{B}_{K,A}:\|Z-Z^*\|_{L_2}\leqslant r^*}\sum_{i=1}^N \sigma_i\mathcal{L}_Z(\widetilde{X}_i), \sup_{Z\in\mathcal{B}_{K,A}:r^*\leqslant\|Z-Z^*\|_{L_2}\leqslant\sqrt{\frac{\nu K}{N}}}\sum_{i=1}^N \sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right).\end{aligned}$$

Let $Z \in \mathcal{B}_{K,A}$ be such that $r^* \leqslant \|Z - Z^*\|_{L_2} \leqslant \sqrt{\frac{\nu K}{N}}$. 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 $\kappa \geqslant 1$, so that $\kappa = \frac{\|Z-Z^*\|_{L_2}}{\|Z_0-Z^*\|_{L_2}} \leqslant \sqrt{\frac{\nu K}{N}}\frac{1}{r^*}$. Moreover, we have by linearity of $\mathcal{L}$ that $\mathcal{L}_{Z_0} = \kappa\mathcal{L}_Z$. Therefore, we obtain

$$\begin{aligned}\sup_{Z\in\mathcal{B}_{K,A}:r^*\leqslant\|Z-Z^*\|_{L_2}\leqslant\sqrt{\frac{\nu K}{N}}}\sum_{i=1}^N \sigma_i\mathcal{L}_Z(\widetilde{X}_i) &\leqslant \sup_{1\leqslant\kappa\leqslant\frac{1}{r^*}\sqrt{\frac{\nu K}{N}}}\ \sup_{Z_0\in\mathcal{B}_{K,A}:\|Z_0-Z^*\|_{L_2}\leqslant r^*}\sum_{i=1}^N \sigma_i\kappa\mathcal{L}_{Z_0}(\widetilde{X}_i)\\ &= \sqrt{\frac{\nu K}{N}}\frac{1}{r^*}\sup_{Z_0\in\mathcal{B}_{K,A}:\|Z_0-Z^*\|_{L_2}\leqslant r^*}\sum_{i=1}^N \sigma_i\mathcal{L}_{Z_0}(\widetilde{X}_i).\end{aligned}$$

Since $C_{K,A} = \nu K/N \geqslant (r^*)^2$, we get, using the definition of r^*:

$$\begin{aligned}&\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^N \sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right] \leqslant \sqrt{\frac{\nu K}{N}}\frac{1}{r^*}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}:\|Z-Z^*\|_{L_2}\leqslant r^*}\sum_{i=1}^N \sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right]\\ &\leqslant \sqrt{\frac{\nu K}{N}}\frac{1}{r^*}\gamma(r^*)^2 N \leqslant C_{K,A}\gamma N.\end{aligned}$$

Finally, we get that whatever the value of $C_{K,A}$ is:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_{K,A}}\sum_{i=1}^N \sigma_i \frac{\mathcal{L}_Z(\widetilde{X}_i)}{C_{K,A}}\right] \leqslant \gamma N.$$

Combining all these inequalities, we finally get that, with probability at least $1-\exp(-12K/625)$, for all $Z \in \mathcal{B}_{K,A}$,

$$\sum_{k\in[K]}\mathbb{1}_{\{|F_Z(\widetilde{W}_k)|>\tau C_{K,A}\}} \leqslant \frac{12}{25}K + \frac{1}{200}K + \frac{8\gamma}{\tau}K \leqslant \frac{49}{400}K$$

From our choice of parameters. This concludes the proof. ■

5.1.4 Proof of Theorem 2.21

The proof of this theorem follows the same lines as the one of the last Theorem 2.15 and 2.18: 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 $\hat{Z}_K^{\rm MOM}$ and r^* for $r^*_{\rm MOM,G}(\gamma)$ for $\gamma = 1/6400$. Consider A and $G : H \to \mathbb{R}$ such that Assumption 2.20 holds. Define

$$\mathcal{C}_{\gamma,G} := \left\{ Z \in \mathcal{C} : G(Z - Z^*) \leqslant (r^*)^2 \right\}.$$

We consider the following event:

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

We first show that on the event Ω_K, $\hat{Z}$ satisfies the statistical bounds announced in Theorem 2.21.

Lemma 5.7 *If there exists $\eta > 0$ such that*

$$\sup_{Z \in \mathcal{C} \backslash \mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) < -\eta \quad \textit{and} \quad \sup_{Z \in \mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leqslant \eta \tag{45}$$

then $G(\hat{Z} - Z^*) \leqslant (r^*)^2$.

Proof Assume that (45) holds. Then:

$$\inf_{Z \in \mathcal{C} \backslash \mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) > \eta. \tag{46}$$

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

$$T_K(Z^*) = \max\left(\sup_{Z \in \mathcal{C} \backslash \mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z), \sup_{Z \in \mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \right) \leqslant \eta \tag{47}$$

and, by definition of $\hat{Z}$, we also have $T_K(\hat{Z}) = \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_Z) \leqslant \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) = T_K(Z^*) \leqslant \eta$. However, by (46), any $Z \in \mathcal{C} \backslash \mathcal{C}_{\gamma,G}$ must satisfy

$$T_K(Z) \geqslant \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) \geqslant \inf_{Z \in \mathcal{C} \backslash \mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) > \eta.$$

Therefore, we necessarily have $\hat{Z} \in \mathcal{C} \cap \mathcal{C}_{\gamma,G}$, that is $G(\hat{Z} - Z^*) \leqslant (r^*)^2$. ■

Lemma 5.8 *Assume that $A < 2$. On the event Ω_K, (45) 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 2.20 that $P\mathcal{L}_{Z_0} \geqslant A^{-1}G(Z_0 - Z^*)$. Moreover, on Ω_K, there are at least $K/2$ blocks B_k on which $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leqslant (r^*)^2/4$ and so $P_{B_k}\mathcal{L}_{Z_0} \geqslant P\mathcal{L}_{Z_0} - (r^*)^2/4 \geqslant A^{-1}G(Z_0 - Z^*) - (r^*)^2/4 \geqslant (r^*)^2/4$ since we assumed that $A < 2$. Now, by linearity of the loss function, we have on these blocks

$$P_{B_k}\mathcal{L}_Z = \alpha P_{B_k}\mathcal{L}_{Z_0} \geqslant \alpha(r^*)^2/4 > (r^*)^2/4.$$

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:

$$\sup_{Z\in\mathcal{C}\backslash\mathcal{C}_{\gamma,G}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leqslant -\frac{(r^*)^2}{4}.$$

This shows the left-hand side inequality of (45) for $\eta = (r^*)^2/4$.

Next, let $Z \in \mathcal{C}$ be such that $G(Z - Z^*) \leqslant (r^*)^2$. On Ω_K, there are at least $K/2$ blocks B_k on which $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leqslant (r^*)^2/4$, that is $-P_{B_k}\mathcal{L}_Z \leqslant (r^*)^2/4 - P\mathcal{L}_Z \leqslant (r^*)^2/4$ since $P\mathcal{L}_Z \geqslant 0$ by definition of Z^*. Then, $\mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) \leqslant (r^*)^2/4$. This holds for all $Z \in \mathcal{C}_{\gamma,G}$, in other words, the right-hand side inequality of (45) holds for $\eta = (r^*)^2/4$ and this concludes the proof. ■

Lemma 5.9 *Assume the conditions of Theorem 2.21 are met. Then, on* Ω_K, $P\mathcal{L}_{\hat{Z}} \leqslant (r^*)^2/2$.

Proof From Assumption 2.20 combined with the fact that $A < 2$, we have from Lemmas 5.7 and 5.8 that $G(\hat{Z} - Z^*) \leqslant (r^*)^2$. Then on Ω_K there exist strictly more than $K/2$ blocks B_k on which $\left|(P_{B_k} - P)\mathcal{L}_{\hat{Z}}\right| \leqslant (r^*)^2/4$, that is:

$$P\mathcal{L}_{\hat{Z}} \leqslant P_{B_k}\mathcal{L}_{\hat{Z}} + \frac{(r^*)^2}{4} \tag{48}$$

Moreover, by (47) and by definition of $\hat{Z}$, we have:

$$\mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) \leqslant \sup_{Z\in\mathcal{C}} \mathrm{MOM}_K(\ell_{\hat{Z}} - \ell_Z) \leqslant \sup_{Z\in\mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) = T_K(Z^*) \leqslant \eta = \frac{(r^*)^2}{4}.$$

As a consequence, there exist at least $K/2$ blocks B_k on which $P_{B_k}(\ell_{\hat{Z}} - \ell_{Z^*}) \leqslant (r^*)^2/4$, that is:

$$P_{B_q}\mathcal{L}_{\hat{Z}} \leqslant \frac{(r^*)^2}{4}. \tag{49}$$

So there must be at least one block B_{k_0} on which (48) and (49) hold simultaneously. On this block, we have:

$$P\mathcal{L}_{\hat{Z}} \leqslant P_{B_{k_0}}\mathcal{L}_{\hat{Z}} + \frac{(r^*)^2}{4} \leqslant \frac{(r^*)^2}{4} + \frac{(r^*)^2}{4} = \frac{(r^*)^2}{2}.$$

■

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 2.21. The final ingredient is to show that, under the conditions of Theorem 2.21, 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 5.2.

Lemma 5.10 *Assume the conditions of Theorem 2.21 are met, with $A < 2$. Then Ω_K holds with probability at least* $1 - \exp(-72K/625)$.

5.1.5 Proof of Theorem 2.24

The proof is structured in the same way as the previous ones: we identify an event on which $\hat{Z}_{K,\lambda}^{\text{RMOM}}$ 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 2.23. For the sake of simplicity, all along this proof we write $\hat{Z}$ for $\hat{Z}_{K,\lambda}^{\text{RMOM}}$ and $r_b^* := r_{\text{RMOM,ER}}^*(\gamma, b\rho^*)$ for both $b \in \{1,2\}$. For $b \in \{1,2\}$, we define $\mathcal{B}_b := \left\{Z \in \mathcal{C} : P\mathcal{L}_Z \leqslant (r_b^*)^2 \text{ and } \|Z - Z^*\| \leqslant b\rho^*\right\}$. Then we define:

$$\Omega_K = \left\{\forall b \in \{1,2\}, \forall Z \in \mathcal{B}_b, \exists J \subset [K], |J| > K/2, \forall k \in J, |(P_{B_k} - P)\mathcal{L}_Z| \leqslant \frac{(r_b^*)^2}{20}\right\}.$$

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 2.24.

Lemma 5.11 *If there exists $\eta > 0$ such that*

$$\sup_{Z \in \mathcal{C}\backslash\mathcal{B}_2} \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda\left(\|Z^*\| - \|Z\|\right) < -\eta \tag{50}$$

and

$$\sup_{Z \in \mathcal{C}} \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda\left(\|Z^*\| - \|Z\|\right) \leqslant \eta \tag{51}$$

then $P\mathcal{L}_{\hat{Z}} \leqslant (r_2^*)^2$ and $\|\hat{Z} - Z^*\| \leqslant 2\rho^*$.

Proof For $Z \in \mathcal{C}$, define $S(Z) = \sup_{Z' \in \mathcal{C}} \text{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\| - \|Z'\|)$. For all $Z \in \mathcal{C}\backslash\mathcal{B}_2$ we have:

$$S(Z) \geqslant \text{MOM}_K(\ell_Z - \ell_Z^*) + \lambda(\|Z\| - \|Z^*\|) \geqslant \inf_{Z \in \mathcal{C}\backslash\mathcal{B}_2} \text{MOM}_K(\ell_Z - \ell_Z^*) + \lambda(\|Z\| - \|Z^*\|) > \eta$$

since (50) holds. Moreover, we have by definition of $\hat{Z}$:

$$S(\hat{Z}) \leqslant S(Z^*) = \sup_{Z \in \mathcal{C}} \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant \eta$$

since (51) holds. This shows that necessarily $\hat{Z} \in \mathcal{B}_2$. ■

We are now looking for $\eta > 0$ such that (50) and (51) hold, which the following Lemma allows us to do.

Lemma 5.12 *Under the assumptions of Theorem 2.24 and on the event event Ω_K, (50) and (51) 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

$$\begin{aligned} P_{B_k}\mathcal{L}_Z^\lambda &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) = \alpha P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|\alpha Z_0 + (1-\alpha)Z^*\| - \|Z^*\|) \\ &\geqslant \alpha P_{B_k}\mathcal{L}_{Z_0} + \lambda\alpha(\|Z_0\| - \|Z^*\|) = \alpha P_{B_k}\mathcal{L}_{Z_0}^\lambda. \end{aligned} \tag{52}$$

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 $P_{B_k}\mathcal{L}_{Z_0} \geqslant P\mathcal{L}_{Z_0} - (r_b^*)^2/20 = (19/20)(r_b^*)^2$. Therefore, on these blocs, we have

$$\begin{aligned} P_{B_k}\mathcal{L}^\lambda_{Z_0} &= P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \geqslant \frac{19}{20}(r_b^*)^2 - \lambda\|Z_0 - Z^*\| \\ &\geqslant \frac{19}{20}(r_b^*)^2 - \lambda b\rho^* = \frac{19}{20}(r_b^*)^2 - \frac{11b}{40}(r_2^*)^2 \geqslant \left\{ \begin{array}{ll} 2(r_2^*)^2/5 & \text{for } b = 2 \\ (r_2^*)^2/5 & \text{for } b = 1. \end{array} \right. \end{aligned} \tag{53}$$

where we used in the case $b = 1$ that $r_1^* \geqslant r_2^*/\sqrt{2}$ thanks to Proposition A.1 from the Appendix.

In the second case *b)*, we have $Z_0 \in \bar{H}_{b\rho^*}$ from Definition 2.23. Since the sparsity equation holds for $\rho = \rho^*$, it also holds for $\rho = b\rho^*$ (see Proposition A.2 in the Appendix). Let $V \in H$ be such that $\|Z^* - V\| \leqslant b\rho^*/20$ and $\Phi \in \partial\|.\|(V)$. We have:

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

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:

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

and then taking the infimum over $\bar{H}_{b\rho^*}$ gives:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geqslant \inf_{Z_0 \in \bar{H}_{b\rho^*}} \|Z_0\| - \|Z^*\| \geqslant \inf_{Z_0 \in \bar{H}_{b\rho^*}} \sup_{\Phi \in \Gamma_{Z^*}(2\rho^*)} \left\langle \Phi, Z_0 - Z^* \right\rangle - \frac{b\rho^*}{10} \\ &= \Delta(b\rho^*) - \frac{b\rho^*}{10} \geqslant \frac{7}{10} b\rho^* \end{aligned} \tag{54}$$

where the last inequality holds since $b\rho^*$ satisfies the sparsity equation. Then, $\lambda(\|Z_0\| - \|Z^*\|) \geqslant (7/10)\lambda b\rho^* = (77/400)b(r_2^*)^2$. Now, since $Z_0 \in \mathcal{B}_b$, on Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leqslant (r_b^*)^2/20$ and so $P_{B_k}\mathcal{L}_{Z_0} \geqslant (r_b^*)^2/20$ - because $P\mathcal{L}_{Z_0} \geqslant 0$. Therefore, on the very same blocks,

$$P_{B_k}\mathcal{L}^\lambda_{Z_0} = P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \geqslant -\frac{1}{20}(r_b^*)^2 + \frac{77}{400}b(r_2^*)^2 \geqslant \left\{ \begin{array}{ll} 134(r_2^*)^2/400 & \text{for } b = 2 \\ 29(r_2^*)^2/400 & \text{for } b = 1 \end{array} \right. \tag{55}$$

where we used that $r_1^* \leqslant r_2^*$ (see Proposition A.1 in the Appendix). As a consequence, it follows from (52), the fact that $\alpha > 1$, (53) and (55) for $b = 2$ that for all $Z \in \mathcal{C} \backslash \mathcal{B}_2$, on more than $K/2$ blocks B_k: $P_{B_k}\mathcal{L}^\lambda_Z \geqslant (134/400)(r_2^*)^2$ and so (50) holds for $\eta \leqslant (134/400)(r_2^*)^2$.

Let us now turn to Equation (51). Let $Z \in \mathcal{B}_1$. On Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leqslant (r_1^*)^2/20$. On these blocks B_k, all $P_{B_k}\mathcal{L}_Z^\lambda$'s are such that

$$\begin{aligned} P_{B_k}\mathcal{L}_Z^\lambda &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \geqslant P\mathcal{L}_Z - \frac{1}{20}(r_1^*)^2 - \lambda\|Z - Z^*\| \geqslant -\frac{1}{20}(r_1^*)^2 - \lambda\rho^* \\ &= -\frac{1}{20}(r_1^*)^2 - \frac{11}{40}(r_2^*)^2 \geqslant -\frac{13}{40}(r_2^*)^2 \end{aligned} \tag{56}$$

because $r_1^* \leqslant r_2^*$ (see Proposition A.1 in the Appendix). Next, it follows from (52), the fact that $\alpha > 1$, (53) for $b = 1$, (55) for $b = 1$ and (56) that

$$\sup_{Z \in \mathcal{C}} \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant \max\left(\frac{-1}{5}, \frac{-29}{400}, \frac{13}{40}\right) r_2^2 = \frac{13}{40}(r_2^*)^2 \tag{57}$$

and so (51) holds for $\eta \geqslant 13(r_2^*)^2/40$. As a consequence, (50) and (51) 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 2.24. In what follows we prove that in the framework of Theorem 2.24, Ω_K holds with exponentially large probability.

