Contributions to variable selection, clustering and statistical estimation inhigh dimension

Mohamed Ndaoud

université PARIS-SACLAY

ENSAE ParisTech
École nationale de la statistique et de l'administration économique

NNT : 2019SACLG005

Thèse de doctorat

Contributions to variable selection, clustering and statistical estimation in high dimension

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale nº 574- École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques Fondamentales

Thèse présentée et soutenue à Palaiseau, le 03 Juillet 2019, par

MOHAMED NDAOUD

Composition du Jury :

Alexandre Tsybakov ENSAE Paris	Directeur de thèse
Alexandra Carpentier Magdeburg University	Rapporteuse
Yihong Wu Yale University	Rapporteur
Cristina Butucea ENSAE Paris	Examinatrice
Christophe Giraud Université Paris-Sud	Président du jury
Enno Mammen Heidelberg University	Examinateur
Nicolas Verzelen INRA	Examinateur

Beyond the mountains,
more mountains.
Haitian Proverb

À mes parents,
mes deux sœurs
et Raouia

Remerciements

Mes premières pensées vont droit à Sacha, mon directeur de thèse. Je tiens à te remercier pour m'avoir encadré, pour m'avoir fait confiance mais surtout pour m'avoir converti. Jeune étudiant hésitant à pousser les portes du monde de la finance, tu as su me donner une chance, et ainsi un nouveau départ, sans vraiment te poser de questions quant à ma formation. Ce geste en dit beaucoup sur ta personne ainsi que ta noble vocation à former la prochaine génération de chercheurs. Ta patience, ta générosité, ta rigueur ainsi que tous les échanges qu'on a eus me marqueront à jamais. Je suis certain que notre collaboration ne s'arrêtera pas après ma soutenance.

Alexandra et Yihong m'ont fait l'immense honneur d'accepter de rapporter cette thèse. Je leur suis très reconnaissant. Je ne saurais assez les remercier. Mes remerciements vont ensuite au jury de thèse. Cristina, pour avoir été un exemple de professionnalisme pendant nos quelques collaborations. Je suis ravi de faire partie de tes coauteurs. Christophe, pour avoir été l'un des premiers à m'encourager à faire de la recherche. Merci pour tes précieux conseils tout au long de mon parcours ainsi que pour m'avoir recommandé chaque fois que j'en avais besoin. Ton livre m'a été d'une grande aide pendant ces trois ans. Enno, pour ton invitation à Heidelberg et notre collaboration en cours. Je suis très honoré de pouvoir apprendre à tes côtés. Nicolas, pour tes nombreux travaux qui ont inspiré plusieurs résultats de cette thèse.

Une pensée particulière est pour mes autres collaborateurs Laëtitia, Natalia et Olivier. J'ai beaucoup profité de travailler avec vous. Je suis très fier d'avoir pu progresser à vos côtés et des résultats que nous avons obtenus ensemble.

Ce fut un réel plaisir d'effectuer ma thèse au département de statistiques de l'ENSAE. Je n'oublierais jamais tous les moments conviviaux passés ensemble, nos pauses café ainsi que nos matchs de foot. Je suis heureux d'y avoir rencontré des collègues que je considère aujourd'hui mes amis, voire ma seconde famille.

Je remercie Arnak pour m'avoir toujours conseillé. Tu es pour moi un chercheur exemplaire sur les deux plans humain ainsi que scientifique. Guillaume pour tous nos échanges scientifiques, surtout pendant le déjeuner. Marco, pour ta gentillesse et nos séances de musculation. Pierre pour avoir été une personne sur qui j'ai pu compter pendant ces trois années. J'espère que notre amitié va durer longtemps. Victor-Emmanuel pour ta bonne humeur. Nos discussions m'ont souvent égayé. J'espère qu'on pourra travailler ensemble un jour.

Ensuite mes remerciements vont droit à mes cobureaux: Gautier et Geoffrey, à mon colocataire et compagnon de thèse Lionel, ainsi qu'à mes coéquipiers de foot: Badr, Monsour, Wasfe, Aurélien et Arshak. Enfin je remercie: Nicolas, Jérémy, Léna, Solenne, Lucie, Mamadou, Suzanne, Julien, Mehdi, Edwin, Alexander, Philip, Avo, Amir, Christophe, Boris, Alexis, Gabriel et François-Pierre.

Merci à mes parents et mes soeurs pour leur soutien infaillible tout au long de ces trois années. J'espère qu'ils sont fiers de moi. Merci à mes amis proches A., J., I.,R., O. pour tous les moments passés ensemble. Puisse notre amitié durer à jamais.

Contents

Résumé Substantiel

Au cours des dernières décennies, la statistique a été au centre de l'attention, et ce, de bien des façons. Grâce à des améliorations technologiques, par exemple l'augmentation des émissions de la performance des ordinateurs et l'essor du partage de la capacité de données, les statistiques à grande dimension ont été extrêmement dynamiques. En conséquence, ce domaine est devenu l'un des principaux piliers du paysage statistique moderne.

Le cadre statistique standard tient compte du cas où la taille de l'échantillon est relativement grande et que la dimension des observations est beaucoup plus petite. Comme on l'a fait remarquer dans Giraud (2014), l'évolution technologique de l'informatique a poussé à un changement de paradigme de la théorie statistique classique à la statistique de grande dimension. Plus précisément, nous caractérisons un problème statistique comme étant hautement dimensionnel lorsque la dimension des observations est beaucoup plus grande que la taille de l'échantillon. Elle est devenue plus courante avec l'augmentation du nombre de caractéristiques accessibles des données.

D'une manière générale, les problèmes statistiques sont mal posés dans un contexte de grande dimension. D'autres hypothèses sur la structure du modèle sous-jacent sont nécessaires afin de rendre le problème plus important. Par exemple, dans un problème de régression à dimensions élevées, on peut supposer que le vecteur à estimer est parcimonieux (c'est à dire que peu de composantes sont non nulles), ou que la matrice du signal est de faible rang lorsqu'il s'agit d'estimer une matrice. Ces hypothèses sont généralement très réalistes et confirmées par des données empiriques. C'est ce que l'on peut qualifier de statistiques structurées de grande dimension.

Dans cette thèse, nous nous sommes concentrés sur certains problèmes spécifiques aux statistiques de grande dimension et à leurs applications à l'apprentissage automatique.

Notre principale contribution porte sur le problème de la sélection de variables dans la régression linéaire à grande dimension. Nous dérivons des limites non-asymptotiques pour le risque minimax de recouvrement du support sous la perte de Hamming en espérance dans le modèle de bruit Gaussien en $\mathbf{R}^d$ pour les classes de vecteurs s-sparse séparés de 0 par une constante $a > 0$. Dans certains cas, on trouve aussi explicitement les sélecteurs minimax correspondants et leurs variantes adaptatives. Comme corollaires, nous caractérisons précisément une transition de phase asymptotique pour le recouvrement presque complet ainsi que le recouvrement exact.

En ce qui concerne le problème de recouvrement du support exact en acquisition comprimée, nous proposons un algorithme de recouvrement du support exact dans le cadre de l'acquisition comprimée bruitée où toutes les entrées de la matrice de compression sont des Gaussiennes i.i.d. Notre méthode est la première procédure en temps

polynomial à atteindre les mêmes conditions de recouvrement exact que le décodeur de recherche exhaustive étudié dans Rad (2011) et Wainwright (2009a). Notre procédure a l'avantage d'être adaptative à tous les paramètres du problème, robuste et calculable en temps polynomial.

Motivé par l'interaction entre l'estimation et le recouvrement du support, nous introduisons une nouvelle notion de minimaxité pour l'estimation parcimonieuse dans le modèle de régression linéaire à grande dimension. Nous présentons des bornes inférieures plus optimistes que celles données par la théorie classique du minimax et améliorons ainsi les résultats existants. Nous récupérons le résultat précis de Donoho et al. (1992) pour la minimaxité globale à la suite de notre étude. En fixant l'échelle du rapport signal/bruit, nous prouvons que l'erreur d'estimation peut être beaucoup plus petite que l'erreur minimax globale. Entre autres, nous montrons que le recouvrement exact du support n'est pas nécessaire pour atteindre la meilleure erreur d'estimation.

En ce qui concerne le problème de clustering dans le modèle de mélange Gaussien à deux composantes, nous fournissons une caractérisation non-asymptotique précise du risque minimax de Hamming. En conséquence, nous récupérons la transition de phase précise pour un recouvrement exact dans ce modèle. À savoir, la transition de phase se produit autour du seuil $\Delta = \bar{\Delta}_n$ tel que

$$\bar{\Delta}_n^2 = \sigma^2 \left(1 + \sqrt{1 + \frac{2p}{n \log n}}\right) \log n.$$

Notre procédure atteint le seuil précédent. C'est une variante de l'algorithme de Lloyd initialisée par une méthode spectrale. Cette procédure est entièrement adaptative, optimale en termes de taux et simple en termes de calcul. Il s'avère que notre procédure est, à notre connaissance, la première méthode rapide pour obtenir un recouvrement exact optimal.

Une autre contribution principale est consacrée à certains effets de l'adaptabilité sous l'hypothèse de parcimonie, où l'adaptabilité est soit par rapport au niveau du bruit, soit par rapport à sa loi nominale. Nous dérivons les taux minimax optimaux et présentons des estimateurs correspondants pour l'estimation de la variance du bruit σ^2 pour différentes classes de bruit. Par exemple, lorsque la distribution du bruit est exactement connue, σ^2 peut être estimée plus précisément si le bruit a des queues polynomiales connues plutôt que d'appartenir à la classe de bruit sous-Gaussienne. Des résultats similaires ont été obtenus pour le problème de l'estimation minimax de $\|\theta\|_2$. Enfin, nous étudions l'optimalité minimax de l'estimation de θ lorsque le bruit appartient à une classe de distributions avec queues polynomiales ou queues exponentielles. Nous calculons les taux minimax pour ces paramètres. Une conclusion inattendue est que dans le modèle à moyenne parcimonieuse, les taux optimaux sont beaucoup plus lents et dépendent de l'indice polynomial du bruit par opposition aux taux en régression avec des régresseurs "bien répartis".

Dans notre dernière contribution, nous proposons une nouvelle approche pour dériver des développements en série pour certains processus Gaussiens basée sur l'analyse harmonique de leur fonction de covariance. En particulier, une nouvelle série simple est dérivée pour le mouvement Brownien fractionnaire. La convergence de cette dernière série se maintient en moyenne quadratique ainsi qu'uniformément presque sûrement, avec un taux optimal de décroissance du reste de la série. Nous développons également un cadre général de séries convergentes pour certaines classes de processus Gaussiens.