Lemma 5.13 *Assume that $K \geqslant 100|\mathcal{O}|$, and let $\rho^* > 0$ be such that it satisfies the sparsity equation from Definition 2.23. Then, Ω_K holds with probability at least $1 - 2\exp(-72K/625)$.*

Proof Let $\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t \geqslant 1\}} + 2(t - 1/2)\mathbb{1}_{\{1/2 \leqslant t \leqslant 1\}}$, so that for any $t \in \mathbb{R}$, $\mathbb{1}_{\{t \geqslant 1\}} \leqslant \phi(t) \leqslant \mathbb{1}_{\{t \geqslant 1/2\}}$. 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: $\widetilde{W}_k := \left\{\tilde{X}_i : i \in B_k\right\}$ and $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - \tilde{P})\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := \frac{K}{N}\sum_{i \in B_k}\mathcal{L}_Z(\tilde{X}_i)$ and $\tilde{P}\mathcal{L}_Z := \mathbb{E}[\mathcal{L}_Z(\tilde{X}_i)]$. For both $b \in \{1, 2\}$, let $Z \to \psi_b(Z) = \sum_{k \in [K]} \mathbb{1}_{\{|F_Z(W_k)| \leqslant (r_b^*)^2/20\}}$. Let $b \in \{1, 2\}$. We want to show that, with high probability, if $Z \in \mathcal{B}_b$, then $\psi_b(Z) > K/2$ which follows if one can proves that

$$\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{20}\right\}} \leqslant \frac{49K}{100}. \tag{58}$$

Indeed, consider $Z \in \mathcal{B}_b$ such that (58) holds. Then, there exist at least $(1 - 49/100)K = 51K/100$ blocks B_k on which $|F_Z(\widetilde{W}_k)| \leqslant (r_b^*)^2/20$. On the other hand, we know that $|\mathcal{O}| \leqslant K/100$, 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:

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \\
&= \sum_{k\in[K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right)\right] \\
&= \sum_{k\in[K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{40}\right\}}\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leqslant \sum_{k\in[K]} \left(\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leqslant \sup_{Z\in\mathcal{B}_b} \left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right)
\end{aligned} \tag{59}$$

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

$$\begin{aligned}
\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) &\leqslant \frac{1600}{(r_b^*)^4}\mathbb{E}\left[F_Z(\widetilde{W}_k)^2\right] = \frac{1600}{(r_b^*)^4}\left(\frac{K}{N}\right)\mathrm{Var}(\mathcal{L}_Z(\widetilde{X})) \\
\leqslant \frac{1600}{(r_b^*)^4}\left(V_K(r_b^*)\right)^2 \leqslant \frac{1}{200}
\end{aligned}$$

Plugging this last result into (59), we get:

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \leqslant \frac{K}{200} + \sup_{Z\in\mathcal{B}_b}\left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right). \tag{60}$$

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

$$\begin{aligned}
&\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) \\
&\leqslant \frac{12K}{25} + \mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right].
\end{aligned} \tag{61}$$

Let now $\epsilon_1, \dots, \epsilon_K$ be Rademacher variables independant from the $\widetilde{X}_i$'s. By the symmetrization Lemma, we have:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right]\leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]. \tag{62}$$

As ϕ is Lipschitz with $\phi(0) = 0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), chapter 4) to get that:

$$\begin{aligned}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]&\leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k\frac{20F_Z(\widetilde{W}_k)}{(r_b^*)^2}\right]\\&=\frac{40}{(r_b^*)^2}\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-\widetilde{P})\mathcal{L}_Z\right]\end{aligned} \tag{63}$$

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:

$$\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]}\epsilon_k(\widetilde{P_{B_k}}-\widetilde{P})\mathcal{L}_Z\right]\leqslant 2\mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\frac{K}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right]\leqslant 2KE(r_b^*, b\rho^*)\leqslant 2K\gamma(r_b^*)^2.$$

Combining this with (61), (62) and (63), we finally get that, with probability at least $1-\exp(-72K/625)$:

$$\sup_{Z\in\mathcal{C}_\gamma}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\leqslant\left(\frac{12}{25}+160\gamma\right)K \tag{64}$$

Plugging that into (60), we conclude that, with probability at least $1-\exp(-72K/625)$, for all $Z\in\mathcal{B}_b$,

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}}\leqslant\left(\frac{1}{200}+\frac{12}{25}+160\gamma\right)K\leqslant\frac{49}{100}K$$

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.

■

5.1.6 Proof of Theorem 2.28

Let $K > 0$ be a divisor of N such that $K \geqslant 100|\mathcal{O}|$. Let $\gamma = 1/32000$. Let $A \in (0, 1]$ and $\rho^* > 0$ be such that Assumption 2.27 holds and satisfying the sparsity equation from Definition 2.26. Define $\nu = 320000A^2$.

For the sake of simplicity, we write all along this proof $\hat{Z}$ for $\hat{Z}_{K,\lambda}^{\mathrm{RMOM}}$. 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 $\mathcal{B}_{K,b} := \left\{Z \in \mathcal{C} : \|Z - Z^*\| \leqslant b\rho^* \text{ and } \|Z - Z^*\|_{L_2} \leqslant \sqrt{C_{K,b}}\right\}$ - we recall that the L_2-norm associated with the good data $\tilde{X}$ is defined as $\|Z\|_{L_2} = \mathbb{E}[\langle \tilde{X}, Z\rangle^2]^{1/2}$. 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:

$$\Omega_K = \left\{\forall b \in \{1,2\}, \forall Z \in \mathcal{B}_{K,b}, \quad \sum_{k=1}^{K} \mathbb{1}_{\left\{|(P_{B_k} - P)\mathcal{L}_Z| \leqslant \frac{C_{K,b}}{20}\right\}} > \frac{K}{2}\right\}.$$

First, we show that on Ω_K, $\hat{Z}$ has the statistical properties announced in Theorem 2.28. Then, we show that Ω_K holds with high probability.

Lemma 5.14 *If there exists $\eta > 0$ such that*

$$\sup_{Z \in \mathcal{C} \backslash \mathcal{B}_{K,2}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) < -\eta \tag{65}$$

and

$$\sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant \eta \tag{66}$$

then $\|Z - Z^*\| \leqslant 2\rho^*$ and $\|Z - Z^*\|_{L_2} \leqslant \sqrt{C_{K,2}}$.

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 $Z \in \mathcal{C} \backslash \mathcal{B}_{K,2}$ we have:

$$S(Z) \geqslant \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) + \lambda(\|Z\| - \|Z^*\|) \geqslant \inf_{Z \in \mathcal{C} \backslash \mathcal{B}_{K,2}} \mathrm{MOM}_K(\ell_Z - \ell_{Z^*}) + \lambda(\|Z\| - \|Z^*\|) > \eta$$

since (65) holds. Moreover, we have by definition of $\hat{Z}$:

$$S(\hat{Z}) \leqslant S(Z^*) = \sup_{Z \in \mathcal{C}} \mathrm{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant \eta$$

since (66) holds. This shows that necessarily $\hat{Z} \in \mathcal{B}_{K,2}$. ■

We are now looking for $\eta > 0$ such that (65) and (66) hold. In the following result we identify such a η on the event Ω_K.

Lemma 5.15 *Under the conditions of Theorem 2.28 and on the event Ω_K, (65) and (66) hold with $\eta = 33C_{K,2}/100$.*

Proof Consider $b \in \{1, 2\}$ and $Z \in \mathcal{C} \backslash \mathcal{B}_{K,b}$. 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

$$\begin{aligned} P_{B_k} \mathcal{L}_Z^\lambda &= P_{B_k} \mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) = \alpha P_{B_k} \mathcal{L}_{Z_0} + \lambda(\|\alpha Z_0 + (1-\alpha) Z^*\| - \|Z^*\|) \\ &\geqslant \alpha P_{B_k} \mathcal{L}_{Z_0} + \lambda \alpha (\|Z_0\| - \|Z^*\|) = \alpha P_{B_k} \mathcal{L}_{Z_0}^\lambda. \end{aligned} \tag{67}$$

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 $P_{B_k} \mathcal{L}_{Z_0} \geqslant P\mathcal{L}_{Z_0} - C_{K,b}/(20)$. But from Assumption 2.27, we have in this case that $AP\mathcal{L}_{Z_0} \geqslant \|Z_0 - Z^*\|_{L_2}^2 = C_{K,b}$, so that, on the same blocks of data, $P_{B_k} \mathcal{L}_{Z_0} \geqslant (1/A) C_{K,b} - (1/20) C_{K,b} \geqslant (19/20) C_{K,b}$, since we assumed that $0 < A \leqslant 1$. Therefore, on these blocs, we have

$$\begin{aligned} P_{B_k} \mathcal{L}_{Z_0}^\lambda &= P_{B_k} \mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \geqslant \frac{19}{20} C_{K,b} - \lambda \|Z_0 - Z^*\| \\ &\geqslant \frac{19}{20} C_{K,b} - \lambda b \rho^* = \frac{19}{20} C_{K,b} - \frac{11b}{40} C_{K,2}. \end{aligned}$$

But thanks to Proposition A.1 from the Appendix, we have that $r_1^* \geqslant r_2^*/\sqrt{2}$, from which we deduce that $C_{K,1} \geqslant C_{K,2}/2$. As a consequence, on the previous blocks, we have

$$P_{B_k} \mathcal{L}_{Z_0}^\lambda \geqslant \begin{cases} 7C_{K,2}/40 & \text{for } b = 1 \\ 16 C_{K,2}/40 & \text{for } b = 2. \end{cases} \tag{68}$$

In the second case *b)*, we have $Z_0 \in \widetilde{H}_{b\rho^*, A}$ from Definition 2.26. Since the sparsity equation is satisfied by ρ^*, it is also satisfied by $b\rho^*$ as well (see Proposition A.2 in the Appendix). Let $V \in H$ be such that $\|Z^* - V\| \leqslant b\rho^*/20$ and $\Phi \in \partial \|.\|(V)$. We have:

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

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:

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

and then taking the infimum over $\widetilde{H}_{b\rho^*, A}$ gives:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geqslant \inf_{Z_0 \in \widetilde{H}_{b\rho^*, A}} \|Z_0\| - \|Z^*\| \geqslant \inf_{Z_0 \in \widetilde{H}_{b\rho^*, A}} \sup_{\Phi \in \Gamma_{Z^*}(2\rho^*)} \left\langle \Phi, Z_0 - Z^* \right\rangle - \frac{b\rho^*}{10} \\ &= \Delta(b\rho^*) - \frac{b\rho^*}{10} \geqslant \frac{7}{10} b\rho^* \end{aligned} \tag{69}$$

where the last inequality holds since $b\rho^*$ satisfies the sparsity equation. Then, $\lambda(\|Z_0\| - \|Z^*\|) \geqslant (7/10)\lambda b\rho^* = (77/400)bC_{K,2}$. Now, since $Z_0 \in \mathcal{B}_{K,b}$, on Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leqslant C_{K,b}/(20)$ and so $P_{B_k}\mathcal{L}_{Z_0} \geqslant -C_{K,b}/(20)$ (because $P\mathcal{L}_{Z_0} \geqslant 0$). Therefore, on the very same blocks,

$$P_{B_k}\mathcal{L}_{Z_0}^\lambda = P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \geqslant -\frac{1}{20}C_{K,b} + \frac{77b}{400}C_{K,2} \geqslant \left\{ \begin{array}{ll} 57(r_2^*)^2/400 & \text{for } b=1 \\ 134(r_2^*)^2/400 & \text{for } b=2 \end{array} \right. \tag{70}$$

where we used that $C_{K,1} \leqslant C_{K,2}$ because $r_1^* \leqslant r_2^*$ (see Proposition A.1 in the Appendix). As a consequence, it follows from (67), the fact that $\alpha > 1$, (68) and (70) for $b = 2$ that, for all $Z \in \mathcal{C}\backslash\mathcal{B}_{K,2}$, on more than $K/2$ blocks B_k: $P_{B_k}\mathcal{L}_Z^\lambda \geqslant (134/400)C_{K,2}$ and so (65) holds for $\eta < (134/400)C_{K,2}$.

Let us now turn to Equation (66). Let $Z \in \mathcal{B}_{K,1}$. On Ω_K there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leqslant C_{K,1}/20$. On these blocks B_k, all $P_{B_k}\mathcal{L}_Z^\lambda$'s are such that

$$\begin{aligned} P_{B_k}\mathcal{L}_Z^\lambda &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \geqslant P\mathcal{L}_Z - \frac{1}{20}C_{K,1} - \lambda\|Z - Z^*\| \geqslant -\frac{1}{20}C_{K,1} - \lambda\rho^* \\ &= -\frac{1}{20}C_{K,1} - \frac{11}{40}C_{K,2} \geqslant -\frac{13}{40}C_{K,2} \end{aligned} \tag{71}$$

where we used the fact that, thanks to Proposition A.1 in the Appendix, $C_{K,1} \leqslant C_{K,2}$. Next, it follows from (67), the fact that $\alpha > 1$, (68) and (67) for $b = 1$ and (71) that

$$\sup_{Z\in\mathcal{C}} \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant \max\left(\frac{-7}{40}, \frac{-57}{400}, \frac{13}{40}\right) C_{K,2} = \frac{13}{40}C_{K,2} \tag{72}$$

and so (66) holds for $\eta \geqslant 13C_{K,2}/40$. As a consequence, (65) and (66) both hold for $\eta = 132C_{K,2}/400$.

■

From Lemmas 5.14 and 5.15, 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 5.16 *Under the conditions of Theorem 2.28 and on the event Ω_K, we have $P\mathcal{L}_{\hat{Z}} \leqslant (27/100)C_{K,2}$.*

Proof From Lemmas 5.14 and 5.15, 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

$$P\mathcal{L}_{\hat{Z}} \leqslant P_{B_k}\mathcal{L}_{\hat{Z}} + C_{K,2}/20. \tag{73}$$

Now, by definition of $\hat{Z}$, (66) and Lemma 5.15 we get

$$\text{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) + \lambda(\|\hat{Z}\| - \|Z^*\|) \leqslant \sup_{Z\in\mathcal{C}} \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant 33C_{K,2}/100.$$

This means that there exist at least $K/2$ blocks B_k on which $P_{B_k}\mathcal{L}_{\hat{Z}} + \lambda(\|\hat{Z}\| - \|Z^*\|) \leqslant 33C_{K,2}/100$. Since $\lambda(\|Z^*\| - \|\hat{Z}\|) \leqslant \lambda\|Z^* - \hat{Z}\| \leqslant 2\lambda\rho^* = 11C_{K,2}/20$, we have on these blocks

$$P_{B_k}\mathcal{L}_{\hat{Z}} \leqslant 33C_{K,2}/100 + 11C_{K,2}/20 = 22C_{K,2}/100. \tag{74}$$

Therefore, there exist at least a block B_{k_0} on which (73) and (74) hold simultaneously. On this block, we can write

$$P\mathcal{L}_{\hat{Z}} \leqslant P_{B_{k_0}}\mathcal{L}_{\hat{Z}} + C_{K,2}/20 \leqslant 22C_{K,2}/100 + C_{K,2}/20 = 27C_{K,2}/100.$$

■

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 2.28. In what follows, we prove that, in the framework of Theorem 2.28, Ω_K holds with exponentially large probability.

Proposition 5.17 *Consider ρ^* that satisfies the sparsity equation from Definition 2.26. Assume that $K \geqslant 100|\mathcal{O}|$. Then, Ω_K holds with probability at least $1 - 2\exp(-72K/625)$.*

Proof Let $b \in \{1, 2\}$. Let $\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t \geqslant 1\}} + 2(t - 1/2)\mathbb{1}_{\{1/2 \leqslant t \leqslant 1\}}$, so that for any $t \in \mathbb{R}$, $\mathbb{1}_{\{t \geqslant 1\}} \leqslant \phi(t) \leqslant \mathbb{1}_{\{t \geqslant 1/2\}}$. 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: $\widetilde{W}_k := \left\{\tilde{X}_i : i \in B_k\right\}$ and $F_Z(\widetilde{W}_k) = (\widetilde{P_{B_k}} - P)\mathcal{L}_Z$, where $\widetilde{P_{B_k}}\mathcal{L}_Z := (K/N)\sum_{i \in B_k}\mathcal{L}_Z(\tilde{X}_i)$. Let $\psi_b(Z) = \sum_{k \in [K]} \mathbb{1}_{\{|F_Z(W_k)| \leqslant C_{K,b}/(20)\}}$. 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 5.2, in our framework this is true if we show that with high probability, for all $Z \in \mathcal{B}_{K,b}$,

$$\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} \leqslant \frac{49K}{100} \tag{75}$$

and this is what we do now. Let $Z \in \mathcal{B}_{K,b}$. We have:

$$\begin{aligned}
&\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} \\
&= \sum_{k \in [K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right)\right] \\
&= \sum_{k \in [K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right\}}\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) \\
&\leqslant \sum_{k \in [K]} \left(\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) \\
&\leqslant \sup_{Z \in \mathcal{B}_{K,b}} \left(\sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right) + \sum_{k \in [K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right)
\end{aligned} \tag{76}$$

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

$$\begin{aligned}
&\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{40}\right) \\
&\leqslant \left(\frac{40}{C_{K,b}}\right)^2 \mathbb{E}\left[|F_Z(\widetilde{W}_k)|^2\right] = \left(\frac{40}{C_{K,b}}\right)^2 \left(\frac{K}{N}\right)^2 \mathbb{E}\left[\left(\sum_{i \in B_k} \mathcal{L}_Z(\widetilde{X}_i) - \mathbb{E}\left[\mathcal{L}_Z(\widetilde{X}_i)\right]\right)^2\right] \\
&\leqslant \left(\frac{40}{C_{K,b}}\right)^2 \frac{K}{N} \|Z - Z^*\|_{L_2}^2 \leqslant \left(\frac{40}{C_{K,b}}\right)^2 \frac{K}{N} C_{K,b} \leqslant 40^2 \nu^{-1} = \frac{1}{200}.
\end{aligned}$$

Plugging this last result into (76), we get:

$$\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{C_{K,b}}{20}\right\}} \leqslant \frac{K}{200} + \sup_{Z \in \mathcal{B}_{K,b}} \left(\sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right). \tag{77}$$

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

$$\begin{aligned}
&\sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) \\
&\leqslant \frac{12K}{25} + \mathbb{E}\left[\sup_{Z \in \mathcal{B}_{K,b}} \sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right].
\end{aligned} \tag{78}$$

Let now $\epsilon_1, \dots, \epsilon_K$ be Rademacher variables independant from the $\widetilde{X}_i$'s. By the symmetrization Lemma, we have:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{Z \in \mathcal{B}_{K,b}} \sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\right] \\
&\leqslant 2\mathbb{E}\left[\sup_{Z \in \mathcal{B}_{K,b}} \sum_{k \in [K]} \epsilon_k \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right].
\end{aligned} \tag{79}$$

As ϕ is Lipschitz with $\phi(0) = 0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), chapter 4) to get that:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{Z \in \mathcal{B}_{K,b}} \sum_{k \in [K]} \epsilon_k \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right] \\
&\leqslant 2\mathbb{E}\left[\sup_{Z \in \mathcal{B}_{K,b}} \sum_{k \in [K]} \epsilon_k \frac{20 F_Z(\widetilde{W}_k)}{C_{K,b}}\right] = \frac{40}{C_{K,b}} \mathbb{E}\left[\sup_{Z \in \mathcal{B}_{K,b}} \sum_{k \in [K]} \epsilon_k (\widetilde{P_{B_k}} - \widetilde{P}) \mathcal{L}_Z\right]
\end{aligned} \tag{80}$$

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:

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

Combining this with (78), (79) and (80), we finally get that, with probability at least $1-\exp(-72K/625)$:

$$\sup_{Z\in\mathcal{B}_{K,b}}\sum_{k\in[K]}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)-\mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{C_{K,b}}\right)\right]\leqslant\left(\frac{12}{25}+160\gamma\right)K \tag{81}$$

Plugging that into (77), we conclude that, with probability at least $1-\exp(-72K/625)$, for all $Z\in\mathcal{B}_{K,b}$,

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{C_{K,b}}{20}\right\}}\leqslant\left(\frac{1}{200}+\frac{12}{25}+160\gamma\right)K\leqslant\frac{49}{100}K$$

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.

■

5.1.7 Proof of Theorem 2.32

The proof is structured in the same way as the previous ones: we identify an event on which $\hat{Z}_{K,\lambda}^{\text{RMOM}}$ has the desired statistical properties, then we show that this event holds with high probability. We place ourselves under the conditions of Theorem 2.32, i.e., we assume the existence of $A\in(0,1]$ such that Assumption 2.31 holds, $\gamma=1/32000$ and ρ^* which satisfies the sparsity equation from Definition 2.30. For $b\in\{1,2\}$ we define $r_b^*=r_{\text{RMOM,G}}^*(\gamma,2\rho^*)$ and $\mathcal{B}_b:=\{Z\in\mathcal{C}:G(Z-Z^*)\leqslant(r_b^*)^2\text{ and }\|Z-Z^*\|\leqslant b\rho^*\}$. With these notation, $\lambda=(11/(40\rho^*))r_2^*$. We consider the event

$$\Omega_{K,G}=\left\{\forall b\in\{1,2\},\forall Z\in\mathcal{B}_b,\quad\sum_{k=1}^{K}\mathbb{1}\left(|(P_{B_k}-P)\mathcal{L}_Z|\leqslant\frac{1}{20}(r_b^*)^2\right)>\frac{K}{2}\right\}$$

For the sake of simplicity, in the rest of the proof we write $\hat{Z}=\hat{Z}_{K,\lambda}^{\text{RMOM}}$.

Lemma 5.18 *If there exists $\eta>0$ such that*

$$\sup_{Z\in\mathcal{C}\backslash\mathcal{B}_2}\text{MOM}_K(\ell_{Z^*}-\ell_Z)+\lambda(\|Z^*\|-\|Z\|)<-\eta \tag{82}$$

and

$$\sup_{Z\in\mathcal{C}}\text{MOM}_K(\ell_{Z^*}-\ell_Z)+\lambda(\|Z^*\|-\|Z\|)\leqslant\eta \tag{83}$$

then $\|Z-Z^*\|\leqslant 2\rho^*$ and $G(Z-Z^*)\leqslant r_{\text{RMOM,G}}^*(\gamma,2\rho^*)^2$.

Proof Let η be such that (82) and (83) hold. For all $Z \in \mathcal{C}$, define $S(Z) = \sup_{Z' \in \mathcal{C}} \text{MOM}_K(\ell_Z - \ell_{Z'}) + \lambda(\|Z\| - \|Z'\|)$. It follows from (82) that for all $Z \in \mathcal{C} \backslash \mathcal{B}_2$,

$$S(Z) \geqslant \text{MOM}_K(\ell_Z - \ell_{Z*}) + \lambda(\|Z\| - \|Z^*\|) > \eta$$

Moreover, it follows from the definition of $\hat{Z}$ and (83) that

$$S(\hat{Z}) \leqslant S(Z^*) = \sup_{Z \in \mathcal{C}} \text{MOM}_K(\ell_{Z*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant \eta$$

This shows that necessarily $\hat{Z} \in \mathcal{B}_2$. ■

Lemma 5.19 *Under the conditions of Theorem 2.32 and on the event $\Omega_{K,G}$, (82) and (83) 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

$$\begin{aligned} P_{B_k} \mathcal{L}_Z^\lambda &= P_{B_k} \mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) = \alpha P_{B_k} \mathcal{L}_{Z_0} + \lambda(\|\alpha Z_0 + (1-\alpha) Z^*\| - \|Z^*\|) \\ &\geqslant \alpha P_{B_k} \mathcal{L}_{Z_0} + \lambda \alpha(\|Z_0\| - \|Z^*\|) = \alpha P_{B_k} \mathcal{L}_{Z_0}^\lambda. \end{aligned} \tag{84}$$

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 $P_{B_k} \mathcal{L}_{Z_0} \geqslant P\mathcal{L}_{Z_0} - (r_b^*)^2/(20)$. But we also have from Assumption 2.31 that $AP\mathcal{L}_{Z_0} \geqslant G(Z_0 - Z^*) = (r_b^*)^2$, so that $P_{B_k} \mathcal{L}_{Z_0} \geqslant (1/A)(r_b^*)^2 - (1/20)(r_b^*)^2 \geqslant (19/20)(r_b^*)^2$, since we assumed that $A \leqslant 1$. Therefore, on these blocs, we have

$$\begin{aligned} P_{B_k} \mathcal{L}_{Z_0}^\lambda &= P_{B_k} \mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \geqslant \frac{19}{20}(r_b^*)^2 - \lambda\|Z_0 - Z^*\| \\ &\geqslant \frac{19}{20}(r_b^*)^2 - \lambda b \rho^* = \frac{19}{20}(r_b^*)^2 - \frac{11b}{40}(r_2^*)^2 \geqslant \begin{cases} (r_2^*)^2/5 & \text{for } b = 1 \\ 2(r_2^*)^2/5 & \text{for } b = 2. \end{cases} \end{aligned} \tag{85}$$

where we used in the case $b = 1$ that $r_1^* \geqslant r_2^*/\sqrt{2}$ thanks to Proposition A.1 from the Appendix.

In the second case *b)*, we have $Z_0 \in \bar{H}_{b\rho^*}$ from Definition 2.30. Since the sparsity equation holds for $\rho = \rho^*$, it also holds for $\rho = b\rho^*$ (see Proposition A.2 in the Appendix). Let $V \in H$ be such that $\|Z^* - V\| \leqslant b\rho^*/20$ and $\Phi \in \partial\|.\|(V)$. We have:

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

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:

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

and then taking the infimum over $\bar{H}_{b\rho^*,A}$ gives:

$$\begin{aligned} \|Z_0\| - \|Z^*\| &\geqslant \inf_{Z_0 \in \bar{H}_{b\rho^*,A}} \|Z_0\| - \|Z^*\| \geqslant \inf_{Z_0 \in \bar{H}_{b\rho^*,A}} \sup_{\Phi \in \Gamma_{Z^*}(2\rho^*)} \left\langle \Phi, Z_0 - Z^* \right\rangle - \frac{b\rho^*}{10} \\ &= \Delta(b\rho^*) - \frac{b\rho^*}{10} \geqslant \frac{7}{10} b\rho^* \end{aligned} \tag{86}$$

where the last inequality holds since $b\rho^*$ satisfies the sparsity equation. Then, $\lambda(\|Z_0\| - \|Z^*\|) \geqslant (7/10)\lambda b\rho^* = (77/400)b(r_2^*)^2$. Now, since $Z_0 \in \mathcal{B}_b$, on $\Omega_{K,G}$ there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_{Z_0}| \leqslant (r_b^*)^2/(20)$ and so $P_{B_k}\mathcal{L}_{Z_0} \geqslant -(r_b^*)^2/(20)$ (because $P\mathcal{L}_{Z_0} \geqslant 0$). Therefore, on the very same blocks,

$$\begin{aligned} P_{B_k}\mathcal{L}_{Z_0}^{\lambda} &= P_{B_k}\mathcal{L}_{Z_0} + \lambda(\|Z_0\| - \|Z^*\|) \geqslant -\frac{1}{20}(r_b^*)^2 + \frac{77}{400} b(r_2^*)^2 \\ &\geqslant \begin{cases} 57(r_2^*)^2/(400) & \text{for } b = 1 \\ 134(r_2^*)^2/(400) & \text{for } b = 2 \end{cases} \end{aligned} \tag{87}$$

where we used that $r_1^* \leqslant r_2^*$ (see Proposition A.1 in the Appendix). As a consequence, it follows from (84), the fact that $\alpha > 1$, (85) and (87) for $b = 2$ that, for all $Z \in \mathcal{C} \backslash \mathcal{B}_2$, on more than $K/2$ blocks B_k: $P_{B_k}\mathcal{L}_Z^{\lambda} \geqslant (134/400)(r_2^*)^2$ and so (82) holds for $\eta < (134/400)(r_2^*)^2$.

Let us now turn to Equation (83). Let $Z \in \mathcal{B}_1$. On $\Omega_{K,G}$ there exist at least $K/2$ blocks B_k such that $|(P_{B_k} - P)\mathcal{L}_Z| \leqslant (r_1^*)^2/(20)$. On these blocks B_k, all $P_{B_k}\mathcal{L}_Z^{\lambda}$'s are such that

$$\begin{aligned} P_{B_k}\mathcal{L}_Z^{\lambda} &= P_{B_k}\mathcal{L}_Z + \lambda(\|Z\| - \|Z^*\|) \geqslant P\mathcal{L}_Z - \frac{1}{20}(r_1^*)^2 - \lambda\|Z - Z^*\| \geqslant -\frac{1}{20}(r_1^*)^2 - \lambda\rho^* \\ &= -\frac{1}{20}(r_1^*)^2 - \frac{11}{40}(r_2^*)^2 \geqslant -\frac{13}{40}(r_2^*)^2 \end{aligned} \tag{88}$$

because $r_1^* \leqslant r_2^*$ (see Proposition A.1 in the Appendix). Next, it follows from (84), the fact that $\alpha > 1$, (85) and (87) for $b = 1$ and (88) that

$$\sup_{Z \in \mathcal{C}} \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda(\|Z^*\| - \|Z\|) \leqslant \max\left(\frac{-1}{5}, \frac{-57}{400}, \frac{13}{40}\right) r_2^2 = \frac{13}{40}(r_2^*)^2 \tag{89}$$

and so (83) holds for $\eta \geqslant 13(r_2^*)^2/(40)$. As a consequence, (82) and (83) both hold for $\eta = 132(r_2^*)^2/(400)$. ■

From Lemmas 5.18 and 5.19, we conclude that on the event $\Omega_{K,G}$, $\hat{Z} \in \mathcal{B}_2$, that is $\|\hat{Z} - Z^*\| \leqslant 2\rho^*$ and $G(\hat{Z} - Z^*) \leqslant (r_2^*)^2$. The following lemma gives us an upper bound on the excess risk $P\mathcal{L}_{\hat{Z}}$.

Lemma 5.20 *Under the conditions of Theorem 2.32, and on the event $\Omega_{K,G}$, we have $P\mathcal{L}_{\hat{Z}} \leqslant (93/100)(r_2^*)^2$.*

Proof From Lemmas 5.18 and 5.19, 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 $\left|(P_{B_k} - P)\mathcal{L}_{\hat{Z}}\right| \leqslant (r_2^*)^2/(20)$, that is:

$$P\mathcal{L}_{\hat{Z}} \leqslant P_{B_k}\mathcal{L}_{\hat{Z}} + (r_2^*)^2/(20). \tag{90}$$

Moreover, by (83), the definition of $\hat{Z}$ and (5.19), we have:

$$\begin{aligned} &\text{MOM}_K(\ell_{\hat{Z}} - \ell_{Z^*}) + \lambda\left(\|\hat{Z}\| - \|Z^*\|\right) \leqslant \sup_{Z\in\mathcal{C}} \quad \text{MOM}_K(\ell_{\hat{Z}} - \ell_Z) + \lambda\left(\|\hat{Z}\| - \|Z\|\right) \\ &\leqslant \sup_{Z\in\mathcal{C}} \quad \text{MOM}_K(\ell_{Z^*} - \ell_Z) + \lambda\left(\|Z^*\| - \|Z\|\right) \leqslant \frac{33}{100}(r_2^*)^2. \end{aligned}$$

As a consequence, there exist at least $K/2$ blocks B_k on which

$$\begin{aligned} P_{B_k}\mathcal{L}_{\hat{Z}} &\leqslant \frac{33}{100}(r_2^*)^2 - \lambda\left(\|\hat{Z}\| - \|Z^*\|\right) \leqslant \frac{33}{100}(r_2^*)^2 + \lambda\|\hat{Z} - Z^*\| \\ &\leqslant \frac{33}{100}(r_2^*)^2 + 2\lambda\rho^* = \frac{88}{100}(r_2^*)^2. \end{aligned} \tag{91}$$

So there must be at least a block B_{k_0} on which (90) and (91) hold simultaneously. On this block, we have

$$P\mathcal{L}_{\hat{Z}} \leqslant P_{B_{k_0}}\mathcal{L}_{\hat{Z}} + \frac{1}{20}(r_2^*)^2 \leqslant \frac{88}{100}(r_2^*)^2 + \frac{1}{20}(r_2^*)^2 = \frac{93}{100}(r_2^*)^2.$$

■

At this stage, we have shown that on the event $\Omega_{K,G}$, the estimator $\hat{Z}$ has the statistical properties announced in Theorem 2.32. In what follows we prove that under the conditions of Theorem 2.32, $\Omega_{K,G}$ holds with exponentially large probability.