Chapter 1

Introduction

The aim of this chapter is to introduce some of the recent topics of interest in high-dimensional statistics, not necessarily related to the results of the thesis. The list of references is not exhaustive and more details are provided in the following chapters. We inform the reader that the notation may change from chapter to chapter.

1.1 Structured High-Dimensional Models

Over the last decades, Statistics has been at the center of attention, in a wide variety of ways. Thanks to technological improvements, for instance the increase of computer performance and the soar of sharing data capacity, high-dimensional statistics has been extremely dynamic. As a consequence, this field became one of the main pillars of the modern statistical landscape.

The standard statistical framework considers the case where the sample size is relatively large and the dimension of the observations substantially smaller. As pointed out in Giraud (2014), the technological evolution of computing has urged a shift of paradigm from classical statistical theory to high-dimensional statistics. More precisely, we characterize a statistical problem as high-dimensional whenever the dimension of the observations is much larger than the sample size. It has become more common with the increase of accessible features of data.

Generally speaking, statistical problems are ill posed in the high-dimensional setting. Further assumptions on the structure of the underlying model are required in order to make the problem more significant. For instance in a problem of high-dimensional regression, we may assume that the vector to estimate is sparse (i.e. only few components are non-zero), or that the signal matrix is of small rank when dealing with matrix estimation. These assumptions are usually very realistic and endorsed by empirical evidence. This is what can described as as Structured High-Dimensional Statistics.

A new paradigm

One way to summarize some paradigms of modern Statistics is the following. For statistical methods to be "successful", they need to fulfill the OCAR criterion, where OCAR stands respectively for Optimality, Computational tractability, Adaptivity and Robustness.

- Optimality

  In order to evaluate and compare algorithms the oldest criterion is probably the statistical optimality. An estimator is said to be optimal if it cannot be improved in some sense. A widely used criterion is minimax optimality. The notion of minimax optimality is relative to some risk. In order to make this notion more transparent, let us assume that we observe i.i.d realizations $X_1, \dots, X_n$ of some random variable X. Suppose moreover that the distribution of X is given by $\mathbf{P}_\theta$ for some parameter $\theta \in \Theta$ we are interested in. Given a semi-distance d, the performance of an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ of θ is measured by the maximum risk of this estimator on Θ:

  $$r(\hat{\theta}_n) = \sup_{\theta \in \Theta} \mathbf{E}_\theta \left(d^2(\hat{\theta}_n, \theta) \right),$$

  where $\mathbf{E}_\theta$ denotes the expectation with respect to $(X_1, \dots, X_n)$. The minimax risk is given by the smallest worst-case risk reached among all measurable estimators. It is given by:

  $$\mathcal{R}_n^* = \inf_{\hat{\theta}_n} r(\hat{\theta}_n),$$

  where the infimum is over all estimators. In practice, we have a general framework to derive minimax lower bounds, cf. Tsybakov (2008). We say that an estimator θ_n^* is non-asymptotically minimax optimal if the following holds

  $$r(\theta_n^*) \leq C\mathcal{R}_n^*,$$

  where $C > 0$ is a constant.

  Given this criterion, we are interested in estimators achieving the minimax optimal rate. As an example, consider the problem of low rank matrix estimation. It turns out that a simple spectral procedure is minimax optimal. Indeed, Koltchinskii et al. (2011) gives a lower bound and a matching upper bound for the problem of minimax low-rank matrix estimation through a nuclear norm penalization procedure. We recall that the notion of minimax optimality is one way to define optimality, and one may think of other criteria, for instance a Bayesian risk instead of the minimax risk.

  We should point out that the notion of minimax risk is not impeccable. In general, this notion is pessimistic since the worst case scenario may be located in a tiny region of Θ. In that case, the worst-case scenario is not likely to be realized. This fact is detailed further in Chapter 5.

- Computational Tractability

  Computational tractability captures whether a given algorithm can be computed in polynomial time. For instance, a method based on a sample of size N that runs in $\mathcal{O}(N^2)$ is practical while another one running in $\mathcal{O}(e^{\sqrt{N}})$ is not. The recent importance of this criterion is due to the explosion of sample sizes versus the limited capacity of our actual machines. Indeed, for many statistical problems computational by non-tractable exhaustive search methods (i.e. greedy methods testing all possible solutions in a finite set of an exponential size) are shown to be optimal from a statistical point of view.

One of the most challenging problems related to tractability of algorithms is related to computational lower bounds. While a large body of techniques is available to derive general lower bounds for minimax risks, not much in known when we restrict the class of estimators to polynomial time methods. Karp (1972) has proved, for the specific problem of detecting the presence of a hidden clique, that there is a non trivial gap between what could be achieved by any method and by polynomial time methods. This breakthrough shows that it is not always possible to reach statistical optimality through polynomial methods. Inspired by the planted clique problem, the previous fact has been extended to Sparse PCA among many other problems, cf. Berthet and Rigollet (2013). Apart from this reduction to the planted clique problem, it is still unclear how to derive general computational lower bounds having the same flavour as information-theoretical lower bounds.

- Adaptivity

  In order to measure the performance of a given estimator, we may assume that the data is generated according to some model. This model is used further to evaluate the algorithm. Usually, a model depends on different parameters, and the proposed estimator may depend on these parameters. The criterion of adaptivity aims to compare two optimal algorithms through their ability to adapt to the parameters of the model. Sometimes optimality and adaptive optimality are slightly different but in many scenarios adaptivity is possible at almost no cost. For instance consider the problem of high-dimensional estimation in linear regression. The performance of LASSO and SLOPE (Bogdan et al. (2015)) estimators is studied in Bellec et al. (2018) under similar conditions on the design. It turns out that a sparsity dependent tuning of LASSO achieves the minimax estimation rate. While LASSO requires a prior knowledge of the sparsity, SLOPE is adaptively minimax optimal. Still, we may argue that SLOPE requires a higher complexity due to the sorting step. This may be seen as the price to pay for adaptation. To the best of our knowledge, the question of minimax adaptive optimality using a fixed complexity has not been addressed so far. Generally speaking adaptation to sparsity can be done through two main techniques, either by a Lepski type method or by sorted thresholding procedures as in the Benjamini-Hochberg procedure.

- Robustness

  There are two popular notions of robustness. The classical robustness is with respect to outliers, in the sense that a small fraction of data is corrupted by outliers. The Huber contamination model is a typical example of it (Huber (1992)). Let $X_1, \dots, X_n$ be n i.i.d random variables and $\bar{p}$ the probability distribution of X_i. There are two probability measures p, q and a real $\epsilon \in [0, 1/2)$ such that

  $$\bar{p} = (1-\epsilon)p + \epsilon q, \quad \forall i \in \{1, \dots, n\}.$$

  This model corresponds to assuming that $(1-\epsilon)$-fraction of observations, called inliers, are drawn from a reference measure p, whereas ϵ-fraction of observations are outliers and are drawn from another distribution q. In general, all the three parameters p, q and ϵ are unknown. The parameter of interest is the reference distribution p, whereas q and ϵ play the role of nuisance parameters. For instance, the particular case where p is the normal distribution with unknown mean θ and

variance 1 has been extensively studied in the last decade, cf. Diakonikolas et al. (2016, 2017) and references therein.

In dimension one, it is clear that the empirical median is a robust alternative to the empirical mean. The problem becomes more complicated in higher dimensions since there are many generalizations of median in dimension larger than two. For the normal mean estimation problem, Chen et al. (2018) show that robust estimation can be achieved in a minimax sense through Tukey's median (Tukey (1975)). Unfortunately, this approach is not computationally efficient. Recently, many efforts has been made to prove similar results using polynomial time methods, for instance, filtering techniques (Diakonikolas et al. (2016)) and group thresholding (Collier and Dalalyan (2017)). We should add here that the outliers may be deterministic, random or even adversarial.

A more recent notion of robustness is with respect to heavy tailed noise. It is pioneered by Catoni (2012). Although the sub-Gaussian noise assumption is not always realistic, it is quite convenient in order to derive non-asymptotic results thanks to concentration properties. These guarantees fail under heavy tail assumptions of the noise. Assume that we observe $X_1, \dots, X_n \in \mathbf{R}^p$ such that

$$X_i = \mu + \xi_i,$$

where ξ_i are i.i.d centered sub-Gaussian random vectors with independent entries. In that case the empirical mean $\hat{X}$ satisfies, for any given confidence level $\delta > 0$, the following:

$$\mathbf{P}\left(\|\hat{X} - \mu\| \geq C\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log 1/\delta}{n}} \right) \right) \leq \delta,$$

where $C > 0$ and $\|.\|$ denotes the ℓ_2 norm. Recently, the Median-Of-Means estimator (Nemirovskii and Yudin (1983)) was shown to achieve similar results under very mild assumptions on the noise in dimension one, cf. Devroye et al. (2016). Generalization to high dimensions through tractable methods has been an active field of research in recent years. The recent paper by Cherapanamjeri et al. (2019), exhibits a new method based on an SDP relaxation achieving similar results in polynomial time. Their algorithm is significantly faster than the one proposed by Hopkins (2018), which was, to the best of our knowledge, the first polynomial method achieving sub-Gaussian guarantees for mean estimation under only the second moment assumption.

To sum up, we have presented some criteria that we believe are in the core of modern Statistics. Following this perspective, the ideal algorithm would satisfy the OCAR. However this is subject to further evolution. Distributional implementation along with storage capacity are already attracting attention, cf. Szabo and van Zanten (2017) and Ding et al. (2019) for recent advances in these directions. If the data keeps growing without improving the speed limitations then at some point distributed algorithms will become to polynomial time methods what today polynomial time methods are for exponential time methods.

1.2 Interaction with other disciplines

In the previous paragraph, we described what we may consider as the modern statistical paradigm. It is also important to recall that modern statistics have been shaped through many interactions with other disciplines. We only investigate here two of these interactions that are of interest in the rest of this manuscript.