Lemma 5.21 *Assume that $K \geqslant 100|\mathcal{O}|$, and let $\rho^* > 0$ be such that it satisfies the sparsity equation from Definition 2.30. Then, Ω_K holds with probability at least $1-2\exp(-72K/625)$.*

Proof Let $\phi : t \in \mathbb{R} \to \mathbb{1}_{\{t\geqslant 1\}} + 2(t - 1/2)\mathbb{1}_{\{1/2\leqslant t\leqslant 1\}}$, so that for any $t \in \mathbb{R}$, $\mathbb{1}_{\{t\geqslant 1\}} \leqslant \phi(t) \leqslant \mathbb{1}_{\{t\geqslant 1/2\}}$. 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 $b \in \{1,2\}$, let $Z \to \psi_b(Z) = \sum_{k\in[K]}\mathbb{1}_{\{|F_Z(W_k)|\leqslant(r_b^*)^2/(20)\}}$. 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 5.2, in our framework this is equivalent to proving that the following inequality occurs with high probability:

$$\sum_{k\in[K]}\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \leqslant \frac{49K}{100}, \tag{92}$$

and this is what we do now. Let $Z \in \mathcal{B}_b$. We have:

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \\
&= \sum_{k\in[K]} \left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) + \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right)\right] \\
&= \sum_{k\in[K]} \left(\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{40}\right\}}\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leqslant \sum_{k\in[K]} \left(\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) \\
&\leqslant \sup_{Z\in\mathcal{B}_b} \left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right) + \sum_{k\in[K]} \mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right)
\end{aligned} \tag{93}$$

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

$$\begin{aligned}
\mathbb{P}\left(|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{40}\right) &\leqslant \frac{1600^2}{(r_b^*)^4}\mathbb{E}\left[F_Z(\widetilde{W}_k)^2\right] = \frac{1600^2}{(r_b^*)^4}\left(\frac{K}{N}\right)\mathrm{Var}(\mathcal{L}_Z(\widetilde{X})) \\
&\leqslant \frac{1600^2}{(r_b^*)^4}\left(V_K(r_b^*)\right)^2 \leqslant \frac{1}{200}
\end{aligned}$$

Plugging this last result into (93), we get:

$$\sum_{k\in[K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)|>\frac{(r_b^*)^2}{20}\right\}} \leqslant \frac{K}{200} + \sup_{Z\in\mathcal{B}_b}\left(\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right). \tag{94}$$

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

$$\begin{aligned}
&\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) \\
&\leqslant \frac{12K}{25} + \mathbb{E}\left[\sup_{Z\in\mathcal{B}_b}\sum_{k\in[K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right].
\end{aligned} \tag{95}$$

Let now $\epsilon_1, \dots, \epsilon_K$ be Rademacher variables independant from the $\widetilde{X}_i$'s. By the symmetrization Lemma, we have:

$$\mathbb{E}\left[\sup_{Z \in \mathcal{B}_b} \sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]\right] \leqslant 2\mathbb{E}\left[\sup_{Z \in \mathcal{B}_b} \sum_{k \in [K]} \epsilon_k \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right]. \tag{96}$$

As ϕ is Lipschitz with $\phi(0) = 0$, we can use the contraction Lemma (see Ledoux and Talagrand (2013), chapter 4) to get that:

$$\begin{aligned}\mathbb{E}\left[\sup_{Z \in \mathcal{B}_b} \sum_{k \in [K]} \epsilon_k \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right] &\leqslant 2\mathbb{E}\left[\sup_{Z \in \mathcal{B}_b} \sum_{k \in [K]} \epsilon_k \frac{20 F_Z(\widetilde{W}_k)}{(r_b^*)^2}\right] \\ &= \frac{40}{(r_b^*)^2}\mathbb{E}\left[\sup_{Z \in \mathcal{B}_b} \sum_{k \in [K]} \epsilon_k (\widetilde{P_{B_k}} - \widetilde{P})\mathcal{L}_Z\right] \end{aligned} \tag{97}$$

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:

$$\mathbb{E}\left[\sup_{Z \in \mathcal{B}_b} \sum_{k \in [K]} \epsilon_k (\widetilde{P_{B_k}} - \widetilde{P})\mathcal{L}_Z\right] \leqslant 2\mathbb{E}\left[\sup_{Z \in \mathcal{B}_b} \frac{K}{N} \sum_{i=1}^{N} \sigma_i \mathcal{L}_Z(\widetilde{X}_i)\right] \leqslant 2K E_G(r_b^*, b\rho^*) \leqslant 2K\gamma(r_b^*)^2.$$

Combining this with (95), (97) and (98), we finally get that, with probability at least $1 - \exp(-72K/625)$:

$$\sup_{Z \in \mathcal{C}_\gamma} \sum_{k \in [K]} \phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right) - \mathbb{E}\left[\phi\left(\frac{20|F_Z(\widetilde{W}_k)|}{(r_b^*)^2}\right)\right] \leqslant \left(\frac{12}{25} + 160\gamma\right) K \tag{98}$$

Plugging that into (94), we conclude that, with probability at least $1 - \exp(-72K/625)$, for all $Z \in \mathcal{B}_b$,

$$\sum_{k \in [K]} \mathbb{1}_{\left\{|F_Z(\widetilde{W}_k)| > \frac{(r_b^*)^2}{20}\right\}} \leqslant \left(\frac{1}{200} + \frac{12}{25} + 160\gamma\right) K \leqslant \frac{49}{100} K$$

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.

■

5.2 Proofs of section 3

5.2.1 Proof of Theorem 3.2

The proof of Theorem 3.2 relies on several Lemmas. We first recall that $kB_1 \cap B_2 \subset 2\text{conv}(U_{k^2} \cap \mathcal{S}_2^{d \times d})$ where $\mathcal{S}_2^{d \times d}$ is the unit sphere of $\ell_2^{d \times d}$ and U_{k^2} is the set of all matrices in

$\mathbb{R}^{d\times d}$ with at most k^2 non zero entries (see, for instance, equation (3.1) in Mendelson et al. (2007)). Hence, for all $A \in \mathbb{R}^{d\times d}$, we have

$$\|A\| \leqslant 2 \sup_{I\subset[d]\times[d]:|I|=k^2} \left(\sum_{(p,q)\in I} A_{pq}^2\right)^{1/2}.$$

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 5.22 *Let $(z_{pq} : p, q \in [d])$ be real-valued random variables (not necessarily independent) and $\lambda, t \geqslant 1$ be two positive constants. We assume that for $r = 2\log(ed/k) + t$, we have $\|z_{pq}\|_{L_r} \leqslant \lambda\sqrt{r}$ for all $p, q \in [d]$. Then, with probability at least $1 - \exp(-t)$,*

$$\sup_{I\subset[d]\times[d]:|I|=k^2} \left(\sum_{(p,q)\in I} z_{pq}^2\right)^{1/2} \leqslant e^2\lambda\sqrt{2k^2\left(\log(ed/k) + t\right)}.$$

Moreover:

$$\mathbb{E}\left[\sup_{I\subset[d]\times[d]:|I|=k^2} \left(\sum_{(p,q)\in I} z_{pq}^2\right)^{1/2}\right] \leqslant e^2\lambda\sqrt{6k^2\log(ed/k)}$$

Proof. We define for all $p, q \in [d]$,

$$Z_{pq} = z_{pq} I(|z_{pq}| \leqslant e\lambda\sqrt{r}) \text{ and } Y_{pq} = z_{pq} I(|z_{pq}| > e\lambda\sqrt{r})$$

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 $I \subset [d] \times [d]$

$$\left(\frac{1}{|I|}\sum_{(p,q)\in I} z_{pq}^2\right)^{t/2} \leqslant \frac{1}{|I|}\sum_{(p,q)\in I} |z_{pq}|^t = \frac{1}{|I|}\sum_{(p,q)\in I} |Z_{pq}|^t + \frac{1}{|I|}\sum_{(p,q)\in I} |Y_{pq}|^t. \tag{99}$$

Let $I \subset [d] \times [d]$ be such that $|I| = k^2$. We have

$$\frac{1}{|I|}\sum_{(p,q)\in I} |Z_{pq}|^t \leqslant (e\lambda\sqrt{r})^t. \tag{100}$$

For the second term in the right hand side inequality of (99), we have

$$\frac{1}{|I|}\sum_{(p,q)\in I} |Y_{pq}|^t \leqslant \frac{1}{|I|}\sum_{(p,q)\in[d]\times[d]} |Y_{pq}|^t$$

and for $\theta := r/t$ and all $p, q \in [d]$, we have

$$\begin{aligned}\mathbb{E}[|Y_{pq}|^t] = \mathbb{E}\left[|z_{pq}|^t I(|z_{pq}| > e\lambda\sqrt{r})\right] &\leqslant \mathbb{E}\left[|z_{pq}|^{t\theta}\right]^{1/\theta} \mathbb{P}\left[|z_{pq}| > e\lambda\sqrt{r}\right]^{1-1/\theta} \\ &\leqslant (\lambda\sqrt{r})^t \left(\frac{\|z_{pq}\|_{L_r}}{e\lambda\sqrt{r}}\right)^{r-t} \leqslant (\lambda\sqrt{r})^t e^{r-t} = (\lambda\sqrt{r})^t \frac{k^2}{e^2d^2}.\end{aligned}$$

It follows that

$$\mathbb{E}\sup_{I\subset[d]\times[d]:|I|=k^2}\frac{1}{|I|}\sum_{(p,q)\in I}|Y_{pq}|^t \leqslant \frac{d^2}{k^2}(\lambda\sqrt{r})^t\frac{k^2}{e^2d^2} = \frac{(\lambda\sqrt{r})^t}{e^2}.$$

Hence, using (99), (100) and the last inequality, we get $\mathbb{E}\mathcal{Z}^t \leqslant (e\lambda\sqrt{r})^t + (\lambda\sqrt{r})^t/e^2 \leqslant 2(e\lambda\sqrt{r})^t$ 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, $\|\mathcal{Z}\|_{L_t} \leqslant e\lambda\sqrt{2r}$ and so, for $t \geqslant 2$ we get by Markov's inequality that $\mathcal{Z} \leqslant e^2\lambda\sqrt{2r}$ with probability at least $1-\exp(-t)$.

Finally, by taking $t = 1$ above we get:

$$\|\mathcal{Z}\|_{L_1} \leqslant e\lambda\sqrt{2r} = e\lambda\sqrt{2(2\log(ed/k)+1)} \leqslant e\lambda\sqrt{6\log(ed/k)}$$

since $k \leqslant d$. As a consequence, $\mathbb{E}[\mathcal{Z}] = \|\mathcal{Z}\|_{L_1} \leqslant e\lambda\sqrt{6\log(ed/k)}$, which concludes the proof. ■

The proof of Theorem 3.2 will follow from Lemma 5.22 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 5.22. 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 3.1. To that end we use a version (see Lemma 2.8 in Lecué and Mendelson (to appear)) 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 5.23 *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 $z_1, \dots, z_N$ be N independent copies of z. Let $p_0 \geqslant 2$ and assume that there exists $\kappa_1 > 0$ and $\alpha \geqslant 1/2$ for which $\|z\|_{L_p} \leqslant \kappa_1 p^\alpha$ for every $2 \leqslant p \leqslant p_0$. If $N \geqslant p_0^{\max\{2\alpha-1,1\}}$ then for every $2 \leqslant p \leqslant p_0$,*

$$\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N z_i\right\|_{L_p} \leqslant c_1(\alpha)\kappa_1\sqrt{p},$$

where $c_1(\alpha) = c_0\exp((2\alpha-1))$.

We use Lemma 5.23 to prove the following moment growth condition on the $\hat{\Sigma}_{pq} - \Sigma_{pq}, p, q \in [d]$.

Lemma 5.24 *There exists an absolute constant c_0 such that the following holds. Grant Assumption 3.1 with parameters w and $t \geqslant 2$. For all $p, q \in [d]$ and all $2 \leqslant r \leqslant 2\log(ed/k) + t$, if $N \geqslant 2\log(ed/k) + t$ then $\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leqslant (c_0 w^2/\sqrt{N})\sqrt{r}$.*

Proof. Let $p, q \in [d]$. It follows from Assumption 3.1 and Lemma 5.23 that for all $p, q \in [d]$ and all $2 \leqslant r \leqslant 2\log(ed/k) + t$,

$$\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leqslant \frac{1}{\sqrt{N}}\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N X_{ip}X_{iq} - \mathbb{E}X_{ip}X_{iq}\right\|_{L_r} \leqslant \frac{c_0 w^2}{\sqrt{N}}\sqrt{r}.$$

■

Proof of Theorem 3.2 We set for all $p, q \in [d]$, $z_{pq} = \hat{\Sigma}_{pq} - \Sigma_{pq}$. It follows from Lemma 5.24 for $\alpha = 1$ that for all $2 \leqslant r \leqslant 2\log(ed/k) + t$, $\|z_{pq}\|_{L_r} \leqslant \lambda\sqrt{r}$ where $\lambda = c_0 w^2/\sqrt{N}$. The result now follows from Lemma 5.22. ■

5.2.2 Proof of Theorem 3.4

The proof of Theorem 3.2 relies on several Lemmas. We first use a decomposition similar to the one from Lecué and Mendelson (2018). We have

$$\begin{aligned}
\|\hat{\Sigma}_N - \Sigma\|_\rho &\leqslant \min\left(\sup_{Z \in B_2} \sum_{p,q=1}^{d} Z_{pq}(\hat{\Sigma}_N - \Sigma)_{pq}, \sup_{Z \in \rho B_{SLOPE}} \sum_{p,q=1}^{d} Z_{pq}(\hat{\Sigma}_N - \Sigma)_{pq}\right) \\
&= \min\left(\sqrt{\sum_{p,q=1}^{d} (\hat{\Sigma}_N - \Sigma)_{pq}^2}, \rho \sup_{Z \in \rho B_{SLOPE}} \sum_{p,q=1}^{d} Z^*_{(p,q)} \beta_{pq} \frac{(\hat{\Sigma}_N - \Sigma)^*_{(p,q)}}{\beta_{pq}}\right) \\
&= \min\left(\sqrt{\sum_{p,q=1}^{d} (\hat{\Sigma}_N - \Sigma)_{pq}^2}, \rho \max_{p,q \in [d]} \frac{(\hat{\Sigma}_N - \Sigma)^*_{(p,q)}}{\beta_{pq}}\right). \qquad (101)
\end{aligned}$$

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 3.3. We just have to use it for $k = d$ to handle the left-hand side term of (101). Therefore, with probability at least $1 - \exp(-t)$,

$$\sqrt{\sum_{p,q=1}^{d} (\hat{\Sigma}_N - \Sigma)_{pq}^2} \leqslant c_0 w^2 \sqrt{\frac{d}{N}}.$$

It only remains to handle the second term in the right-hand side inequality of (101). To that end, we start with the following result.