Interaction with statistical physics and mechanics

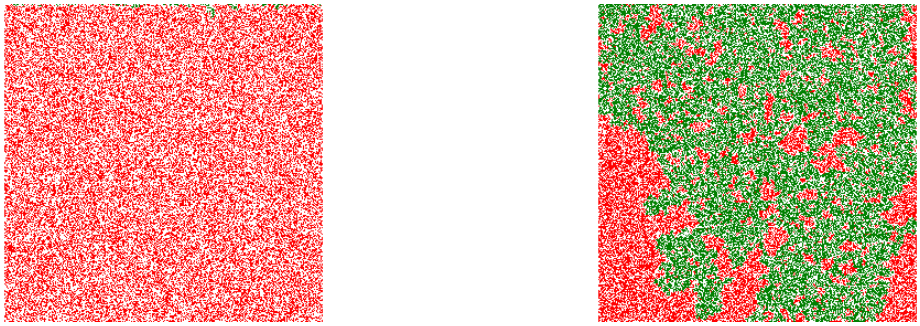
Matter exists in different phases, different states of matter with qualitatively different properties. A phase transition is a singularity in its thermodynamic behavior. As one changes the macroscopic variables of a system, sometimes its properties will abruptly change, often in a dramatic way. We devote this section to discuss some popular phase transitions that has inspired research in Statistics.

Percolation

Among the most popular models involving phase transitions in physics are the Ising model (a model for magnetic solids) and percolation. We only consider here the latter model for simplicity of its phase transition.

The percolation model, is a model meant to study the spread of fluid through a random medium. More precisely, assume that the medium of interest has different channels with different wideness. The fluid will only spread through channels that are wide enough. A question of interest here is whether the fluid can reach the origin starting from outside a large region.

The first mathematical model of percolation was introduced in Broadbent and Hammersley (1957). More precisely, the channels are the edges or bonds between adjacent sites on the integer lattice in the plane $\mathbf{Z}^2$. Each bond is passable with probability p (and hence blocked with probability $q = 1 - p$), and all the bonds are independent of each other. The fundamental question of percolation theory is for which p is there an infinite open cluster?



(a) $p = 0.45$ (b) $p = 0.6$

Figure 1.1: Illustration of percolation on a square lattice of size 500×500. Bonds are red if open, white if blocked and percolation paths are in green.

If we set E_∞ to be the event that there is an infinite open cluster, then it is trivial that $\mathbf{P}(E_\infty)$ is non-decreasing with respect to p. Moreover, by Kolmogorov's 0/1-law

$\mathbf{P}(E_\infty)$ is 0 or 1 for any p. Hence, it is natural to expect the existence of a phase transition around some critical threshold p_c.

Inspired by a line of works, to name Harris (1960), Kesten (1980) proved the long conjectured result that the critical probability is exactly 1/2. In other words, the observed phase transition on the square lattice is the following:

- For $p > 1/2$, there is with probability one a unique infinite open cluster.
- For $p < 1/2$, just the opposite occurs, and percolation is impossible.

Connectivity in the Erdős-Rényi Model

In their seminal paper, Erdős and Rényi (1960) introduced and studied several properties of what we call today the Erdős-Rényi (ER) Model as one of the most popular models in graph theory.

In the ER model $G(n, p)$, n nodes are constructed randomly, then each edge is included in the graph with probability p independent from every other edge. This provides us with a model described by a single parameter. This model has been (and still is) a source of intense research activity, in particular due to its phase transition phenomenon. A very interesting question is related to connectivity of this graph.

Obviously, as p grows the graph is most likely to be connected. One may wonder what is the probability above which the graph is connected almost surely. Erdős and Rényi (1960) show the existence of a sharp phase transition for the described phenomenon as n tends to infinity. For any $\epsilon \in (0, 1)$, they prove the following

- For $p < (1-\epsilon)\frac{\log n}{n}$, then a graph in $G(n, p)$ almost surely contains isolated vertices, and thus is disconnected.
- For $p > (1+\epsilon)\frac{\log n}{n}$, then a graph in $G(n, p)$ is almost surely connected.

Thus $\frac{\log n}{n}$ is a sharp threshold for the connectivity of $G(n, p)$.

There is some similarity between percolation and connectivity in the ER Model. Indeed, we may see the latter problem as a percolation problem on the complete graph instead of the lattice. One may also notice that the critical probability in the ER Model is smaller than the one in percolation and the difference is mainly due to their different geometric structures - more degrees of freedom are allowed in the ER Model. The precise link between the two problems is beyond the scope of this introduction.

Spiked Wigner Model

Random matrices are in the heart of many problems in multivariate statistics such as estimation of covariance matrices and low rank matrix estimation, to name a few. We present here a phase transition phenomenon in the Wigner Model.

Wigner was the first to observe a very interesting property of the spectrum of random matrices. Suppose that W is drawn from the $n \times n$ GOE (Gaussian Orthogonal Ensemble), i.e. W is a random symmetric matrix with off-diagonal entries $\mathcal{N}(0, \frac{1}{n})$, diagonal entries $\mathcal{N}(0, \frac{2}{n})$, and all entries independent (except for symmetry $W_{ij} = W_{ji}$). Set μ_n to be the empirical spectral measure of W such that

$$\mu_n(A) = \frac{1}{n} \left|\{\text{eigenvalues of } W \text{ in } A\}\right|, \quad \forall A \subset \mathbf{R}.$$

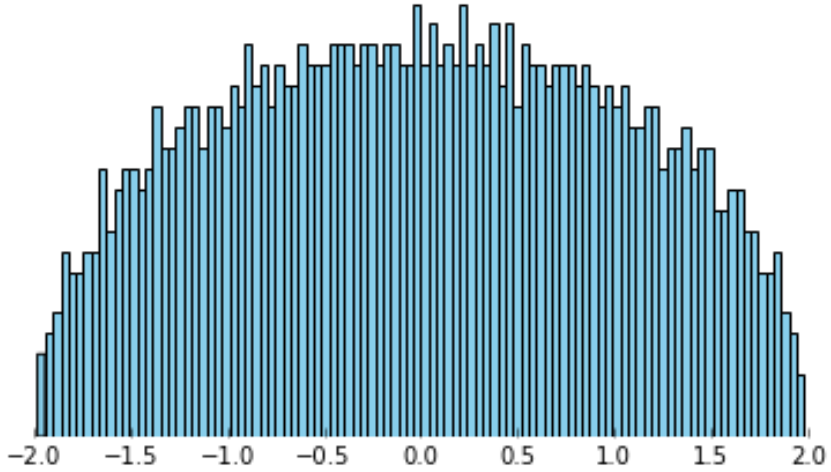


Figure 1.2: The empirical spectral distribution of a matrix drawn from the 1500×1500 GOE.

The limit of the empirical spectral measure for Wigner matrices, as n tends to infinity, is the Wigner semicircle distribution, as described by Wigner (1958). The semicircle distribution is supported on $[-2, 2]$ and is given by the density

$$\mu(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \quad \forall x \in [-2, 2].$$

This phenomenon is more general and universal in the sense that it holds not necessarily for matrices with Gaussian entries. From a physical point of view, we may view the eigenvalues of W as an interacting system where it is possible to characterize these interactions precisely. The previous result, in particular, states that the eigenvalues are confined in the compact set $[-2, 2]$ with high probability (this holds even almost surely).

A very interesting question is about the behaviour of the spectrum of W under an external action. More precisely, for $\lambda > 0$, and a spike vector $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$, we define the spiked Wigner model as follows:

$$Y = \lambda \mathbf{x}\mathbf{x}^\top + W.$$

As above, it is not difficult to observe that the top eigenvalue of Y is 2 if $\lambda = 0$ and that it tends to infinity as λ goes to infinity. Hence, we may wonder at which power λ of the spike the spectrum gets affected. This question was solved by Féral and Péché (2007) where a new phase transition phenomenon arises as n tends to infinity.

- For $\lambda \leq 1$, the top eigenvalue of Y converges almost surely to 2.
- For $\lambda > 1$, the top eigenvalue converges almost surely to $\lambda + 1/\lambda > 2$.

It was further shown in Benaych-Georges and Nadakuditi (2011) that the top (unit-norm) eigenvector $\hat{\mathbf{x}}$ of Y has non-trivial correlation with the spike vector almost surely, if and only if $\lambda > 1$. This phase transition is probably at the heart of a better understanding of spectral methods. Interestingly, in the regime where $\lambda < 1$, we may not

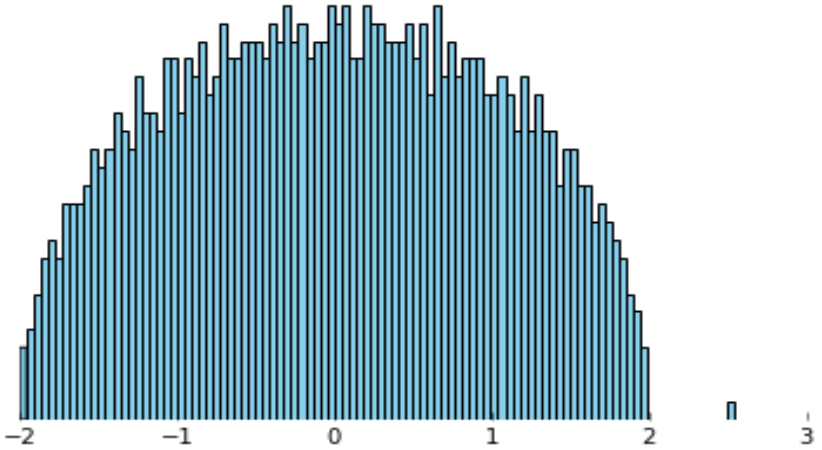


Figure 1.3: The empirical spectral distribution of a spiked Wigner matrix with $n = 1500$ and $\lambda = 2$.

distinguish the presence of a spike from its absence. In Perry et al. (2018), this question is studied in detail and it turns out that, under the Gaussian random spike, the spectral test based on the top eigenvalue of Y is, in a certain sense, the best test that one should use in order to test the presence of a spike.

This result is exciting but also alarming. Indeed, the spiked model is a simple model to study the performance of Principal Component Analysis (PCA). Basically, we assume the presence of a low rank signal corrupted by some noise, and want to recover the underlying signal. The previous results show that, in the regime where $\lambda < 1$, there is no hope to capture non-trivial correlation with the signal and hence the resulting PCA is meaningless. Such facts are not always known to practitioners and we believe that a good use and interpretation of PCA should always start with a safety test asserting whether PCA is meaningful or not.

Interaction with optimization

Statistics and Learning Theory are intimately linked with optimization when it comes to algorithms. Indeed, in Learning Theory, a classical problem is to minimize some empirical risk. In order to evaluate the performance of a predictor or a classifier we rely on a specific choice of some loss function and evaluate the performance with respect to it. In that sense the notion of goodness of training is relative to the choice of loss. As long as we can optimize the objective loss function, we may be able to derive a good predictor that is usually optimal in some sense.

For statisticians, the main interest is statistical accuracy. It is still important to wonder whether we can in practice find an optimizer. For this purpose, we rely on the wealth of results developed by the optimization community that is mostly designed for convex objective functions (where we have existence of global minima) defined on convex sets. Hence, as long as the goal is to minimize some convex function on a convex set, we may assume given the corresponding minimizer.

Unfortunately, in many important problems convexity is not granted. A popular approach is to convexify the objective function and to optimize its convex counterpart. This is one important example of the influence of optimization on statistics. By doing

so, one still has to prove that a good solution of the convex problem leads to a good predictor for the original problem. In classification, SVM and boosting are examples supporting that the convexification trick can be optimal. The previous remark does not hold in general. As an example, Sparse PCA is a non-convex problem where its SDP relaxation (cf. d'Aspremont et al. (2005)) leads to a strict deterioration of the rate of estimation compared to a solution of the original problem. In simple words, convexification is not always optimal.

Most interestingly, the interplay between statistics and optimization becomes more beneficial when combining both realms. A popular approach handling non-convex problems is given by variants of greedy descents over the objective function. This is usually how iterative algorithms are designed. Some popular examples are Lloyds for K-means clustering (Lloyd (1982)) and Iterative Hard Thresholding for sparse linear regression (Shen and Li (2017); Foucart and Lecué (2017); Liu and Barber (2018)).

These algorithms are non-convex counterparts of gradient descent methods. One may argue that studying the statistical performance of iterative algorithms can be richer than analyzing minimizers of objective functions. We discuss below this fact.

- Advantages of iterative algorithms:

  It is common in the statistical community to evaluate the performance of estimators given as solutions of optimization problems. In practice, we do not have access to one of the optimizers but only to an approximation. The corresponding approximation error is missing in the statistical analysis and is referred to as the optimization error. By studying iterative algorithms, we get a precise control of both errors after a certain number of steps.

  The optimization error can be made as small as possible by running a large number of iterations. Usually the number of iterations is set at some precision level. While the optimization error is vanishing and depends on the speed of the algorithm, the statistical error is intrinsic. Hence, the latter is in all cases a lower bound of the global error. That being said, there is no need to run an infinite number of iterations and we can always stop the algorithm once the statistical accuracy is reached. Combining statistics and optimization, we may consider early stopping rules saving time and efforts from the optimization side.

  Following the previous fact, there is no need of global convergence of the algorithm. Existence of global minima is not even required. Indeed, as long as local minima of the objective function are below the optimal statistical error, then the algorithm succeeds.

  To sum up, on one hand, iterative algorithms may tolerate non-convexity for some statistical problems, while on the other hand statistical limits and assumptions allow for more flexibility and improvements of the algorithms.

- Defects of iterative algorithms:

  Although the benefits of combining optimization and statistics are numerous, there are still some drawbacks in using iterative algorithms. The first one is probably related to ignoring which deterministic descent algorithm to mimic in order to construct an iterative procedure. Objective function minimization is a good guideline for that purpose, as long as we only care about the statistical performance. To

the best of our knowledge, the choice of iterative methods is empirical and there is no general rule to master it.

Another drawback and probably the most limiting part in the analysis of iterative algorithms is the dependence between the steps. Indeed, while the noise is usually assumed to have independent entries, after one iteration the estimator usually starts depending on the noise in a complex way. It is not always trivial to handle these dependencies but in some important examples we can use the contraction of the objective function in order to get around it, cf. for example, Lu and Zhou (2016).

We conclude this section by mentioning a potential direction to explore as a consequence of the interplay between the fields of Optimization and Statistics. It is well known that in both fields, the notion of adaptivity is of high importance. An adaptive algorithm solving an optimization problem is probably not adaptive from a statistical perspective and vice versa. Simultaneous analysis, through iterative algorithms for instance, may lead to a more accurate notion of adaptivity.

1.3 Variable Selection and Clustering

Variable selection is an important task in Statistics that has many applications. Consider the linear regression model in high dimensions, and assume that the underlying vector $\boldsymbol{\theta}$ is sparse. There are three questions of interest that could be treated independently.

- Detection: Test the presence of a sparse signal against the hypothesis that the observation is pure noise.
- Estimation: Estimate the sparse signal vector.
- Support recovery: Recover the set of non-zero variables (support of the true sparse signal).

In what follows, the question of variable selection is equivalent to support recovery. While detection and support recovery require additional separation conditions, in order to get meaningful results, it is not the case for estimation. The three problems cannot always be compared in terms of hardness. It is obvious that detection is easier than support recovery in the sense that it requires weaker separation conditions, but there are no general rules to compare these tasks.

At this stage, we want to emphasize that these three tasks may be combined for a specific purpose but they can also be used independently following the statistical application. Let us consider an illustrative example of these tasks. An internet service provider routinely collects statistics of server to determine if there are abnormal blackouts. While this monitoring is performed on a large number of servers, only very few servers are believed to be experiencing problems if there are any. This represents the sparsity assumption. The detection problem is then equivalent to determining if there are any anomalies among all servers; the estimation problem is equivalent to associating weights (probability of failure) to every server, while support recovery is equivalent to identifying the servers with anomalies.

Probably one of the most exciting applications of high-dimensional variable selection is in genetic epidemiology. Typically, the number of subjects n, is in thousands, while p ranges from tens of thousands to hundreds of thousands of genetic features. The number of genes exhibiting a detectable association with a trait is extremely small in practice. For example, in disease classification, it is commonly believed that only tens of genes are responsible for a disease. Selecting tens of genes helps not only statisticians in constructing a more reliable classification rule, but also biologists to understand molecular mechanisms.

Apart from its own interest, variable selection can be used in estimation procedures. For instance it can be used as a first step in sparse estimation reducing the dimension of the problem. The simple fact that estimating a vector on its true support has a significantly smaller error than on the complete vector, has encouraged practitioners to proceed to variable selection as a first step. Wasserman and Roeder (2009) is an example of works studying the theoretical aspects of methods in the same spirit. Long story short, the main idea of this approach is to first find a good sparse estimator and then keep its support. This step is also known as model selection. Once the support is estimated, the dimension of the problem is much smaller than the initial dimension p. The second step is then to estimate the signal solving a simple least-square problem on the estimated support.

Examples of Variable Selection methods

The issues of variable selection for high-dimensional statistical modeling has been widely studied in last decades. We give here a brief summary of some well-known procedures. All these procedures are thresholding methods. Logically, selecting a variable is only possible if we can distinguish it from noise. As long as we can estimate the noise level, the variables that are above this level are more likely to be relevant. This also explains why the threshold usually depends on the noise and not necessarily on the significant variables.

In order to present some of these thresholding methods, assume that we observe $\mathbf{x} \in \mathbf{R}^p$, such that

$$\mathbf{x} = \boldsymbol{\theta} + \xi,$$

where $\boldsymbol{\theta}$ is a sparse vector, and ξ is a vector of identically distributed noise variables, not necessarily independent. A thresholding procedure typically returns an estimated set $\hat{S}$ of the form

$$\hat{S} = \{j : \quad |x_j| > t(\mathbf{x})\},$$

where x_j is the j-th component of $\mathbf{x}$. Note that the threshold $t(.)$ may depend on the vector $\mathbf{x}$. Here are examples of thresholding procedures, where the two first are deterministic while the others are data-dependent.

- Bonferroni's procedure: Bonferroni's procedure with family-wise error rate (FWER) at most $\alpha \in (0, 1)$ is the thresholding procedure that uses the threshold

$$t(\mathbf{x}) = F^{-1}(1 - \alpha/2p).$$

  We use the abusive notation F^{-1} to denote the generalized inverse of the c.d.f of the component of ξ.

- Sidák's procedure (Šidák (1967)): This procedure is more aggressive (i.e. gives higher threshold) than Bonferroni's procedure. It uses the threshold

$$t(\mathbf{x}) = F^{-1}((1-\alpha/2)^{1/p}).$$

Consider the non-increasing arrangement of coordinates of $\mathbf{x}$ in absolute value such that $|x|_{(1)} \geq \cdots \geq |x|_{(p)}$. The next procedures are data-dependent and are shown to be strictly more powerful than Bonferroni's procedure.

- Holm's procedure (Holm (1979)): Let k be the largest index such that

$$|x|_{(i)} \geq F^{-1}(1-\alpha/2(p-i+1)), \quad \forall i \leq k.$$

Holm's procedure with FWER controlled at α is the thresholding procedure that uses

$$t(\mathbf{x}) = |x|_{(k)}. \tag{1.1}$$

- Hochberg's procedure (Hochberg (1988)): More aggressive than Holm's, the corresponding threshold is given by (1.1) where k is the largest index i such that

$$|x|_{(i)} \geq F^{-1}(1-\alpha/2(p-i+1)).$$

Theoretical properties of these methods, in a more general setting, are analyzed in the recent work of Gao and Stoev (2018).

Community Detection

Learning community structures is a central problem in machine learning and computer science. The simplest clustering setting is the one where we observe n agents (or nodes) that are partitioned into two classes. Depending on available data, we may proceed differently. When observed data is interactions among agents (e.g., social, biological, computer or image networks), then clustering is node-based. While, when observed data is spacial position of agents, then clustering is vector-based. In both cases, the goal is to infer, from the provided observations, communities that are alike or complementary. A very popular model for the first case is the Stochastic Block Model (Holland et al. (1983)), while the two component Gaussian mixture model is the equivalent for the second case. The problem of community recovery can be reformulated as recovering the set of labels belonging to the same class and hence can be seen as a variable selection problem.

As the study of community detection grows at the intersections of various fields, the notions of clusters and the models vary significantly. As a result, the comparison and validation of clustering algorithms remains a major challenge.

Stochastic Block Model (SBM)

The SBM has been at the center of attention in a large body of literature. It can be seen as an extension of the ER model, described previously. Recall that in the ER model, edges are placed independently with probability p, providing a model described by a single parameter. It is however well known to be too simplistic to model real networks,

in particular due to its strong homogeneity and absence of community structure. The SBM is based on the assumption that agents in a network connect not independently but based on their profiles, or equivalently, on their community assignment. Of particular interest is the SBM with two communities and symmetric parameters, also known as the planted bisection model, denoted here by $\mathcal{G}(n, p, q)$, with an integer n denoting the number of vertices.