Lemma 5.25 *Let $\boldsymbol{z} := (z_{pq} : p, q \in [d])$ be real-valued random variables (not necessarily independent) and $\lambda, t \geqslant 1$ 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 $\boldsymbol{z}$. Let $p_0, q_0 \in [d]$. We assume that for $r = \log[ed^2/(p_0 q_0)] + t$, we have $\|z_{pq}\|_{L_r} \leqslant \lambda\sqrt{r}$ for all $p, q \in [d]$. Then,*

$$\|\frac{z^*_{(p_0,q_0)}}{\beta_{p_0 q_0}}\|_{L_t} \leqslant e^2 \lambda.$$

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 $j \in [d^2]$. Let $j_0 \in [d^2]$ and set $r_0 = \log(ed^2/j_0) + t$. We assume that $\|z_j\|_{L_{r_0}} \leqslant \lambda\sqrt{r_0}$ for all $p, q \in [d]$. We want to prove that $\|z^*_{j_0}/\beta_{j_0}\|_{L_t} \leqslant e^2\lambda$. We first remark that

$$\frac{z^*_{j_0}}{\beta_{j_0}} \leqslant \max_{I \subset [d^2] : |I| = j_0} \frac{1}{\beta_{j_0}|I|} \sum_{j \in I} |z_j| := \mathcal{Z}. \qquad (102)$$

We define for all $j \in [d^2]$,

$$Z_j = z_j I\left(|z_j| \leqslant e\lambda\sqrt{r_0}\right) \text{ and } Y_j = z_j I\left(|z_j| > e\lambda\sqrt{r_0}\right).$$

It follows from the convexity of $x \in \mathbb{R}_+ \to x^t$ and the definitions above that

$$\mathbb{E}\mathcal{Z}^t \leqslant \max_{I\subset[d^2]:|I|=j_0} \frac{1}{\beta_{j_0}^t |I|}\sum_{j\in I}|z_j|^t \leqslant \left(\frac{e\lambda\sqrt{r_0}}{\beta_{j_0}}\right)^t + \frac{1}{j_0}\sum_{j=1}^{d^2}\frac{\mathbb{E}|Y_j|^t}{\beta_{j_0}^t}. \tag{103}$$

Next, for the second term in the right-hand side inequality of (103) for $\theta := r_0/t$ and all $j \in [d^2]$, we have

$$\begin{aligned}\mathbb{E}|Y_j|^t &= \mathbb{E}\left[|z_j|^t I(|z_j| > e\lambda\sqrt{r_0})\right] \leqslant \mathbb{E}\left[|z_j|^{t\theta}\right]^{1/\theta} \mathbb{P}\left[|z_j| > e\lambda\sqrt{r_0}\right]^{1-1/\theta}\\ &\leqslant (\lambda\sqrt{r_0})^t\left(\frac{\|z_j\|_{L_{r_0}}}{e\lambda\sqrt{r_0}}\right)^{r_0-t} \leqslant (e\lambda)^t r_0^{t/2} e^{-r_0} = (e\lambda)^t\beta_{j_0}^t e^{-t}\frac{j_0}{ed^2}.\end{aligned}$$

We end up in (103) with $\mathbb{E}\mathcal{Z}^t \leqslant (e\lambda)^t + \lambda^t \leqslant (e^2\lambda)^t$. ■

Lemma 5.26 *Let $\boldsymbol{z} := (z_{pq} : p, q \in [d])$ be real-valued random variables (not necessarily independent) and $\lambda \geqslant 0, t \geqslant 3$ 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 $\boldsymbol{z}$. Let $r_0 = \log(ed^2) + t$ and assume that $\|z_{pq}\|_{L_r} \leqslant \lambda\sqrt{r}$ for all $p, q \in [d]$ and $2 \leqslant r \leqslant r_0$. Let $k \in [d]$ and $\gamma \geqslant 1$. Then, when $t \geqslant \max\left(2\log(\lceil\log(k^2)\rceil), \gamma\log(ed^2/k^2)\right)$, with probability at least $1 - \exp(-t/2)$,*

$$\max_{p,q\in[d^2]}\left(\frac{z^*_{(p,q)}}{\beta_{pq}}\right) \leqslant \sqrt{2}e^3\lambda.$$

Proof. We use the same 'vectorial' notation as the one introduced in the proof of Lemma 5.25. We remark that for all $j \in [d^2]$, we have $(1/\beta_{2j}) \leqslant \sqrt{2}/\beta_j$ when $t \geqslant 3$ and for all $j \geqslant k^2$, $1/\beta_j \leqslant \sqrt{2}/\beta_{k^2}$ when $t \geqslant \gamma\log(ed^2/k^2)$, hence,

$$\max_{j\in[d^2]}\left(\frac{z_j^*}{\beta_j}\right) \leqslant \sqrt{2}\max\left(\frac{z^*_{2^j}}{\beta_{2^j}} : j = 0, 1, \ldots, \lceil\log(k^2)\rceil\right).$$

Il follows from Lemma 5.25 that for all $j = 0, 1, \ldots, \lceil\log(k^2)\rceil$, we have $\|z^*_{2^j}/\beta_{2^j}\|_{L_t} \leqslant e^2\lambda$ and so by Markov's inequality with probability at least $1 - \exp(-t)$, $z^*_{2^j}/\beta_{2^j} \leqslant e^3\lambda$. The union bound yields that with probability at least $1 - \lceil\log(k^2)\rceil\exp(-t)$,

$$\max\left(z^*_{2^j}/\beta_{2^j} : j = 0, 1, \ldots, \lceil\log(k^2)\rceil\right) \leqslant e^3\lambda.$$

■

The proof of Theorem 3.2 will follow from Lemma 5.26 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 5.26. 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 3.3. To that end we use Lemma 5.23 to prove the following moment growth condition on the $\hat{\Sigma}_{pq} - \Sigma_{pq}, p, q \in [d]$.

Lemma 5.27 *There exists an absolute constant c_0 such that the following holds. Grant Assumption 3.3 with parameters w and $t \geqslant 3$. For all $p, q \in [d]$ and all $2 \leqslant r \leqslant 2\log(ed^2)+t$, if $N \geqslant 2\log(ed^2) + t$ then $\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leqslant (c_0 w^2/\sqrt{N})\sqrt{r}$.*

Proof. Let $p, q \in [d]$. It follows from Assumption 3.3 and Lemma 5.26 that for all $p, q \in [d]$ and all $2 \leqslant r \leqslant 2\log(ed^2) + t$,

$$\|\hat{\Sigma}_{pq} - \Sigma_{pq}\|_{L_r} \leqslant \frac{1}{\sqrt{N}} \| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_{ip}X_{iq} - \mathbb{E}X_{ip}X_{iq}\|_{L_r} \leqslant \frac{c_0 w^2}{\sqrt{N}}\sqrt{r}.$$

■

Proof of Theorem 3.2 We set for all $p, q \in [d]$, $z_{pq} = \hat{\Sigma}_{pq} - \Sigma_{pq}$. It follows from Lemma 5.27 for $\alpha = 1$ that for all $2 \leqslant r \leqslant 2\log(ed^2)+t$, $\|z_{pq}\|_{L_r} \leqslant \lambda\sqrt{r}$ where $\lambda = c_0 w^2/\sqrt{N}$. The result now follows from Lemma 5.26. ■

5.3 Proofs of section 4

5.3.1 Proof of Lemma 4.2

Let $Z \in \mathcal{C}$ and consider its SVD $Z = \sum_i \sigma_i u_i u_i^\top$. We have

$$\begin{aligned}\langle \Sigma, (\beta^*)(\beta^*)^\top - Z\rangle &= \theta\langle (\beta^*)(\beta^*)^\top, (\beta^*)(\beta^*)^\top - Z\rangle + \langle I_d, (\beta^*)(\beta^*)^\top - Z\rangle \\ &\overset{(i)}{=} \theta\langle (\beta^*)(\beta^*)^\top, (\beta^*)(\beta^*)^\top - \sum_i \sigma_i u_i u_i^\top\rangle \\ &= \theta\left(1 - \sum_i \sigma_i \langle u_i, \beta^*\rangle^2\right) = \theta\sum_i \sigma_i(1 - \langle u_i, \beta^*\rangle^2) \overset{(ii)}{\geqslant} 0\end{aligned}$$

where we used in *(i)* that $\langle I_d, (\beta^*)(\beta^*)^\top - Z\rangle = \mathrm{Tr}\left((\beta^*)(\beta^*)^\top\right) - \mathrm{Tr}\,(Z) = 0$, and in *(ii)* that $|\langle u_i, \beta^*\rangle| \leqslant 1$ (by Cauchy-Schwart). Hence, $\beta^*(\beta^*)^\top$ is a solution to the problem $\max\left(\langle \Sigma, Z\rangle, Z \in \mathcal{C}\right)$. 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.

5.3.2 Proof of Lemma 4.3

Let $Z \in \mathcal{C}$ and consider its SVD $Z = \sum_i \sigma_i u_i u_i^\top$. In the proof of Lemma 4.2, 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

$$\|Z^*-Z\|_2^2 = \mathrm{Tr}\left((Z^*-Z)(Z^*-Z)^\top\right) = \mathrm{Tr}\left(\left(\sum_i \sigma_i((\beta^*)(\beta^*)^\top - u_iu_i^\top)\right)^2\right)$$

$$= \sum_{i,j}\sigma_i\sigma_j \mathrm{Tr}\left((u_iu_i^\top - (\beta^*)(\beta^*)^\top)(u_ju_j^\top - (\beta^*)(\beta^*)^\top)\right) = \sum_i \sigma_i\left(\sigma_i - 2\langle u_i,\beta^*\rangle^2 + 1\right)$$

$$= 2\sum_i \sigma_i(1-2\langle u_i,\beta^*\rangle^2) + \left(\sum_i \sigma_i^2 - \sigma_i\right) = \frac{2}{\theta}\langle \Sigma, Z^*-Z\rangle + \left(\|Z\|_2^2 - \|Z\|_*\right) \leqslant \frac{2}{\theta}\langle \Sigma, Z^*-Z\rangle.$$

5.3.3 Proof of Lemma 4.4

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 $|I| \leqslant k^2$. 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:

$$\begin{aligned}\langle \Phi^*, Z-Z^*\rangle &= \langle \Phi^*, P_I(Z-Z^*)\rangle + \langle \Phi^*, P_{I^c}(Z-Z^*)\rangle \geqslant -|\langle \Phi^*, P_I(Z-Z^*)\rangle| + \|P_{I^c}Z\|_1 \\ &\geqslant -\|\Phi^*\|_\infty\|P_I(Z-Z^*)\|_1 + \|P_{I^c}Z\|_1 = -\|P_I(Z-Z^*)\|_1 + \|P_{I^c}(Z-Z^*)\|_1 \\ &= \|Z-Z^*\|_1 - 2\|P_I(Z-Z^*)\|_1 = \rho - 2\|P_I(Z-Z^*)\|_1.\end{aligned}$$

Now, we have $\|P_I(Z-Z^*)\|_1 \leqslant k\|P_I(Z-Z^*)\|_2 \leqslant k\|Z-Z^*\|_2 \leqslant k\sqrt{r^*(\rho)}$. We conclude that $\langle \Phi^*, Z-Z^*\rangle \geqslant \rho - 2k\sqrt{r^*(\rho)}$. Then, $\sup_{\Phi \in \Gamma_{Z^*}(\rho)} \langle \Phi, Z-Z^*\rangle \geqslant \langle \Phi^*, Z-Z^*\rangle \geqslant \rho - 2k\sqrt{r^*(\rho)}$. Since this is true for any $Z \in H_{\rho,A}$, we conclude that $c \geqslant \rho - 2k\sqrt{r^*(\rho)}$, where $P_I(Z-Z^*)$ is the quantity introduced in Definition 2.11. Then, if we choose ρ such that $\rho \geqslant 10k\sqrt{r^*(\rho)}$, we have $\Delta(\rho, A) \geqslant (4/5)\rho$, and the A-sparsity equation is satisfied by such a ρ.

5.3.4 Proof of Lemma 4.5

From Lemma 4.3, we get that Assumption 2.10 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 3.1 is granted for $t = \log(ed/10k)$ and $w \geqslant 0$. Let then $c_0 > 0$ be the constant provided by Theorem 3.2, 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:

$$100k^2r^* = 100k^2bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\frac{(ed)^4}{4.10^4k^4\log(\frac{ed}{k})}\right)} \leqslant 200k^2bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\left(\frac{ed}{k}\right)^4\right)} = (\rho^*)^2, \tag{104}$$

so that $\rho^* \geqslant 10k\sqrt{r^*}$. Let us then define $k^* := \rho^*/\sqrt{r^*}$. Since $k^* > k$, any $2 \leqslant r \leqslant 2\log(ed/k^*) + t$ satisfies $2 \leqslant r \leqslant 2\log(ed/k) + t$, so that Assumption 3.1 holds with w, t and k^*. We are then in measure to apply Theorem 3.2 with those parameters. As a consequence, as soon as $N \geqslant 2\log(ed/k^*) + t$, one has with probability at least $1 - \exp(-t)$:

$$\|\hat{\Sigma}_N - \Sigma\|_{k^*} \leqslant c_0w^2\sqrt{\frac{(k^*)^2\left(\log\left(\frac{ed}{k^*}\right) + t\right)}{N}} \tag{105}$$

where $\|.\|_{k^*}$ is the ℓ_1/ℓ_2 interpolation norm defined in (13). Now, we have:

$$\sup_{Z\in\mathcal{C}\cap(Z^*+\rho^*B_1\cap\sqrt{r^*}B_2)} |(P-P_N)\mathcal{L}_Z| \leqslant \sqrt{r^*}\sup_{Z\in\mathcal{C}\cap(Z^*+k^*B_1\cap B_2)} |(P-P_N)\mathcal{L}_Z|$$
$$= \sqrt{r^*}\left\|\frac{1}{N}\sum_{i=1}^N X_iX_i^\top - \mathbb{E}[X_iX_i^\top]\right\|_{k^*} = \sqrt{r^*}\|\hat{\Sigma}_N - \Sigma\|_{k^*}. \tag{106}$$

Combining it with (105), we get that with probability at least $1 - \exp(-t)$:

$$\sup_{Z\in\mathcal{C}\cap(Z^*+\rho^*B_1\cap\sqrt{r^*}B_2)} |(P-P_N)\mathcal{L}_Z| \leqslant c_0w^2\sqrt{\frac{(\rho^*)^2\left(\log\left(\frac{ed}{k^*}\right)+t\right)}{N}} \leqslant c_0w^2\sqrt{\frac{2(\rho^*)^2}{N}\log\left(\frac{ed}{10k}\right)} \tag{107}$$

since $k^* \geqslant 10k$. Now, we have:

$$r^* = bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\frac{b^2A^2(ed)^4}{N(\rho^*)^2}\right)} = bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\frac{(ed)^4}{200^2k^4\log(ed/10k)}\right)}$$
$$\geqslant bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\frac{(ed)^3}{200^2k^3}\right)} \geqslant bA\sqrt{\frac{(\rho^*)^2}{N}\log\left(\frac{ed}{10k}\right)}$$

where the last inequality holds since we assumed that $k \leqslant ed/200$. Combining it with (107), we conclude that:

$$\sup_{Z \in \mathcal{C} \cap (Z^* + \rho^* B_1 \cap \sqrt{r^*} B_2)} |(P - P_N)\mathcal{L}_Z| \leqslant \frac{r^*}{3A}$$

which allows us to conclude that $r^*_{RERM,G}(A, \rho^*, e^{-t}) \leqslant r^*$. Moreover, we have from (104) that

$$\rho^* \geqslant 10k\sqrt{r^*} \geqslant 10k\sqrt{r^*_{RERM,G}(A, \rho^*, e^{-t})}$$

that is, ρ^* satisfies the A-sparsity equation from Definition 2.11. These results are valid provided that $N \geqslant 2\log(ed/k^*) + t$, which is ensured by the assumption that $N \geqslant 3\log(ed/10k)$, given that $k^* \geqslant 10k$. This concludes the proof.

5.3.5 Proof of Theorem 4.6

From Lemma 4.3, we get that Assumption 2.10 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 $N \geqslant 3\log(ed/10k)$, Lemma 4.5 applies and so, for ρ^* and $r^*(\rho^*)$ (defined in (20)) we have $r^*_{\text{RERM,G}}(A, \rho^*, 10k/ed) \leqslant r^*(\rho^*)$ and ρ^* satisfies the A-sparsity equation from Definition 2.11. We are then in position to apply Theorem 2.12, provided that λ satisfies (9). 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:

$$bA\sqrt{\frac{3}{N}\log\left(\frac{ed}{200^{2/3}k}\right)} \leqslant \frac{r^*(\rho^*)}{\rho^*} \leqslant bA\sqrt{\frac{4}{N}\log\left(\frac{ed}{200^{1/2}\log(200)^{1/4}k}\right)},$$

since we assumed that $k \leqslant ed/200$. As a consequence, (9) is satisfied as soon as:

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

which is the assumption made in (21). We only have to check that this authorized interval for λ is not empty, which is ensured as soon as $ed/k \geqslant 200^{48/47}/\log(200)^{25/47}$, which is granted by the assumption that $k \leqslant ed/200$.

We are then in measure to apply Theorem 2.12, which enables us to state that, with probability at least $1 - 10k/ed$:

$$\|\hat{Z}_\lambda^{\text{RERM}} - Z^*\|_1 \leqslant \rho^* = 200Abk^2\sqrt{\frac{1}{N}\log\left(\frac{ed}{k}\right)} = 400bk^2\sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)},$$

$$\|\hat{Z}_\lambda^{\text{RERM}} - Z^*\|_2 \leqslant \sqrt{r^*_{\text{RERM,G}}\left(A, \rho^*, 10k/ed\right)} \leqslant \frac{\rho^*}{10k} = 40b\sqrt{\frac{k^2}{N\theta^2}\log\left(\frac{ed}{k}\right)}$$

and

$$P\mathcal{L}_{\hat{Z}_\lambda^{\text{RERM}}} \leqslant A^{-1} r^*_{\text{RERM,G}}\left(A, \rho^*, 10k/ed\right) \leqslant \frac{(\rho^*)^2}{100k^2 A} = 800b^2 \frac{k^2}{N\theta}\log\left(\frac{ed}{k}\right).$$

This concludes the proof.

5.3.6 Proof of Corollary 4.7

From Theorem 4.6, we get the existence of a universal constant $C > 0$ such that with probability at least $1 - 20\,(k/ed)^{3/4}$, $\|\hat{Z}_\lambda^{\text{RERM}} - Z^*\|_2 \leqslant C\sqrt{k^2(N\theta^2)\log(ed/k)}$. 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 $\sin(\Theta(\hat{\beta}, \beta^*)) = (1/\sqrt{2})\|\hat{\beta}\hat{\beta}^\top - \beta^*(\beta^*)^\top\|_2 \leqslant (c_0/g)\|\hat{Z}_\lambda^{\text{RERM}} - Z^*\|_2$ 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 $D = \sqrt{2}c_0 \times C$.