More precisely, each node v in the graph is assigned a label $\sigma_v \in \{-1, 1\}$, and each pair of nodes (u, v) is connected with probability p within the clusters of labels and q across the clusters. Upon observing the graph (without labels), the goal of community detection is to reconstruct the label assignments. Of course, one can only hope to recover the communities up to a global flip of labels. When $p = q$, it is clearly impossible to recover the communities, whereas for $p > q$ or $p < q$, one may hope to succeed in certain regimes. While this is a toy model, it captures some of the central challenges for community detection. In particular, it represents a phase transition similar to the ER model. In the independent works of Abbe et al. (2014) and Mossel et al. (2015), the phase transition of exact recovery is characterized precisely. For any $\epsilon > 0$, we observe the following as n tends to infinity,

- For $\sqrt{p} - \sqrt{q} < (1 - \epsilon)\sqrt{\frac{2 \log n}{n}}$, then exact recovery of labels in $\mathcal{G}(n, p, q)$ is impossible.
- For $\sqrt{p} - \sqrt{q} > (1 + \epsilon)\sqrt{\frac{2 \log n}{n}}$, then exact recovery of labels in $\mathcal{G}(n, p, q)$ is possible and is achieved through a polynomial time method.

One method achieving the sharp phase transition is based on a spectral method followed by a rounding procedure. In order to get more intuition on the construction of such a method, let us observe that the graph adjacency matrix A can be decomposed as follows

$$A = \frac{p + q}{2} \mathbf{1}\mathbf{1}^\top + \frac{p - q}{2} \eta\eta^\top + W,$$

where $\mathbf{1}$ is the vector of ones in $\mathbf{R}^n$, $\eta \in \{-1, 1\}^n$ is the vector of labels up to a sign change and W is a centered sub-Gaussian random matrix with independent entries. The first term, also known as the mean component can be removed if $p + q$ is known or simply by projecting the adjacency matrix on the orthogonal of $\mathbf{1}$. As a consequence, the observation A has a similar behaviour as the Spiked Wigner model. It can be shown that spectral methods are efficient to detect the presence of the planted bisection structure and also to get non-trivial correlation with the vector of labels. The rounding step can be seen as a cleaning step that helps finding the exact labels once a non-trivial correlation is captured.

Gaussian Mixture Model (GMM)

The Gaussian Mixture Model is one of the most popular statistical models. Of particular interest is the two component Gaussian Mixture Model with two balanced communities. Similary to SBM, assume that $\eta \in \{-1, 1\}^n$ is a vector of labels. In a GMM with labels η, the random vectors Y_i are assumed to be independent and to follow a Gaussian distribution with variance 1 centered at θ_1 if $\eta_i = 1$ or centered at θ_2 if $\eta_i = -1$, where $\theta_1, \theta_2 \in \mathbf{R}^p$ are the unknown center vectors. In this setting variables belonging

to the same group are close to each other and we may rely on distances between the observations to cluster them.

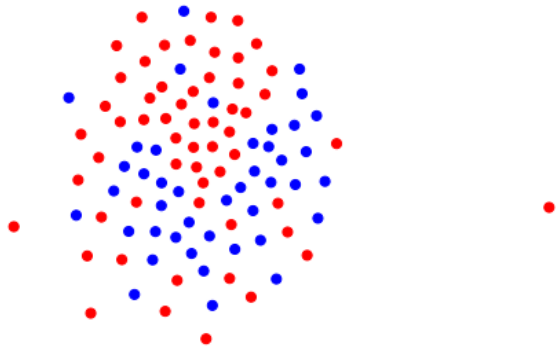


Figure 1.4: A two-dimensional projection of a two component Gaussian mixture with $n = 100$ and $p = 1000$.

As for SBM, spectral methods used for GMM are inspired by the following fact. Notice that the observation matrix $Y \in \mathbf{R}^{p\times n}$ can be decomposed as follows

$$Y = \frac{\theta_1 + \theta_2}{2}\mathbf{1}^\top + \frac{\theta_1 - \theta_2}{2}\eta^\top + W,$$

where W is a centered Gaussian random matrix with independent columns. The mean component $\frac{\theta_1+\theta_2}{2}\mathbf{1}^\top$ can be handled in the same way as for SBM. As a consequence, clustering is possible when $\|\theta_1 - \theta_2\|$ is large enough. The observation Y can be viewed as an asymmetric equivalent of the Spiked Wigner Model. We refer the reader to the recent work of Banks et al. (2018) for phase transitions in this model.

Apart from community detection, the problem of center estimation in GMM is of high interest as well. Although many iterative procedures, in the same spirit as K-means, operate a center estimation step, it is not always true that the center estimation is necessary to achieve exact recovery. As we will argue in Chapter 6 the two problems may be considered independently.

Chapter 2

Overview of the Results

This thesis deals with the following statistical problems: Variable selection in high-Dimensional Linear Regression, Clustering in the Gaussian Mixture Model, Some effects of adaptivity under sparsity and Simulation of Gaussian processes. The goal of this chapter is to provide a motivation for these statistical problems, to explain how these areas are connected and to give an overview of the results derived in the next chapters. Each chapter can be read independently of the others.

2.1 Variable Selection in High-Dimensional Linear Regression

In recent years, the problem of variable selection in high-dimensional regression models has been extensively studied from the theoretical and computational viewpoints. In making effective high-dimensional inference, sparsity plays a key role. With regard to variable selection in sparse high-dimensional linear regression, the Lasso, Dantzig selector, other penalized techniques as well as marginal regression were analyzed in detail. In some cases, practitioners are more interested in the pattern or support of the signal rather than its estimation. It turns out, that the problem of variable selection or support recovery are highly dependent on the separation between entries of the signal on its support and zero. One may define optimality in a minimax sense for this problem. In particular, the study of phase transitions in support recovery with respect to the separation parameter is of high interest. The model of interest is the one of high-dimensional linear regression

$$Y = X\boldsymbol{\theta} + \sigma\xi.$$

When X is an orthogonal matrix the model corresponds to the sparse vector model, while when X is random it corresponds to noisy Compressed Sensing. Some papers on this topic include Meinshausen and Bühlmann (2006); Candes and Tao (2007); Wainwright (2009b); Zhao and Yu (2006); Zou (2006); Fan and Lv (2008); Gao and Stoev (2018).

Chapter 3 is devoted to derive non-asymptotic bounds for the minimax risk of support recovery under expected Hamming loss in the Gaussian mean model in $\mathbf{R}^d$ for classes of s-sparse vectors separated from 0 by a constant $a > 0$. Namely, we study the problem of variable selection in the following model:

$$Y_j = \theta_j + \sigma\xi_j, \qquad j = 1, \dots, d,$$

where $\xi_1, \dots, \xi_d$ are i.i.d. standard Gaussian random variables, $\sigma > 0$ is the noise level, and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ is an unknown vector of parameters to be estimated. For $s \in \{1, \dots, d\}$ and $a > 0$, we assume that $\boldsymbol{\theta}$ is (s, a)-*sparse*, which is understood in the sense that $\boldsymbol{\theta}$ belongs to the following set:

$$\Theta_d(s, a) = \big\{\boldsymbol{\theta} \in \mathbf{R}^d : \text{ there exists a set } S \subseteq \{1, \dots, d\} \text{ with at most } s \text{ elements such that } |\theta_j| \geq a \text{ for all } j \in S, \text{ and } \theta_j = 0 \text{ for all } j \notin S\big\}.$$

We study the problem of selecting the relevant components of $\boldsymbol{\theta}$, that is, of estimating the vector

$$\eta = \eta(\boldsymbol{\theta}) = \big(I(\theta_j \neq 0)\big)_{j=1,\dots,d},$$

where $I(\cdot)$ is the indicator function. As estimators of η, we consider any measurable functions $\hat{\eta} = \hat{\eta}(Y_1, \dots, Y_d)$ of $(Y_1, \dots, Y_d)$ taking values in $\{0, 1\}^d$. Such estimators will be called *selectors*. We characterize the loss of a selector $\hat{\eta}$ as an estimator of η by the Hamming distance between $\hat{\eta}$ and η, that is, by the number of positions at which $\hat{\eta}$ and η differ:

$$|\hat{\eta} - \eta| \triangleq \sum_{j=1}^{d} |\hat{\eta}_j - \eta_j| = \sum_{j=1}^{d} I(\hat{\eta}_j \neq \eta_j).$$

Here, $\hat{\eta}_j$ and $\eta_j = \eta_j(\theta)$ are the jth components of $\hat{\eta}$ and $\eta = \eta(\theta)$, respectively. The expected Hamming loss of a selector $\hat{\eta}$ is defined as $\mathbf{E}_\theta|\hat{\eta} - \eta|$, where $\mathbf{E}_\theta$ denotes the expectation with respect to the distribution $\mathbf{P}_\theta$ of $(Y_1, \dots, Y_d)$. We are interested in the minimax risk

$$\inf_{\tilde{\eta}} \sup_{\boldsymbol{\theta} \in \Theta_d(s,a)} \mathbf{E}_{\boldsymbol{\theta}} |\tilde{\eta} - \eta|.$$

In some cases, we get exact expressions for the non-asymptotic minimax risk as a function of d, s, a and find explicitly the corresponding minimax selectors. These results are extended to dependent or non-Gaussian observations and to the problem of crowdsourcing. Analogous conclusions are obtained for the probability of wrong recovery of the sparsity pattern. As corollaries, we characterize precisely an asymptotic sharp phase transition for both almost full and exact recovery. We say that almost full recovery is possible if there exists a selector $\hat{\eta}$ such that

$$\lim_{d \to \infty} \sup_{\boldsymbol{\theta} \in \Theta_d(s,a)} \frac{1}{s} \mathbf{E}_{\boldsymbol{\theta}} |\tilde{\eta} - \eta| = 0.$$

Moreover, we say that exact recovery is possible if there exists a selector $\hat{\eta}$ such that

$$\lim_{d \to \infty} \sup_{\boldsymbol{\theta} \in \Theta_d(s,a)} \mathbf{E}_{\boldsymbol{\theta}} |\tilde{\eta} - \eta| = 0.$$

Among other results, we prove that necessary and sufficient conditions for almost full and exact recovery are respectively

$$a > \sigma \sqrt{2 \log(d/s - 1)}$$

and

$$a > \sigma \sqrt{2 \log(d - s)} + \sigma \sqrt{2 \log(s)}. \tag{2.1}$$

Moreover, we propose data-driven selectors that provide almost full and exact recovery adaptively to the parameters of the classes.

As a generalization to high-dimensional linear regression, we study in Chapter 4 the problem of exact support recovery in noisy Compressed Sensing. Assume that we have the vector of measurements $Y \in \mathbf{R}^n$ satisfying

$$Y = X\boldsymbol{\theta} + \sigma\xi$$

where $X \in \mathbf{R}^{n\times p}$ is a given design or sensing matrix, $\boldsymbol{\theta} \in \mathbf{R}^p$ is the unknown signal, and $\sigma > 0$. Similarily to the Gaussian sequence model case, we assume that $\boldsymbol{\theta}$ belongs to $\Theta_p(s, a)$.We are interested in the Hamming minimax risk, and therefore in sufficient and necessary conditions for exact recovery in Compressed Sensing. Table 2.1 summarizes

SNR	Upper bound for ML	Lower bound
$a/\sigma = \mathcal{O}(1/\sqrt{s})$	$\frac{\sigma^2 \log(p-s)}{a^2}$	
$a/\sigma = \mathcal{O}(1)$ and $a/\sigma = \Omega(1/\sqrt{s})$	$\frac{s\log(\frac{p}{s})}{\log(1+s\frac{a^2}{\sigma^2})} \vee \frac{\log(p-s)}{\log(1+\frac{a^2}{\sigma^2})}$	
$a/\sigma = \Omega(1)$	$s\log(\frac{p}{s})$	$\frac{s\log(p/s)}{\log(1+sa^2/\sigma^2)}$

Table 2.1: Phase transitions in Gaussian setting.

known sufficient and necessary conditions for exact recovery in the setting where both X and ξ are Gaussian.

We propose an algorithm for exact support recovery in the setting of noisy compressed sensing where all entries of the design matrix are i.i.d standard Gaussian. This algorithm is the first polynomial time procedure to achieve the same conditions of exact recovery as the exhaustive search (maximal likelihood) decoder that was studied in Rad (2011), Wainwright (2009a). In particular, we prove that, in the zone $a/\sigma = \mathcal{O}(1)$, our sufficient condition for exact recovery has the form

$$n = \Omega\left(s\log\left(\frac{p}{s}\right) \vee \frac{\sigma^2\log(p-s)}{a^2}\right),$$

where we write $x_n = \Omega(y_n)$ if there exists an absolute constant $C > 0$ such that $x_n \geq Cy_n$. Our procedure has an advantage over the exhaustive search of being adaptive to all parameters of the problem and computable in polynomial time. The core of our analysis consists in the study of the non-asymptotic minimax Hamming risk of variable selection. This allows us to derive a procedure, which is nearly optimal in a non-asymptotic minimax sense. We develop its adaptive version, and propose a robust variant of our method to handle datasets with outliers and heavy-tailed distributions of observations. The resulting polynomial time procedure is near optimal, adaptive to all parameters of the problem and also robust.

Another topic of interest is the interplay between estimation and support recovery. As described previously, and depending on applications, practitioners may be interested either in accurate estimation of the signal or recovering its support. It is not clear how the two problems are connected. When the support of the signal is known, one

may achieve better rates of estimation but it is not clear whether the knowledge of the support is necessary for that.

A fairly neglected problem by practitioners is the bias in high-dimensional estimation. Despite their popularity, the l_1-regularization methods suffer from some drawbacks. For instance, it is well known that penalized estimators suffer from an unavoidable bias as pointed out in Zhang and Huang (2008), Bellec (2018). A sub-optimal remedy is to threshold the resulting coefficients as suggested in Zhang (2009). However, this approach requires additional tuning parameters, making the resulting procedures more complex and less robust. It turns out that under some separation of the components this bias can be removed. Zhang (2010) propose a concave penalty based estimator in order to deal with the bias term. A variant of this method that is adaptive to sparsity is given in Feng and Zhang (2017).

Alternatively, other techniques were introduced through greedy algorithms such as Orthogonal Matching Pursuit: Cai and Wang (2011); Joseph (2013); Tropp and Gilbert (2007); Zhang (2011b). For Forward greedy algorithm, also referred to as matching pursuit in the signal processing community, it was shown that the irrepresentable condition of Zhao and Yu (2006) for l_1-regularization is necessary to effectively select features. For Backward greedy algorithm, although widely used by practitioners, not much was known concerning its theoretical analysis in the literature before Zhang (2011a). A combination of these two algorithms is presented in Zhang (2011a) and turns out to be successful removing the bias when it is possible. Again, we emphasize that a precise characterization of necessary and sufficient conditions for the bias to be removed is missing and will certainly complement the state of the art results.

Chapter 5 is devoted to the interplay between estimation and support recovery. We introduce a new notion of scaled minimaxity for sparse estimation in high-dimensional linear regression model. The scaled minimax risk is given by

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}\in\Theta_d(s,a)} \mathbf{E}_{\boldsymbol{\theta}} \left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right),$$

where the infimum is taken over all possible estimators $\hat{\boldsymbol{\theta}}$ and $\|.\|$ is the ℓ_2-norm,. We present more optimistic lower bounds than the one given by the classical minimax theory and hence improve on existing results. We recover the sharp result of Donoho et al. (1992) for the global minimaxity in the Gaussian sequence model as a consequence of our study, namely

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{|\boldsymbol{\theta}|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}} \left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right) = 2\sigma^2 s \log(d/s)(1 + o(1)) \quad \text{as } \frac{s}{d} \to 0. \tag{2.2}$$

Here $|.|_0$ is the number of non-zero components and $\mathbf{E}_{\boldsymbol{\theta}}$ denotes the expectation with respect to the distribution of Y in the Gaussian sequence model. Fixing the scale of the signal-to-noise ratio, we prove that estimation error can be much smaller than the global minimax error. We also study sufficient and necessary conditions for an estimator $\hat{\boldsymbol{\theta}}$ to achieve *exact estimation* that corresponds in the Gaussian sequence model to the following property:

$$\lim_{s/d\to 0} \frac{\sup_{\boldsymbol{\theta}\in\Theta_d(s,a)} \mathbf{E}_{\boldsymbol{\theta}} \left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right)}{\sigma^2 s} = 1.$$

The notion of exact estimation is closely related to the bias of estimation, since achieving exact estimation is only possible if the bias is avoidable. Finally, we construct a new optimal estimator for the scaled minimax sparse estimation and derive its adaptive variant.

Among other findings, we show that exact support recovery is not necessary to achieve the scaled minimax error. Indeed, in Chapter 5, we obtain a new phase transition related to sparsity. Recall that we show in Chapter 3 that the necessary and sufficient condition to achieve exact recovery is given by

$$a > \sigma\sqrt{2\log(d-s)} + \sigma\sqrt{2\log(s)},$$

cf. (2.1). To achieve exact estimation, we prove that a necessary condition is given by

$$a > \sigma\sqrt{2\log(d/s-1) + 2\log\log(d/s-1)} + \sigma\sqrt{2\log\log(d/s-1)}.$$

Hence exact recovery is not necessary for exact estimation. In fact, when $s \gg \log(d)$ then exact estimation is easier and when $s \ll \log(d)$ exact recovery becomes easier. This shows that there is no direct implication of exact recovery on exact estimation, and the latter task should be considered as a separate problem.

2.2 Clustering in Gaussian Mixture Model

Gaussian Mixture Model (GMM) is one of the most popular vector clustering models. The simple case of two components mixture is modeled as follows:

$$Y_i = \eta_i \boldsymbol{\theta} + \sigma\xi_i, \quad \forall i = 1, \dots, n, \tag{2.3}$$

where $\boldsymbol{\theta} \in \mathbf{R}^p$ is the center vector, $\eta = (\eta_1, \dots, \eta_n) \in \{-1, 1\}^n$ is the labels vector and $(\xi_i)_{1\le i\le n}$ is a sequence of standard independent random Gaussian vectors. Two important questions arising in this model are related to center estimation and community detection. Performance guarantees for several algorithmic alternatives have emerged, including expectation maximization (Dasgupta and Schulman (2007)), spectral methods (Vempala and Wang (2004); Kumar and Kannan (2010); Awasthi and Sheffet (2012)), projections (random and deterministic in Moitra and Valiant (2010); Sanjeev and Kannan (2001)) and the method of moments (Moitra and Valiant (2010)).

While Moitra and Valiant (2010) and Mixon et al. (2016) are interested in center estimation, Vempala and Wang (2004); Kumar and Kannan (2010); Awasthi and Sheffet (2012) are interested in recovering correctly the clusters. We only focus on the latter question in this manuscript.

The paper of Lu and Zhou (2016) is one of the first works that study theoretical guarantees of Lloyd's algorithm in order to recover communities in the sub-Gaussian Mixture Model. It, particularly, succeeds to handle the dependence between different steps of the latter algorithm. Recently, Fei and Chen (2018) and Giraud and Verzelen (2018) have investigated the clustering performance of SDP relaxed K-means in the setting of sub-Gaussian Mixture Model. After identifying the appropriate signal-to-noise ratio (SNR), that is different from the one given by Lu and Zhou (2016) when $p > n$, Giraud and Verzelen (2018) prove that the misclassification error decays exponentially

fast with respect to this SNR. These recovery bounds for SDP relaxed K-means improve upon the results previously known in the GMM setting.

In high dimensional regime, the exact recovery phase transition is not known. Also, in the same regime, there is a strict performance gap between known results for fast iterative algorithms and SDP relaxation methods. It is of interest to know whether this gap is crucial or not. Eventual positive answers will complement the state of the art results.