5.3.7 Proof of Lemma 4.8

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 4.4.

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 $|I| \leqslant k^2$. Let P_I (resp. P_{I^c}) be the coordinate projection on I (resp. on I^c).

We know that for $Z \neq 0$:

$$\partial\|.\|_{SLOPE}(Z) = \left\{\Phi \in S^*_{SLOPE} : \left\langle \Phi, Z\right\rangle = \|Z\|_{SLOPE}\right\},$$

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:

$$\sup_{\Phi\in\Gamma_{Z^*}(\rho)} \left\langle \Phi, Z - Z^*\right\rangle \geqslant \sup_{\Phi\in\partial\|.\|_{SLOPE}(Z^*)} \left\langle \Phi, Z - Z^*\right\rangle.$$

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 $1 \leqslant p, q \leqslant d$,

$$\Phi^*_{p,q} = \begin{cases} \text{sgn}(Z^*_{p,q})\, b_{\sigma^{-1}(p,q)} & \text{if } (p,q) \leqslant (k,k); \\ \text{sgn}((Z - Z^*)_{p,q})\, b_{\pi^{-1}(p,q)} & \text{otherwise} \end{cases}$$

and

$$\widetilde{\Phi}^*_{p,q} = \text{sgn}((Z - Z^*)_{p,q})\, b_{\pi^{-1}(p,q)}.$$

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:

$$\begin{aligned} \left\langle \Phi^*, Z - Z^* \right\rangle &= \left\langle \Phi^*, P_I(Z-Z^*) \right\rangle + \left\langle \Phi^*, P_{I^c}(Z-Z^*) \right\rangle \\ &= \sum_{p,q=1}^{k} \operatorname{sgn}(Z^*_{\sigma(p,q)}) b_{p,q} (Z-Z^*)_{\sigma(p,q)} + \left\langle \Phi^*, P_{I^c}(Z-Z^*) \right\rangle . \end{aligned} \tag{108}$$

Regarding the first term, we have:

$$\left| \sum_{p,q=1}^{k} \operatorname{sgn}(Z^*_{\sigma(p,q)}) b_{p,q} (Z-Z^*)_{\sigma(p,q)} \right| \leqslant \sum_{p,q=1}^{k} b_{p,q} \left| (Z-Z^*)_{\pi(p,q)} \right| = \sum_{p,q=1}^{k} b_{p,q} (Z-Z^*)^{\sharp}_{p,q},$$

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:

$$\sum_{p,q=1}^{k} \operatorname{sgn}(Z^*_{\sigma(p,q)}) b_{p,q} (Z-Z^*)_{\sigma_{p,q}} \geqslant - \sum_{p,q=1}^{k} b_{p,q} (Z-Z^*)^{\sharp}_{p,q}. \tag{109}$$

Concerning the second term in (108):

$$\begin{aligned} \left\langle \Phi^*, P_{I^c}(Z-Z^*) \right\rangle = \left\langle \widetilde{\Phi}^*, P_{I^c}(Z-Z^*) \right\rangle &= \left\langle \widetilde{\Phi}^*, Z-Z^* \right\rangle - \left\langle \widetilde{\Phi}^*, P_I(Z-Z^*) \right\rangle \\ &= ||Z-Z^*||_{SLOPE} - \sum_{p,q=1}^{k} b_{\pi^{-1}\circ\sigma(p,q)} (Z-Z^*)^{\sharp}_{\pi^{-1}\circ\sigma(p,q)} \\ &\geqslant ||Z-Z^*||_{SLOPE} - \sum_{p,q=1}^{k} b_{p,q} (Z-Z^*)^{\sharp}_{p,q}. \end{aligned} \tag{110}$$

Putting (108), (109) and (110) together, we obtain

$$\left\langle \Phi^*, Z - Z^* \right\rangle \geqslant ||Z-Z^*||_{SLOPE} - 2\sum_{p,q=1}^{k} b_{p,q} (Z-Z^*)^{\sharp}_{p,q} = \rho - 2\sum_{p,q=1}^{k} b_{p,q} (Z-Z^*)^{\sharp}_{p,q}. \tag{111}$$

Now, since $\|Z-Z^*\|_2 \leqslant \sqrt{r^*_G}$, we can show that for any $k \in [d]$, $(Z-Z^*)^{\sharp}_{kk} \leqslant \frac{\sqrt{r^*_G}}{k}$. Indeed, assume the existence of $k_0 \in [d]$ such that $(Z-Z^*)^{\sharp}_{k_0k_0} > \frac{\sqrt{r^*_G}}{k_0}$. Then by construction we have that for any $(p,q) \leqslant (k_0,k_0)$, $(Z-Z^*)^{\sharp}_{pq} \geqslant (Z-Z^*)^{\sharp}_{k_0k_0}$, so that

$$\|Z-Z^*\|_2^2 = \|(Z-Z^*)^{\sharp}\|_2^2 \geqslant \sum_{(p,q)\leqslant(k_0,k_0)} ((Z-Z^*)^{\sharp}_{pq})^2 > \sum_{(p,q)\leqslant(k_0,k_0)} \frac{r^*_G}{k_0^2} = r^*_G,$$

since the k_0^2 largest elements of $(Z-Z^*)^\sharp$ belong to $[k_0]^2$, as a result of which

$$|\{(p,q):\,(p,q)\leqslant(k_0,k_0)\}|\leqslant k_0^2.$$

This is inconsistent with the fact that $\|Z-Z^*\|_2\leqslant\sqrt{r_G^*}$.

As a consequence, we have:

$$\begin{aligned}&\sum_{p,q=1}^k b_{pq}(Z-Z^*)^\sharp_{pq}=\sum_{\ell=1}^{k-1}\sum_{(\ell,\ell)\leqslant(p,q)<(\ell+1,\ell+1)}b_{pq}(Z-Z^*)^\sharp_{\ell\ell}+b_{kk}(Z-Z^*)^\sharp_{kk}\\ &\leqslant\sum_{\ell=1}^{k-1}|\{(\ell,\ell)\leqslant(p,q)<(\ell+1,\ell+1)\}|\,b_{\ell\ell}(Z-Z^*)^\sharp_{\ell\ell}+b_{kk}\frac{\sqrt{r_G^*}}{k}\\ &\leqslant\sum_{\ell=1}^{k-1}(2\ell+1)b_{\ell\ell}\frac{\sqrt{r_G^*}}{\ell}+b_{kk}\frac{\sqrt{r_G^*}}{k}\leqslant3\sqrt{r_G^*}\sum_{\ell=1}^{k-1}b_{\ell\ell}+b_{kk}\frac{\sqrt{r_G^*}}{k}\leqslant3\sqrt{r_G^*}\sum_{\ell=1}^{k}b_{\ell\ell}=\sqrt{r_G^*}\Gamma_k.\end{aligned}$$

Then, under the assumption that $\rho\geqslant10\Gamma_k\sqrt{r_G^*(A,\rho,\delta)}$, we get from (111) that $\langle\Phi^*,Z-Z^*\rangle\geqslant(4/5)\rho$. and then:

$$\sup_{\Phi\in S^*_{SLOPE}}\langle\Phi,Z-Z^*\rangle\geqslant\langle\Phi^*,Z-Z^*\rangle\geqslant\frac{4}{5}\rho.$$

Since this is true for any $Z\in H(\rho,A)$, we conclude that:

$$\Delta(\rho,A)=\inf_{Z\in H(\rho,A)}\sup_{\Phi\in S^*_{SLOPE}}\langle\Phi^*,Z-Z^*\rangle\geqslant\frac{4}{5}\rho.$$

that is, ρ satisfies the A-sparsity equation from Definition 2.11.

5.3.8 Proof of Lemma 4.9

From Lemma 4.3, we get that Assumption 2.10 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 $\mathcal{C}_{r,\rho}:=\{Z\in\mathcal{C}:\|Z-Z^*\|_{SLOPE}\leqslant\rho,\|Z-Z^*\|_2\leqslant\sqrt{r}\}$. Let $A>0$. For any ρ and $r>0$. We have

$$\begin{aligned}&\sup_{Z\in\mathcal{C}_{r,\rho}}|\langle\Sigma-\hat\Sigma_N,Z-Z^*\rangle|\leqslant\sup_{Z\in(\rho B_{SLOPE}\cap\sqrt{r}B_2)}|\langle\Sigma-\hat\Sigma_N,Z\rangle|\\ &=\sqrt{r}\sup_{Z\in(\frac{\rho}{\sqrt{r}}B_{SLOPE}\cap B_2)}|\langle\Sigma-\hat\Sigma_N,Z\rangle|=\sqrt{r}\|\Sigma-\hat\Sigma_N\|_{\frac{\rho}{\sqrt{r}}}\end{aligned}\tag{112}$$

where $\|.\|_{\rho/\sqrt{r}}$ is the $SLOPE/\ell_2$ interpolation norm defined in (14). Assumption 3.3 is granted for $t=2\log(ed^2/k^2)$. Let us now check that $k\leqslant d/(e^2\log(d))$: we have that $k^2\log(ek^2)\leqslant ed^2$, hence,

$$2\log(\lceil\log(k^2)\rceil)\leqslant2\log(\log(k^2)+1)=2\log(\log(ek^2))\leqslant2\log\left(\frac{ed^2}{k^2}\right),$$

that is, $t \geqslant \max\left(2\log(\lceil \log(k^2)\rceil), 2\log(ed^2/k^2)\right)$. We are then in position to apply Theorem 3.4 with $\gamma = 2$ and $t = 2\log(ed^2/k^2)$: there exists a universal constant $c_0 > 0$ such that, provided that $N \geqslant \log\left(ed^2\right) + t$, one has with probability at least $1 - 2\exp(-t/2)$:

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

Plugging this last result into (112), we get that:

$$\sup_{Z \in \mathcal{C}_{r,\rho}} |\langle \Sigma - \hat{\Sigma}_N, Z - Z^* \rangle| \leqslant \sqrt{r}\frac{c_0 w^2}{\sqrt{N}} \min\left(\frac{\rho}{\sqrt{r}}, d\right) \tag{113}$$

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 $(\sqrt{r^*(\rho)}c_0 w^2/\sqrt{N}) \min\left(\rho/\sqrt{r^*(\rho)}, d\right) \leqslant r^*(\rho)/3A$ whatever the value of ρ is. From (113) we then deduce that $r^*_{\text{RERM,G}}(A, \rho, 2e^{-t/2}) \leqslant r^*(\rho)$. 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 $\rho^* \geqslant 10\Gamma^*_k r^*(\rho^*)^{1/2} \geqslant 10\Gamma^*_k r^*_{\text{RERM,G}}\left(A, \rho^*, 2e^{-t^*/2}\right)^{1/2}$ which, according to Lemma 4.8, guarantees that ρ^* satisfies the A-sparsity equation from Definition 2.11. Finally, plugging the expression of ρ^* into the one of $r^*(\rho^*)$, we get that $r^*(\rho^*) = (b^2A^2/N) \min\left(d, 10\Gamma^*_k\right)^2$. Finally, the previous results hold provided that $N \geqslant \log\left(ed^2\right) + t$, which is granted by the assumption that $N \geqslant 3\log(ed^2)$. This concludes the proof, noting that $2\exp(-t/2) = 2k^2/(ed^2)$.

5.3.9 Proof of Theorem 4.10

From Lemmas 4.2 and 4.3, we get that Assumption 2.10 holds with $G : Z \to \|Z\|_2^2$ and $A = 2/\theta$. From Lemma 4.9, we get the existence of a constant $b > 0$ such that, provided that $N \geqslant 3\log(ed^2)$, 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 $r^*_{\text{RERM,G}}(A, \rho^*, 2k^2/ed^2) \leqslant r^*$ and ρ^* satsifies the A-sparsity equation from Definition 2.11. Let us now upper bound Γ^*_k:

$$\begin{aligned}\Gamma^*_k = 3\sum_{\ell=1}^k b_{\ell\ell}\left(2\log\left(\frac{ed^2}{k^2}\right)\right) &\leqslant 3\left(\sum_{\ell=1}^k \sqrt{\log\left(\frac{ed^2}{\ell^2}\right)} + \sum_{\ell=1}^k \sqrt{2\log\left(\frac{ed^2}{k^2}\right)}\right) \\ &\leqslant 3\sum_{\ell=1}^k \sqrt{\log\left(\frac{ed^2}{\ell^2}\right)} + 3k\sqrt{2\log\left(\frac{ed^2}{k^2}\right)}.\end{aligned} \tag{114}$$

Concerning the first term in this last inequality, we have:

$$\left(\sum_{\ell=1}^{k}\sqrt{\log\left(\frac{ed^2}{\ell^2}\right)}\right)^2 = 2\sum_{m<\ell}\sqrt{\log\left(\frac{ed^2}{\ell^2}\right)}\sqrt{\log\left(\frac{ed^2}{m^2}\right)} + \sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right)$$
$$\leqslant 2\sum_{m<\ell}\log\left(\frac{ed^2}{\ell^2}\right) + \sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right) \leqslant 3k\sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right). \tag{115}$$

Moreover, we have:

$$\sum_{\ell=1}^{k}\log\left(\frac{ed^2}{\ell^2}\right) \leqslant \sum_{\ell=1}^{k}\int_{u=\ell-1}^{\ell}\log\left(\frac{ed^2}{u^2}\right)\mathrm{d}u = \int_{u=0}^{k}\log\left(\frac{ed^2}{u^2}\right) = k\log(ed^2) - 2\left[u\log(u)-u\right]_0^k$$
$$= k\log\left(\frac{ed^2}{k^2}\right) + 2k \leqslant 3k\log\left(\frac{ed^2}{k^2}\right). \tag{116}$$

Combining (114), (115) and (116), we finally get that $\Gamma_k^* \leqslant (9+3\sqrt{2})k\sqrt{\log\left(\frac{ed^2}{k^2}\right)} \leqslant 14k\sqrt{\log\left(ed^2/k^2\right)}$. As a consequence, we have

$$10\Gamma_k^* \leqslant 140k\sqrt{\log\left(\frac{ed^2}{k^2}\right)} \leqslant 140k\sqrt{2\log\left(d\right)} \leqslant 140k\sqrt{\frac{2d}{e^2k}} \leqslant d,$$

since we assumed that $k \leqslant \min\left(d/(e^2\log(d)), (e/(140\sqrt{2}))^2 d\right)$. We conclude that $\min\left(10\Gamma_k^*, d\right) = 10\Gamma_k^* \leqslant 140k\sqrt{\log\left(ed^2/k^2\right)}$. Plugging this result into the expression of r^* and ρ^*, we finally get that:

$$r^* \leqslant 140^2 b^2 A^2 \frac{k^2}{N}\log\left(\frac{ed^2}{k^2}\right) \quad \text{and} \quad \rho^* \leqslant 140^2 bA\frac{k^2}{\sqrt{N}}\log\left(\frac{ed^2}{k^2}\right)$$

so that $r^*/\rho^* = bA/\sqrt{N}$. As a consequence, (9) is satisfied as soon as:

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

which is (23). We are then in position to apply Theorem 2.12, which allows us to conclude that, with probability at least $1-2k^2/ed^2$:

$$\|\hat{Z}_\lambda^{RERM} - Z^*\|_{SLOPE} \leqslant \rho^* \quad , \quad G(\hat{Z}_\lambda^{RERM} - Z^*) \leqslant r^* \quad \textit{and} \quad P\mathcal{L}_{\hat{Z}_\lambda^{RERM}} \leqslant A^{-1}r^*.$$

This concludes the proof.

5.3.10 Proof of Corollary 4.11

The proof follows exactly the same lines as the one of Corollary 4.7, so we do not detail it here.