In Chapter 6, we consider the problem of exact recovery of clusters in the two components Gaussian Mixture Model (2.3). We denote by $\mathbf{P}_{(\boldsymbol{\theta},\eta)}$ the distribution of Y and by $\mathbf{E}_{(\boldsymbol{\theta},\eta)}$ the corresponding expectation. We assume that $(\boldsymbol{\theta}, \eta)$ belongs to the set

$$\Omega_\Delta = \{\boldsymbol{\theta} \in \mathbf{R}^p : \|\boldsymbol{\theta}\| \geq \Delta\} \times \{-1, 1\}^n,$$

where $\Delta > 0$ is a given constant. We consider the following Hamming loss of selector $\hat{\eta}$:

$$r(\hat{\eta}, \eta) := \min_{\nu \in \{-1,1\}} |\hat{\eta} - \nu\eta|,$$

and its expected loss defined as $\mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\hat{\eta}, \eta)$. We are interested in the following minimax risk:

$$\Psi_\Delta := \inf_{\tilde{\eta}} \sup_{(\boldsymbol{\theta},\eta)\in\Omega_\Delta} \frac{1}{n} \mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\tilde{\eta}, \eta),$$

where $\inf_{\tilde{\eta}}$ denotes the infimum over all estimators $\tilde{\eta}$ with values in $\{-1, 1\}^n$. After identifying the appropriate signal-to-noise ratio (SNR) $\mathbf{r}_n$ of the problem:

$$\mathbf{r}_n = \frac{\Delta^2/\sigma^2}{\sqrt{\Delta^2/\sigma^2 + p/n}},$$

our main contribution is to prove that

$$\Psi_\Delta \asymp e^{-\mathbf{r}_n^2(1/2+o(1))},$$

where $o(1)$ denotes a bounded sequence that vanishes as n goes to infinity. As a consequence we recover the sharp phase transition for exact recovery in the Gaussian mixture model. Namely, we show that the phase transition occurs around the threshold $\Delta = \bar{\Delta}_n$ such that

$$\bar{\Delta}_n^2 = \sigma^2 \left(1 + \sqrt{1 + \frac{2p}{n \log n}}\right) \log n. \tag{2.4}$$

Moreover, we propose a procedure achieving this threshold. It is a variant of Lloyd's iterations initialized by a spectral method. This procedure is a fully adaptive, rate optimal and computationally simple. The main difference in the proposed method compared to the classical EM style algorithms is the following. Most of these algorithms are based on estimating the center at each step, and this is exactly where they loose optimality. In the high SNR regime, there is no hope to achieve even a non-trivial correlation with the true center vector, but exact recovery of labels is still possible. This means that in high dimension, any algorithm relying on estimation of the center must be suboptimal.

To the best of our knowledge, the suggested procedure is the first fast method to achieve optimal exact recovery. In addition, it achieves sharp optimality since we derive the threshold of (2.4) with precise constant. Moreover, this procedure is as fast as any spectral method in terms of complexity. In other words, the proposed procedure takes the best both of the realm of SDP and of the spectral methods.

2.3 Adaptive robust estimation in sparse vector model

For the sparse vector model

$$Y = \boldsymbol{\theta} + \sigma\xi,$$

estimation of the target vector $\boldsymbol{\theta}$ and of its ℓ_2-norm are classical problems that are of interest to the statistical community. In the case where ξ is Gaussian and σ is known, these questions are well understood. A crucial issue arises when the noise level σ and/or the noise distribution are unknown. Then, one is also interested in the problem of estimation of σ.

The classical Gaussian sequence model corresponds to the case where the noise ξ is standard Gaussian, and the noise level σ is known. Then, the optimal rate of estimation of $\boldsymbol{\theta}$ under the quadratic loss in a minimax sense on the class of s-sparse vectors is given in (2.2) and it is attained by thresholding estimators, cf. Donoho et al. (1992). Also, for the Gaussian sequence model with known σ, minimax optimal estimator of the norm $\|\boldsymbol{\theta}\|$ as well as the corresponding minimax rate are available from Collier et al. (2017). It remains to characterize the effects of ignoring some of these parameters on different minimax estimation rates.

We emphasize here, that the parameter $\boldsymbol{\theta}$ can play two roles. Either it is the parameter of interest to estimate or a nuisance parameter if we are interested in estimation of σ^2. Chen et al. (2018) explore the problem of robust estimation of variance and of covariance matrix under Huber's contamination model. This problem has similarities with estimation of noise level in our setting. Another aspect of robust estimation of scale is analyzed by Wei and Minsker (2017) who consider classes of heavy tailed distributions, rather than the contamination model. The main aim in Wei and Minsker (2017) is to construct estimators having sub-Gaussian deviations under weak moment assumptions. In the sparse linear model, estimation of variance is discussed in Sun and Zhang (2012) where some upper bounds for the rates are given, while estimation of the ℓ_2-norm is discussed in Carpentier et al. (2018). We also mention the recent papers of Collier et al. (2018); Carpentier and Verzelen (2019) that discuss estimation of other functionals than the ℓ_2-norm in the sparse vector model when the noise is Gaussian with unknown variance.

In Chapter 7, we consider separately the setting of Gaussian noise, or when the distribution of ξ_i and the noise level σ are both unknown. For the unknown distribution of ξ_1, we denote by P_ξ the unknown distribution of ξ_1 and consider two types of assumptions. Either P_ξ belongs to a class $\mathcal{G}_{a,\tau}$, i.e. for some $a, \tau > 0$,

$$P_\xi \in \mathcal{G}_{a,\tau} \quad \text{iff} \quad \mathbf{E}(\xi_1) = 0,\ \mathbf{E}(\xi_1^2) = 1 \text{ and } \quad \forall t \geq 2,\ \mathbf{P}\big(|\xi_1| > t\big) \leq 2e^{-(t/\tau)^a},$$

which includes for example sub-Gaussian distributions ($a = 2$), or to a class of distributions with polynomially decaying tails $\mathcal{P}_{a,\tau}$, i.e. for some $\tau > 0$ and $a \geq 2$,

$$P_\xi \in \mathcal{P}_{a,\tau} \quad \text{iff} \quad \mathbf{E}(\xi_1) = 0,\ \mathbf{E}(\xi_1^2) = 1 \text{ and } \quad \forall t \geq 2,\ \mathbf{P}\big(|\xi_1| > t\big) \leq \Big(\frac{\tau}{t}\Big)^a.$$

We are interested in the following maximal risk functions over classes of s-sparse vectors:

$$\sup_{|\boldsymbol{\theta}|_0 \leq s} \sqrt{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|}{\sigma}\right)^2}, \tag{2.5}$$

for estimation of the vector $\boldsymbol{\theta}$,

$$\sup_{|\boldsymbol{\theta}|_0 \le s} \sqrt{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma} \left(\frac{\hat{T} - \|\boldsymbol{\theta}\|}{\sigma} \right)^2}, \tag{2.6}$$

for estimation of the ℓ_2-norm of $\boldsymbol{\theta}$, and

$$\sup_{|\boldsymbol{\theta}|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma} \frac{|\hat{\sigma}^2 - \sigma^2|}{\sigma^2}, \tag{2.7}$$

for estimation of σ^2. The studied minimax risks are obtained by taking $\sup_{P \in \mathcal{P}}$ in (2.5)-(2.7) for $\mathcal{P} = \mathcal{G}_{a,\tau}$, $\mathcal{P} = \mathcal{P}_{a,\tau}$ and $\mathcal{P} = \{\mathcal{N}(0,1)\}$, and/or $\sup_{\sigma>0}$ and then taking the infimum over all estimators.

We summarize the major points that constitute the novelty of our contribution.

- Estimation of σ^2:

  We consider this problem in a minimax setting. We obtain the minimax optimal rates and exhibit minimax estimators for two classes of noise $\mathcal{G}_{a,\tau}$ and $\mathcal{P}_{a,\tau}$ when P_ξ is known or unknwon. We show that when the noise distribution is exactly known (and satisfies rather general assumption, not necessarily Gaussian - can have polynomial tails), then, surprisingly, σ^2 can be estimated faster than on the class of sub-Gaussian noise. We establish minimax rate of estimation of σ^2 for the case of pure Gaussian noise, which is even faster than the rate of the previous item.

- Estimation of the norm $\|\boldsymbol{\theta}\|$:

  The non-asymptotic minimax rate for this problem was known only when the noise is Gaussian with known variance σ^2, cf. Collier et al. (2017). We find minimax rates and rate optimal estimators when the variance σ^2 is either known or unknown, while noise distribution belongs to one of the following classes:

  1. Noise with polynomial tails, $P_\xi \in \mathcal{P}_{a,\tau}$.
  2. Noise with exponential tails, $P_\xi \in \mathcal{G}_{a,\tau}$. An unexpected finding here is that the minimax rate on the class of sub-Gaussian noise is substantially different from the minimax rate when the noise is Gaussian. In particular, the rate does not exhibit an elbow between the "dense" and the "sparse" zones characteristic for minimax rates of estimation of functionals in Gaussian noise.
  3. Gaussian noise with unknown variance.

- Estimation of $\boldsymbol{\theta}$:

  We study the minimax optimality for this problem when the noise belongs to a class of distributions $\mathcal{P}_{a,\tau}$ or $\mathcal{G}_{a,\tau}$. We derive the minimax rates for these settings. An unexpected conclusion is about the dramatic difference between the rates in sparse regression with "well spread" regressors and in sparse mean model. It is known from Gautier and Tsybakov (2013), Belloni et al. (2014) that in sparse regression with "well spread" regressors (i.e. having positive variance), one can

attain sub-Gaussian rates even when the noise has polynomial tails. We show that the situation is completely different in the sparse mean model, where the optimal rates are much slower and depend on the polynomial index of the noise.

We summarize all our findings in Table 2.2.

	Gaussian noise model	Noise in class $\mathcal{G}_{a,\tau}$,	Noise in class $\mathcal{P}_{a,\tau}$,
$\boldsymbol{\theta}$	$\sqrt{s\log(ed/s)}$ known σ Donoho et al. (1992) unknown σ Verzelen (2012)	$\sqrt{s}\log^{\frac{1}{a}}(ed/s)$ unknown σ	$\sqrt{s}(d/s)^{\frac{1}{a}}$ unknown σ
$\|\boldsymbol{\theta}\|_2$	$\sqrt{s\log(1+\frac{\sqrt{d}}{s})}\wedge d^{1/4}$ known σ Collier et al. (2017) $\sqrt{s\log(1+\frac{\sqrt{d}}{s})}\vee\sqrt{\frac{s}{1+\log_+(s^2/d)}}$ unknown σ	$\sqrt{s}\log^{\frac{1}{a}}(ed/s)\wedge d^{1/4}$ known σ $\sqrt{s}\log^{\frac{1}{a}}(ed/s)$ unknown σ	$\sqrt{s}(d/s)^{\frac{1}{a}}\wedge d^{1/4}$ known σ $\sqrt{s}(d/s)^{\frac{1}{a}}$ unknown σ
σ^2	$\frac{1}{\sqrt{d}}\vee\frac{s}{d(1+\log_+(s^2/d))}$	$\frac{1}{\sqrt{d}}\vee\frac{s}{d}\log^{\frac{2}{a}}\left(\frac{ed}{s}\right)$	$\frac{1}{\sqrt{d}}\vee\left(\frac{s}{d}\right)^{1-\frac{2}{a}}$

Table 2.2: Rates of convergence of the minimax risks.