5.3.11 Proof of Lemma 4.12

Consider $A = 2/\theta$ and $\gamma > 0$. In the rest of the proof we write $r^*(\rho)$ for $r^*_{\text{RMOM,G}}(A, \gamma, \rho)$. For any $J \subset [d]^2$, let P_J be the coordinate projection on J. Consider $\rho > 0$. Let $I := \text{supp}(Z^*)$ be the set of non-zero coefficients of Z^*. From Lemma 4.2, we have that $|I| \leqslant k^2$. Moreover, we know that for any $Z \neq 0$, $\partial \|.\|_1(Z) = \{\Phi \in S_\infty : \langle \Phi, Z \rangle = \|Z\|_1\}$, 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 $Z \in \bar{H}_{\rho,A}$ that is $\|Z - Z^*\|_1 = \rho$ and $\|Z - Z^*\|_2 \leqslant \sqrt{2/\theta} r^*(\rho)$. 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:

$$\begin{aligned}\langle \Phi^*, Z - Z^* \rangle &= \langle \Phi^*, P_I(Z - Z^*) \rangle + \langle \Phi^*, P_I^c(Z - Z^*) \rangle \geqslant -\|\Phi^*\|_\infty \|P_I(Z - Z^*)\|_1 + \langle \Phi^*, P_I^c(Z) \rangle \\ &= -(\|Z - Z^*\|_1 - \|P_I(Z - Z^*)\|_1) + \|P_I^c(Z - Z^*)\|_1 \\ &= 2\|P_I^c(Z - Z^*)\|_1 - \|Z - Z^*\|_1 = \|Z - Z^*\|_1 - 2\|P_I(Z - Z^*)\|_1 = \rho - 2\|P_I(Z - Z^*)\|_1 \end{aligned} \tag{117}$$

where we used the fact that $\|\Phi^*\|_\infty = 1$. Then, since $Z \in \bar{H}_{\rho,A}$, we have:

$$\|P_I(Z - Z^*)\|_1 \leqslant k \|P_I(Z - Z^*)\|_2 \leqslant k\|Z - Z^*\|_2 \leqslant k\sqrt{\frac{2}{\theta}} r^*(\rho) \tag{118}$$

Combining (117) and (118), we finally get that:

$$\langle \Phi^*, Z - Z^* \rangle \geqslant \rho - 2k\sqrt{\frac{2}{\theta}} r^*(\rho). \tag{119}$$

As a consequence, $\sup_{\Phi \in \Gamma_{Z^*}(\rho)} \langle \Phi, Z - Z^* \rangle \geqslant \langle \Phi^*, Z - Z^* \rangle \geqslant \rho - 2k\sqrt{2/\theta} r^*(\rho)$. This being true whatever $Z \in \bar{H}_{\rho,A}$, it follows that $\bar{\Delta}(\rho) \geqslant \rho - 2k\sqrt{2/\theta} r^*(\rho)$. We conclude that any ρ such that $\rho \geqslant 10k\sqrt{2/\theta} r^*(\rho)$ satisfies $\bar{\Delta}(\rho) \geqslant (4/5)\rho$.

5.3.12 Proof of Lemma 4.13

Consider $\gamma > 0$. From Lemma 4.3, we get that Assumption 2.31 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 3.1 is granted for $t = 1$ and $w \geqslant 0$. Let then $c_0 > 0$ be the constant provided by Theorem 3.2, and consider $B := 3c_0 w^2$ and $D := 1600 w^2$. 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:

$$(1) \quad \rho \geqslant 10k\sqrt{\frac{2B\rho}{\theta\gamma}}\left(\frac{6}{N}\log\left(\frac{2B(ed)^2}{\gamma\theta\rho}\sqrt{\frac{6}{N}}\right)\right)^{1/4} \quad \text{and} \quad (2) \quad \rho \geqslant 10kD\sqrt{\frac{2K}{N\theta^2}}, \tag{120}$$

so that $\rho^* \geqslant 10k\sqrt{2/\theta}r^*$. Let us define $k^* = \sqrt{\theta/2}\rho^*/r^*$. We have $\log(ed/k^*) + 1 \leqslant \log(ed/10k) + 1$, so that Assumption 3.1 still holds with w, $t = 1$ and k^*. Then, since we assumed that $N \geqslant 2\log(ed/k) + 1 \geqslant 2\log(ed/k^*) + 1$, Theorem 3.2 applies and allows us to affirm that

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_i\widetilde{X}_i^\top - \mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top]\right\|_{k^*}\right] \leqslant c_0 w^2\sqrt{\frac{6(k^*)^2\log(ed/k^*)}{N}}, \tag{121}$$

where $\|\cdot\|_{k^*}$ is the ℓ_1/ℓ_2 interpolation norm defined in (13) for $k = k^*$.

For r and $\rho > 0$, define $\mathcal{C}_{r,\rho} := \left\{Z \in \mathcal{C} : \|Z - Z^*\|_1 = \rho \text{ and } \|Z - Z^*\|_2 \leqslant \sqrt{2/\theta}r^*(\rho)\right\}$. Let us now upper bound $E_G(r^*, \rho^*)$ and $V_{K,G}(r^*, \rho^*)$ from Definition 2.29.

Bounding the complexity term $E_G(r^*, \rho^*)$. Let $\sigma_1, \dots, \sigma_N$ be *i.i.d.* rademacher variables independent from the $\widetilde{X}_i$'s. We have $\mathcal{C}_{r^*,\rho^*} \subset \sqrt{2/\theta}r^*\left(k^*B_1 \cap B_2\right)$. As a consequence:

$$\begin{aligned}\sup_{Z\in\mathcal{C}_{r^*,\rho^*}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right| &\leqslant \sqrt{\frac{2}{\theta}}r^*\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}_Z(\widetilde{X}_i)\right| \\ &= \sqrt{\frac{2}{\theta}}r^*\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\frac{1}{N}\sum_{i=1}^{N}\sigma_i\widetilde{X}_i\widetilde{X}_i^\top, Z\right\rangle\right|. \end{aligned}\tag{122}$$

Now, it follows from the desymmetrization inequality (see Theorem 2.1 in Koltchinskii (2011b)) that:

$$\begin{aligned}&\mathbb{E}\left[\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\frac{1}{N}\sum_{i=1}^{N}\sigma_i\widetilde{X}_i\widetilde{X}_i^\top, Z\right\rangle\right|\right] \\ &\leqslant 2\mathbb{E}\left[\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_i\widetilde{X}_i^\top - E[\widetilde{X}_i\widetilde{X}_i^\top], Z\right\rangle\right|\right] \\ &\quad + \frac{2}{\sqrt{N}}\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right| \\ &\leqslant 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_i\widetilde{X}_i^\top - E[\widetilde{X}_i\widetilde{X}_i^\top]\right\|_{k^*}\right] + \frac{2}{\sqrt{N}}\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right| \\ &\leqslant 2c_0w^2\sqrt{\frac{6(k^*)^2\log(ed/k^*)}{N}} + \frac{2}{\sqrt{N}}\sup_{Z\in(k^*B_1\cap B_2)\cap(\mathcal{C}-Z^*)}\left|\left\langle\mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right|, \end{aligned}\tag{123}$$

where we used (121) in the last inequality.

Concerning the second term in (123), we have for any $Z \in (k^* B_1 \cap B_2) \cap (\mathcal{C} - Z^*)$:

$$\left\langle \mathbb{E}\left[\widetilde{X}\widetilde{X}^\top, Z\right]\right\rangle = \left\langle \theta\beta^*(\beta^*)^\top + Id, Z\right\rangle \overset{(i)}{=} \theta\left\langle \beta^*(\beta^*)^\top, Z\right\rangle \overset{(ii)}{\leqslant} \theta\|\beta^*\|_2^2\|Z\|_2 \leqslant \theta$$

where we used the fact that $\left\langle Id, Z\right\rangle = \mathrm{Tr}(Z) = 0$ in (i) and Cauchy-Schwarz in (ii). as a consequence:

$$\sup_{Z \in (k^* B_1 \cap B_2) \cap (\mathcal{C} - Z^*)} \left|\left\langle \mathbb{E}\left[\widetilde{X}\widetilde{X}^\top\right], Z\right\rangle\right| \leqslant \theta. \tag{124}$$

Combining (122), (123) and (124), we finally get that:

$$\begin{aligned} E_G(r^*, \rho^*) = \mathbb{E}\left[\sup_{\mathcal{C}_{r^*,\rho^*}} \left|\frac{1}{N}\sum_{i=1}^N \sigma_i \mathcal{L}_Z(\widetilde{X}_i)\right|\right] &\leqslant \sqrt{\frac{2}{\theta}} r^* \left(2c_0 w^2 \sqrt{\frac{6(k^*)^2\log(ed/k^*)}{N}} + \frac{2\theta}{\sqrt{N}}\right) \\ &\leqslant \sqrt{2\theta} r^* \left(3c_0 w^2 \sqrt{\frac{6(k^*)^2\log(ed/k^*)}{\theta^2 N}}\right), \end{aligned} \tag{125}$$

where we used the assumption that $\theta \leqslant k \leqslant k^*$.

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:

$$\begin{aligned} &\boldsymbol{V}ar(\mathcal{L}_Z(\widetilde{X})) = \mathbb{E}[((\mathcal{L}_Z(\widetilde{X}) - P(\mathcal{L}_Z(\widetilde{X}))^2] = \mathbb{E}[\left\langle \widetilde{X}\widetilde{X}^\top - \mathbb{E}[\widetilde{X}\widetilde{X}^\top], Z - Z^*\right\rangle^2] \\ &= \sum_{p,q,s,t=1}^d \mathbb{E}\left[(\widetilde{X}^{(p)}\widetilde{X}^{(q)} - \mathbb{E}[\widetilde{X}^{(p)}\widetilde{X}^{(q)}])(\widetilde{X}^{(s)}\widetilde{X}^{(t)} - \mathbb{E}[\widetilde{X}^{(s)}\widetilde{X}^{(t)}])\right](Z - Z^*)_{pq}(Z - Z^*)_{st} \\ &= \sum_{p,q,s,t=1}^d T_{p,q,s,t}(Z - Z^*)_{pq}(Z - Z^*)_{st} \end{aligned}$$

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 $1 \leqslant p, q, s, t \leqslant d$. Remembering that Assumption 3.1 is granted, we have:

$$T_{p,q,s,t} = \begin{cases} \|(\widetilde{X}^{(p)})^2 - \mathbb{E}[(\widetilde{X}^{(p)})^2]\|_{L_2}^2 \leqslant (2w^2)^2 & \text{if } p = q = s = t \\ \|\widetilde{X}^{(p)}\widetilde{X}^{(q)} - \mathbb{E}[\widetilde{X}^{(p)}\widetilde{X}^{(q)}]\|_{L_2}^2 \leqslant (2w^2)2 & \text{if } (p, q) = (s, t), p \neq q \\ 0 \text{ otherwise.} & \end{cases}$$

Then:

$$\begin{aligned} \boldsymbol{V}ar(\mathcal{L}_Z(\widetilde{X})) &\leqslant \sum_{p=1}^d 4w^4(Z - Z^*)_{pp}^2 + \sum_{p \neq q} 4w^4(Z - Z^*)_{pq}^2 \\ &= \sum_{p,q=1}^q 4w^4(Z - Z^*)_{pq}^2 = 4w^4\|Z - Z^*\|_2^2 \leqslant (8w^4/\theta)(r^*)^2. \end{aligned}$$

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:

$$V_{K,G}(r,\rho) \leqslant 2w^2\sqrt{\frac{2K}{N\theta}}r^*. \tag{126}$$

Combining (125) and (126), we finally get that:

$$\max\left(\frac{E_G(r^*,\rho^*)}{\gamma}, 400\sqrt{2}V_{K,G}(r^*,\rho^*)\right) \leqslant \max\left(\frac{B}{\gamma}\sqrt{\frac{6(\rho^*)^2}{N}\log\left(\sqrt{\frac{2}{\theta}}\frac{edr^*}{\rho^*}\right)}, D\sqrt{\frac{K}{N\theta}}r^*\right)$$

Now, one can check that r^* satisfies both of the two conditions below:

$$(3)\quad \frac{B}{\gamma}\sqrt{\frac{6(\rho^*)^2}{N}\log\left(\sqrt{\frac{2}{\theta}}\frac{edr^*}{\rho^*}\right)} \leqslant (r^*)^2 \quad \text{and} \quad (4)\quad D\sqrt{\frac{K}{N\theta}}r^* \leqslant (r^*)^2$$

Then, we have:

$$\max\left(\frac{E_G(r^*,\rho^*)}{\gamma}, 400\sqrt{2}V_{K,G}(r^*,\rho^*)\right) \leqslant (r^*)^2$$

which, according to Definition 2.29, allows us to conclude that $r^*_{\text{RMOM,G}}(\gamma,\rho^*) \leqslant r^*$. Moreover, we have from (120) that $\rho^* \geqslant \sqrt{2/\theta}10kr^* \geqslant \sqrt{2/\theta}10kr^*_{\text{RMOM,G}}(\gamma,\rho^*)$, that is, ρ^* satisfies the sparsity equation from Definition 2.30. This concludes the proof.

5.3.13 Proof of Theorem 4.14

The assumptions of Lemma 4.13 are met, which gives us the existence of two positive constants B and D such that, defining

$$\rho^* := \max\left(400\sqrt{3}Bk^2\gamma^{-1}\sqrt{(N\theta^2)^{-1}\log(ed/k)}; 10Dk\sqrt{2K(N\theta^2)^{-1}}\right)$$

and

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

one has $r^*_{\text{RMOM,G}}(\gamma,\rho^*) \leqslant r^*(\gamma,\rho^*)$ and ρ^* satisfies the sparsity equation from Definition 2.30. From Lemma 4.3, we get that Assumption 2.31 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 2.32 are met. Then, fixing $\gamma = 1/32000$ and defining $\lambda = (11r^*(\gamma,2\rho^*))/(40\rho^*)$, it is true that with probability at least $1 - 2\exp(-72K/625)$,

$$\|\hat{Z}^{\text{RMOM}}_{K,\lambda} - Z^*\|_1 \leqslant 2\rho^*,\ P\mathcal{L}_{\hat{Z}^{\text{RMOM}}_{K,\lambda}} \leqslant \frac{93}{100}(r^*(\gamma,2\rho^*))^2 \ \text{ and } \|\hat{Z}^{\text{RMOM}}_{K,\lambda} - Z^*\|_2 \leqslant \sqrt{\frac{2}{\theta}}r^*(\gamma,2\rho^*). \tag{127}$$

Now, we can write:

$$\rho^* \leqslant D_1\frac{k}{\sqrt{N\theta^2}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right) \tag{128}$$

with $D_1 := \max(400\sqrt{3}B\gamma^{-1}, 10\sqrt{2}D)$. On the other hand, since $d \geqslant k$, we get that:

$$\rho^* \geqslant D_2 k^2 \sqrt{\frac{1}{N\theta^2}\log\left(\frac{ed}{k}\right)} \geqslant D_2 \frac{k^2}{\sqrt{N\theta^2}}$$

where $D_2 := 400\sqrt{3}B\gamma^{-1}$. As a consequence, we have:

$$\log\left(\frac{B(ed)^2}{\gamma\theta\rho*}\sqrt{\frac{6}{N}}\right) \leqslant \log\left(\frac{B(ed)^2}{\gamma\theta}\sqrt{\frac{6}{N}}\,\frac{\sqrt{N\theta^2}}{D_2k^2}\right) = \log\left(\frac{1}{4\sqrt{5}}\left(\frac{ed}{k}\right)^2\right) \leqslant 2\log\left(\frac{ed}{k}\right),$$

so that:

$$\begin{aligned}
r^*(\gamma, 2\rho^*) &\leqslant \max\left(\sqrt{\frac{2B\rho^*}{\gamma}}\left(\frac{12}{N}\log\left(\frac{ed}{k}\right)\right)^{1/4}; D\sqrt{\frac{K}{N\theta}}\right) \\
&\leqslant \max\left(\sqrt{\frac{2B}{\gamma}}\left(\frac{12}{N}\log\left(\frac{ed}{k}\right)\right)^{1/4}\frac{\sqrt{D_1k}}{(N\theta^2)^{1/4}}\max\left(\sqrt{k}\log\left(\frac{ed}{k}\right)^{1/4}; K^{1/4}\right); D\sqrt{\frac{K}{N\theta}}\right) \\
&\leqslant \max\left(\sqrt{\frac{2BD_1}{\gamma N\theta}}12^{1/4}\max\left(\sqrt{k}\log\left(\frac{ed}{k}\right)^{1/4}; K^{1/4}\right)^2; D\sqrt{\frac{K}{N\theta}}\right) \\
&\leqslant \frac{C}{\sqrt{N\theta}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right), \qquad (129)
\end{aligned}$$

where $C := \max\left(12^{1/4}\sqrt{2BD_1\gamma^{-1}}; D\right)$. Combining (127), (128) and (129), we finally get that, with probability at least $1 - 2\exp(-72K/625)$:

$$\|\hat{Z}_{K,\lambda}^{\mathrm{RMOM}} - Z^*\|_1 \leqslant 2D_1\frac{k}{\sqrt{N\theta^2}}\max\left(k\sqrt{\log\left(\frac{ed}{k}\right)}; \sqrt{K}\right)$$

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

and

$$P\mathcal{L}_{\hat{Z}_{K,\lambda}^{\mathrm{RMOM}}} \leqslant \frac{93C^2}{100N\theta}\max\left(k^2\log\left(\frac{ed}{k}\right); K\right).$$

This concludes the proof.

5.3.14 Proof of Corollary 4.15

From Theorem 4.14, we get the existence of a universal constant $C_2 > 0$ such that with probability at least $1 - \exp(-72K/625)$, $\|\hat{Z}_{K,\lambda}^{\mathrm{RMOM}} - Z^*\|_2 \leqslant C_2(N\theta^2)^{-1/2}\max\left(k\sqrt{\log(ed/k)}; \sqrt{K}\right)$. 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 $\sin(\Theta(\hat{\beta}, \beta^*)) = (1/\sqrt{2})\|\hat{\beta}\hat{\beta}^\top - \beta^*(\beta^*)^\top\|_2 \leqslant (c_0/g)\|\hat{Z}_{K,\lambda}^{\mathrm{RMOM}} - Z^*\|_2$ 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 $D = \sqrt{2}c_0 \times C_2$.

Appendix A.

A.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 1 is convex. We recall the definition of this set:

$$\mathcal{C} := \left\{ Z \in \mathbb{R}^{d \times d} : Z \succeq 0, \sum_{i,j=1}^{M} \left\langle (Y_i - Y_j)\,(Y_i - Y_j)^\top, Z \right\rangle^{1/2} \geqslant 1 \right\}$$

where $(Y_i)_{i=1}^N$ are N given points in $\mathbb{R}^d$. Fot the sake of simplicity, we define, for $(i,j) \in [d]^2$, $V_{ij} = (Y_i - Y_j) \in \mathbb{R}^d$. 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:

$$\begin{aligned}
\left(\sum_{i,j=1}^{M} \left\langle V_{ij} V_{ij}^\top, Z' \right\rangle^{1/2} \right)^2 &= \sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z' \right\rangle + \sum_{(i,j)\neq(pq)} \left\langle V_{ij} V_{ij}^\top, Z' \right\rangle^{1/2} \left\langle V_{pq} V_{pq}^\top, Z' \right\rangle^{1/2} \\
&\geqslant \sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z' \right\rangle = t \sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z_1 \right\rangle + (1-t) \sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z_2 \right\rangle \\
&= t \left(\sqrt{\sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z_1 \right\rangle} \right)^2 + (1-t) \left(\sqrt{\sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z_2 \right\rangle} \right)^2 \\
&\geqslant t \left(\sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z_1 \right\rangle^{1/2} \right)^2 + (1-t) \left(\sum_{(i,j)\in[N]^2} \left\langle V_{ij} V_{ij}^\top, Z_2 \right\rangle^{1/2} \right)^2 \\
&\geqslant t + (1-t) = 1
\end{aligned}$$

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.

A.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 $\mathcal{C}_{\rho,r} = \{Z \in \mathcal{C} : \|Z - Z^*\| \leqslant \rho, G(Z - Z^*) \leqslant r\}$ 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

$$V(r,\rho) = \sup_{Z\in\mathcal{C}_{\rho,r}} \sqrt{\operatorname{Var}(\mathcal{L}_Z)}.$$

Let θ and τ be two positive constants. We consider a local complexity fixed point: for all $\rho > 0$,

$$r^*(\rho) = \inf\left(r > 0 : \max\left(\theta E(r,\rho), \tau V(r,\rho)\right) \leqslant r^2 \right).$$

Proposition A.1 *We assume that $\mathcal{C}$ is star-shaped in Z^*. We assume that G is such that for all $\alpha \geqslant 1$ and all $Z \in \mathcal{C}, G(\alpha(Z - Z^*)) \geqslant \alpha G(Z - Z^*)$. Then, for all $\rho > 0$ and $b \geqslant 1$, we have $r^*(\rho) \leqslant r^*(b\rho) \leqslant \sqrt{b} r^*(\rho)$.*

Proof. Let $\rho > 0$ and $b \geqslant 1$. For all $r > 0$, $\mathcal{C}_{\rho,r} \subset \mathcal{C}_{b\rho,r}$ and so $r^*(\rho) \leqslant r^*(b\rho)$. Let us now prove the second inequality.

We start with some homogeneity property of the complexity and variance terms:

$$E(\sqrt{b}r, b\rho) \leqslant bE(r, \rho) \text{ and } V(\sqrt{b}r, b\rho) \leqslant bV(r, \rho). \tag{130}$$

We prove (130) 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 $b \geqslant 1$ and $\mathcal{C}$ is star-shaped in Z^*, $Z_0 \in \mathcal{C}$. Moreover, $b\|Z_0 - Z^*\| = \|Z - Z^*\| \leqslant b\rho$ and, by the property of G, $bG(Z_0 - Z^*) \leqslant G(Z - Z^*) \leqslant br^2$. 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

$$\sup_{Z \in \mathcal{C}_{b\rho,\sqrt{b}r}} \left| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathcal{L}_Z(X_i) \right| \leqslant b \sup_{Z_0 \in \mathcal{C}_{\rho,r}} \left| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathcal{L}_{Z_0}(X_i) \right| \tag{131}$$

and so (130) holds for the complexity term. It also holds for the variance using similar tools.

Next, it follows from (130) that

$$\begin{aligned} r^*(b\rho) &= \inf\left(r > 0 : \max\left(\theta E(r, b\rho), \tau V(r, b\rho)\right) \leqslant r^2\right) \\ &= \inf\left(r > 0 : \max\left(\theta E\left(\sqrt{b}\frac{r}{\sqrt{b}}, b\rho\right), \tau V\left(\sqrt{b}\frac{r}{\sqrt{b}}, b\rho\right)\right) \leqslant r^2\right) \\ &\leqslant \inf\left(r > 0 : \max\left(\theta E\left(\frac{r}{\sqrt{b}}, \rho\right), \tau V\left(\frac{r}{\sqrt{b}}, \rho\right)\right) \leqslant \left(\frac{r}{\sqrt{b}}\right)^2\right) \leqslant \sqrt{b} r^*(\rho). \end{aligned}$$

■

A.3 A property of the sparsity equation

We consider the same setup as in Section A.2 and define for all $\rho > 0$,

$$H_\rho = \left\{Z \in \mathcal{C} : \|Z - Z^*\| = \rho, G(Z - Z^*) \leqslant (r^*(\rho))^2\right\}, \Gamma_{Z^*}(\rho) = \bigcup_{Z : \|Z - Z^*\| \leqslant \rho/20} \partial\|\cdot\|(Z)$$

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 $\Delta(\rho) \geqslant c_0\rho$ 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 A.2 *We assume that $\mathcal{C}$ is star-shaped in Z^*. We assume that G is such that for all $\alpha \geqslant 1$ and all $Z \in \mathcal{C}, G(\alpha(Z - Z^*)) \geqslant \alpha G(Z - Z^*)$. Let $0 < c_0 < 1$. Then, for all $\rho > 0$ and $b \geqslant 1$, if ρ is such that $\Delta(\rho) \geqslant c_0\rho$ then $\Delta(b\rho) \geqslant c_0 b\rho$.*

Proof. Let $\rho > 0$ be such that $\Delta(\rho) \geqslant c_0\rho$ and let $b \geqslant 1$. Let $Z \in H_{b\rho}$. Let us show that there exists $\Phi \in \Gamma_{Z^*}(b\rho)$ such that $\left\langle \Phi, Z - Z^* \right\rangle \geqslant c_0 b\rho$.

Let Z_0 be such that $Z = Z^* + b(Z_0 - Z^*)$. Since $b \geqslant 1$ 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 A.1, $bG(Z_0 - Z^*) \leqslant G(Z - Z^*) \leqslant (r^*(b\rho))^2 \leqslant b(r^*(\rho))^2$. Therefore, we have $Z_0 \in H_\rho$. But, since we assumed that $\Delta(\rho) \geqslant c_0\rho$, there exists $\Phi \in \Gamma_{Z^*}(\rho)$ such that $\left\langle \Phi, Z_0 - Z^* \right\rangle \geqslant c_0\rho$ and so $\left\langle \Phi, Z - Z^* \right\rangle \geqslant c_0 b\rho$. We conclude the proof by noting that $\Gamma_{Z^*}(\rho) \subset \Gamma_{Z^*}(b\rho)$ and so $\Phi \in \Gamma_{Z^*}(b\rho)$. ■

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147, 1999. ISSN 0022-0000. doi: 10.1006/jcss.1997.1545. URL https://doi.org/10.1006/jcss.1997.1545. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005. ISSN 0091-1798. doi: 10.1214/009117905000000233. URL https://doi.org/10.1214/009117905000000233.

A.S Bandeira, S. Boumal, and A. Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, 2016.

Nikhil Bansal, Daniel Dadush, and Shashwat Garg. An algorithm for Komlós conjecture matching Banaszczyk's bound. *SIAM J. Comput.*, 48(2):534–553, 2019. ISSN 0097-5397,1095-7111. doi: 10.1137/17M1126795. URL https://doi.org/10.1137/17M1126795.

Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets Lasso: improved oracle bounds and optimality. *Ann. Statist.*, 46(6B):3603–3642, 2018. ISSN 0090-5364,2168-8966. doi: 10.1214/17-AOS1670. URL https://doi.org/10.1214/17-AOS1670.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005. ISSN 1292-8100,1262-3318. doi: 10.1051/ps:2005018. URL https://doi.org/10.1051/ps:2005018.

Stéphane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration inequalities : a non asymptotic theory of independence*. Oxford University Press, 2013. URL https://hal.inria.fr/hal-00942704.

Stephen Boyd and Lieven Vandenberghe. *Semidefinite Programming Relaxations of Non-Convex Problems in Control and Combinatorial Optimization*, pages 279–287. Springer US, Boston, MA, 1997. ISBN 978-1-4615-6281-8. doi: 10.1007/978-1-4615-6281-8_15. URL https://doi.org/10.1007/978-1-4615-6281-8_15.

Florentina Bunea, Christophe Giraud, Xi Luo, Martin Royer, and Nicolas Verzelen. Model assisted variable clustering: Minimax-optimal recovery and algorithms, 2018.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, November 2012. ISSN 0246-0203. doi: 10.1214/11-AIHP454. URL https://projecteuclid.org/journals/annales-de-linstitut-henri-poincare-probabilites-et-statistiques/volume-48/issue-4/Challenging-the-empirical-mean-and-empirical-variance--A-deviation/10.1214/11-AIHP454.full. Publisher: Institut Henri Poincaré.

Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37 of *Panoramas et Synthèses [Panoramas and Syntheses]*. Société Mathématique de France, Paris, 2012. ISBN 978-2-85629-370-6.

Geoffrey Chinot, Lecué Guillaume, and Lerasle Matthieu. Statistical learning with lipschitz and convex loss functions. *arXiv preprint arXiv:1810.01090*, 2018.

Stéphane Chrétien, Mihai Cucuringu, Guillaume Lecué, and Lucie Neirac. Learning with semi-definite programming: statistical bounds based on fixed point analysis and excess risk curvature. Technical report, Université Lyon 2, Alan Turing Institute, Oxford University, CREST-ENSAE, 2020.

Stéphane Chrétien, Mihai Cucuringu, Guillaume Lecué, and Lucie Neirac. Learning with semi-definite programming: statistical bounds based on fixed point analysis and excess risk curvature. *J. Mach. Learn. Res.*, 22:Paper No. 230, 64, 2021. ISSN 1532-4435,1533-7928.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, 49(3): 434–448, 2007. ISSN 0036-1445. doi: 10.1137/050645506. URL https://doi.org/10.1137/050645506.

Jules Depersin and Guillaume Lecué. Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probab. Theory Related Fields*, 183 (3-4):997–1025, 2022. ISSN 0178-8051,1432-2064. doi: 10.1007/s00440-022-01127-y. URL https://doi.org/10.1007/s00440-022-01127-y.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 655–664. IEEE Computer Soc., Los Alamitos, CA, 2016.

Yingjie Fei and Yudong Chen. Exponential error rates of SDP for block models: beyond Grothendieck's inequality. *IEEE Trans. Inform. Theory*, 65(1):551–571, 2019. ISSN 0018-9448.

Elisabeth Gaar, Melanie Siebenhofer, and Angelika Wiegele. An SDP-based approach for computing the stability number of a graph. *Mathematical Methods of Operations Research*, 95(1):141–161, feb 2022. doi: 10.1007/s00186-022-00773-1. URL https://doi.org/10.1007%2Fs00186-022-00773-1.

Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. ISBN 978-1-4214-0794-4; 1-4214-0794-9; 978-1-4214-0859-0.

Stefano Gualandi. k-clustering minimum biclique completion via a hybrid cp and sdp approach. In Willem-Jan van Hoeve and John N. Hooker, editors, *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pages 87–101, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

Samuel C. Gutekunst and David P. Williamson. Semidefinite programming relaxations of the traveling salesman problem and their integrality gaps, 2019.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011. ISSN 0036-1445. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

Christoph Helmberg, Franz Rendl, Robert J. Vanderbei, and Henry Wolkowicz. An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, 6(2):342–361, May 1996. ISSN 1052-6234. doi: 10.1137/0806020. Copyright: Copyright 2017 Elsevier B.V., All rights reserved.

Daniel Hong, Hyunwoo Lee, and Alex Wei. Optimal solutions and ranks in the max-cut sdp, 2021.

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009. ISBN 978-0-470-12990-6. doi: 10.1002/9780470434697. URL https://doi.org/10.1002/9780470434697.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986. ISSN 0304-3975. doi: 10.1016/0304-3975(86)90174-X. URL https://doi.org/10.1016/0304-3975(86)90174-X.

Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, 104(486):682–693, 2009a. ISSN 0162-1459. doi: 10.1198/jasa.2009.0121. URL https://doi.org/10.1198/jasa.2009.0121.

Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *arXiv preprint arXiv:0901.4392*, 2009b.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11:517–553, 2010. ISSN 1532-4435.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, Berlin, 2011a.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. 01 2011b. ISBN 978-3-642-22146-0. doi: 10.1007/978-3-642-22147-7.

Rafał Latała et al. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 1997.

G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.*, to appear. ArXiv:1401.2188.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *Ann. Statist.*, 48(2):906–931, 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1828. URL https://doi.org/10.1214/19-AOS1828.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.*, 46(2):611–641, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1562. URL https://doi.org/10.1214/17-AOS1562.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery, 2017.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. A Series of Modern Surveys in Mathematics. Springer-Verlag, Berlin Heidelberg GmbH, 2013. ISBN 978-3-642-20211-7.

Claude Lemaréchal and François Oustry. Semidefinite relaxations and lagrangian duality with application to combinatorial optimization. *INRIA, rapport de recherche*, (3710), 2018.

Zhi-Quan Tom Luo and Wei Yu. An introduction to convex optimization for communications and signal processing. *IEEE Journal on Selected Areas in Communications*, 24:1426–1438, 2006. URL https://api.semanticscholar.org/CorpusID:490413.

Malik Magdon-Ismail. Np-hardness and inapproximability of sparse pca, 2015.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999. ISSN 0090-5364,2168-8966. doi: 10.1214/aos/1017939240. URL https://doi.org/10.1214/aos/1017939240.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4): 1248–1282, 2007. ISSN 1016-443X. doi: 10.1007/s00039-007-0618-7. URL https://doi.org/10.1007/s00039-007-0618-7.

Renato Monteiro. First- and second-order methods for semidefinite programming. *Math. Program.*, 97:209–244, 07 2003. doi: 10.1007/s10107-003-0451-1.

A.S. Nemirovskii and Yu.E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(85)90100-4. URL https://www.sciencedirect.com/science/article/pii/0041555385901004.

A. S. Nemirovsky and D. B. and Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. ISBN 0-471-10345-4. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Yurii Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods & Software*, 9:141–160, 1998. URL https://api.semanticscholar.org/CorpusID:121309892.

Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007. doi: 10.1137/050641983. URL https://doi.org/10.1137/050641983.

F. Rendl. Semidefinite relaxations for partitioning, assignment and ordering problems. *Ann. Oper. Res.*, 240(1):119–140, 2016. ISSN 0254-5330,1572-9338. doi: 10.1007/s10479-015-2015-1. URL https://doi.org/10.1007/s10479-015-2015-1.

V. N. Vapnik and A. Ya. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya.* Izdat. "Nauka", Moscow, 1974.

Vladimir N. Vapnik. *The nature of statistical learning theory.* Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000. ISBN 0-387-98780-0. doi: 10.1007/978-1-4757-3264-1. URL https://doi.org/10.1007/978-1-4757-3264-1.

Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming, 2013.

Tengyao Wang, Quentin Berthet, and Richard J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, 44(5):1896–1930, 2016a. ISSN 0090-5364. doi: 10.1214/15-AOS1369. URL https://doi.org/10.1214/15-AOS1369.

Tengyao Wang, Quentin Berthet, and Richard J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5): 1896–1930, Oct 2016b. ISSN 0090-5364. doi: 10.1214/15-aos1369. URL http://dx.doi.org/10.1214/15-AOS1369.

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 521–528, Cambridge, MA, USA, 2002. MIT Press.

Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. On stein's identity and near-optimal estimation in high-dimensional index models, 2018.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians, 2014.