2.4 Simulation of Gaussian processes

Stochastic processes are very popular for modeling dynamics, in particular, they are used to model price fluctuations in finance. It is well known that Gaussian processes can be discretized, then simulated when their covariance is known. Unlike the case of Brownian motion, simulation is costly when increments of the process are no longer independent. The papers by Ayache and Taqqu (2003); Dzhaparidze and Van Zanten (2004) propose optimal methods, in a precise sense, to simulate fractional Brownian motion using different series expansions. These series are complicated and use special functions not giving any intuition on their construction which makes it difficult to generalize them to other processes.

In Chapter 8, we present a new approach to derive series expansions for some Gaussian processes based on harmonic analysis of their covariance function. In particular, a new simple rate-optimal series expansion is derived for fractional Brownian motion:

$$B_t^H = \sqrt{c_0}tZ_0 + \sum_{k=1}^{\infty}\sqrt{\frac{-c_k}{2}}\left(Z_k\sin\frac{k\pi t}{T} + Z_{-k}\left(1-\cos\frac{k\pi t}{T}\right)\right), \quad t\in[0,T],$$

where

$$\begin{cases} c_0 := 0, & H < 1/2 \\ c_0 := HT^{2H-2}, & H > 1/2, \end{cases}$$

$$\forall k \geq 1, \quad \begin{cases} c_k := \frac{2}{T}\int_0^T t^{2H}\cos\frac{k\pi t}{T}\mathrm{d}t, & H < 1/2 \\ c_k := -\frac{4H(2H-1)T}{(k\pi)^2}\int_0^T t^{2H-2}\cos\frac{k\pi t}{T}\mathrm{d}t, & H > 1/2, \end{cases}$$

and $(Z_k)_{k\in\mathbf{Z}}$ is a sequence of independent standard Gaussian random variables. The convergence of the latter series holds in mean square and uniformly almost surely, with a rate-optimal decay of the remainder term of the series. We also develop a general framework of convergent series expansion for certain classes of Gaussian processes.

The main Chapters of this thesis are based, respectively, on the following works:

Butucea, C., Ndaoud, M., Stepanova, N. A., and Tsybakov, A. B. (2018). Variable selection with hamming loss. *The Annals of Statistics, 46(5):1837-1875.*

Ndaoud, M. and Tsybakov, A. B. (2018). Optimal variable selection and adaptive noisy compressed sensing. *arXiv preprint arXiv:1809.03145.*

Ndaoud, M. (2019). Interplay of minimax estimation and minimax support recovery under sparsity. *ALT 2019.*

Ndaoud, M. (2018b). Sharp optimal recovery in the two component Gaussian mixture model. *arXiv preprint arXiv:1812.08078.*

Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. (2018). Adaptive robust estimation in sparse vector model. *arXiv preprint arXiv:1802.04230v3.*

Ndaoud, M. (2018a). Harmonic analysis meets stationarity: A general framework for series expansions of special Gaussian processes. *arXiv preprint arXiv:1810.11850.*

Part I

Variable Selection in High-Dimensional Linear Regression

Chapter 3

Variable selection with Hamming loss

We derive nonasymptotic bounds for the minimax risk of variable selection under expected Hamming loss in the Gaussian mean model in $\mathbb{R}^d$ for classes of at most s-sparse vectors separated from 0 by a constant $a > 0$. In some cases, we get exact expressions for the nonasymptotic minimax risk as a function of d, s, a and find explicitly the minimax selectors. These results are extended to dependent or non-Gaussian observations and to the problem of crowdsourcing. Analogous conclusions are obtained for the probability of wrong recovery of the sparsity pattern. As corollaries, we derive necessary and sufficient conditions for such asymptotic properties as almost full recovery and exact recovery. Moreover, we propose data-driven selectors that provide almost full and exact recovery adaptively to the parameters of the classes.

Based on Butucea et al. (2018): Butucea, C., Ndaoud, M., Stepanova, N. A., and Tsybakov, A. B. (2018). Variable selection with hamming loss. *The Annals of Statistics, 46(5):1837-1875.*

3.1 Introduction

In recent years, the problem of variable selection in high-dimensional regression models has been extensively studied from the theoretical and computational viewpoints. In making effective high-dimensional inference, sparsity plays a key role. With regard to variable selection in sparse high-dimensional regression, the Lasso, Dantzig selector, other penalized techniques as well as marginal regression were analyzed in detail; see, for example, Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Wainwright (2009b); Lounici (2008); Wasserman and Roeder (2009); Zhang (2010); Meinshausen and Bühlmann (2010); Genovese et al. (2012); Ji and Jin (2012) and the references cited therein. Several other recent papers deal with sparse variable selection in nonparametric regression; see, for example, Lafferty and Wasserman (2008); Bertin and Lecué (2008); Comminges and Dalalyan (2012); Ingster and Stepanova (2014); Butucea and Stepanova (2017).

In this chapter, we study the problem of variable selection in the Gaussian sequence model

$$X_j = \theta_j + \sigma\xi_j, \qquad j = 1, \dots, d, \tag{3.1}$$

where $\xi_1, \ldots, \xi_d$ are i.i.d. standard Gaussian random variables, $\sigma > 0$ is the noise level, and $\theta = (\theta_1, \ldots, \theta_d)$ is an unknown vector of parameters to be estimated. We assume that θ is (s, a)-*sparse*, which is understood in the sense that θ belongs to one of the following sets:

$$\Theta_d(s,a) = \big\{\theta \in \mathbb{R}^d : \text{ there exists a set } S \subseteq \{1,\ldots,d\} \text{ with at most } s \text{ elements such that } |\theta_j| \geq a \text{ for all } j \in S, \text{ and } \theta_j = 0 \text{ for all } j \notin S\big\}$$

or

$$\Theta_d^+(s,a) = \big\{\theta \in \mathbb{R}^d : \text{ there exists a set } S \subseteq \{1,\ldots,d\} \text{ with at most } s \text{ elements such that } \theta_j \geq a \text{ for all } j \in S, \text{ and } \theta_j = 0 \text{ for all } j \notin S\big\}.$$

Here, $a > 0$ and $s \in \{1, \ldots, d\}$ are given constants.

We study the problem of selecting the relevant components of θ, that is, of estimating the vector

$$\eta = \eta(\theta) = \big(I(\theta_j \neq 0)\big)_{j=1,\ldots,d},$$

where $I(\cdot)$ is the indicator function. As estimators of η, we consider any measurable functions $\hat{\eta} = \hat{\eta}(X_1, \ldots, X_n)$ of $(X_1, \ldots, X_n)$ taking values in $\{0,1\}^d$. Such estimators will be called *selectors*. We characterize the loss of a selector $\hat{\eta}$ as an estimator of η by the Hamming distance between $\hat{\eta}$ and η, that is, by the number of positions at which $\hat{\eta}$ and η differ:

$$|\hat{\eta} - \eta| \triangleq \sum_{j=1}^{d} |\hat{\eta}_j - \eta_j| = \sum_{j=1}^{d} I(\hat{\eta}_j \neq \eta_j).$$

Here, $\hat{\eta}_j$ and $\eta_j = \eta_j(\theta)$ are the jth components of $\hat{\eta}$ and $\eta = \eta(\theta)$, respectively. The expected Hamming loss of a selector $\hat{\eta}$ is defined as $\mathbf{E}_\theta|\hat{\eta} - \eta|$, where $\mathbf{E}_\theta$ denotes the expectation with respect to the distribution $\mathbf{P}_\theta$ of $(X_1, \ldots, X_n)$ satisfying (3.1). Another well-known risk measure is the probability of wrong recovery $\mathbf{P}_\theta(\hat{S} \neq S(\theta))$, where $\hat{S} = \{j : \hat{\eta}_j = 1\}$ and $S(\theta) = \{j : \eta_j(\theta) = 1\}$. It can be viewed as the Hamming distance with an indicator loss and is related to the expected Hamming loss as follows:

$$\mathbf{P}_\theta\big(\hat{S} \neq S(\theta)\big) = \mathbf{P}_\theta\big(|\hat{\eta} - \eta| \geq 1\big) \leq \mathbf{E}_\theta|\hat{\eta} - \eta|. \tag{3.2}$$

In view of the last inequality, bounding the expected Hamming loss provides a stronger result than bounding the probability of wrong recovery.

Most of the literature on variable selection in high dimensions focuses on the recovery of the sparsity pattern, that is, on constructing selectors such that the probability $\mathbf{P}_\theta(\hat{S} \neq S(\theta))$ is close to 0 in some asymptotic sense (see, e.g., Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Wainwright (2009b); Lounici (2008); Wasserman and Roeder (2009); Zhang (2010); Meinshausen and Bühlmann (2010)). These papers consider high-dimensional linear regression settings with deterministic or random covariates. In particular, for the sequence model (3.1), one gets that if $a > C\sigma\sqrt{\log d}$ for some $C > 0$ large enough, then there exist selectors such that $\mathbf{P}_\theta(\hat{S} \neq S(\theta))$ tends to 0, while this is not the case if $a < c\sigma\sqrt{\log d}$ for some $c > 0$ small enough. More insight into variable selection was provided in Genovese et al. (2012); Ji and Jin (2012) by considering a Hamming risk close to the one we have defined above. Assuming that $s \sim d^{1-\beta}$ for some $\beta \in (0,1)$, the papers Genovese et al. (2012); Ji and Jin (2012) establish an

asymptotic in d "phase diagram" that partitions the parameter space into three regions called the exact recovery, almost full recovery, and no recovery regions. This is done in a Bayesian setup for the linear regression model with i.i.d. Gaussian covariates and random θ. Note also that in Genovese et al. (2012); Ji and Jin (2012) the knowledge of β is required to construct the selectors, so that in this sense the methods are not adaptive. The selectors are of the form $\hat{\eta}_j = I(|X_j| \geq t)$ with threshold $t = \tau(\beta)\sigma\sqrt{\log d}$ for some function $\tau(\cdot) > 0$. More recently, these asymptotic results were extended to a combined minimax–Bayes Hamming risk on a certain class of vectors θ in Jin et al. (2014).

The present paper makes further steps in the analysis of variable selection with a Hamming loss initiated in Genovese et al. (2012); Ji and Jin (2012). Unlike Genovese et al. (2012); Ji and Jin (2012), we study the sequence model (3.1) rather than Gaussian regression and analyze the behavior of the minimax risk rather than that of the Bayes risk with a specific prior. Furthermore, we consider not only $s \sim d^{1-\beta}$ but general s and derive nonasymptotic results that are valid for any sample size. Remarkably, we get an exact expression for the nonasymptotic minimax risk of separable (coordinate-wise) selectors and find explicitly the separable minimax selectors. Finally, we construct data-driven selectors that are simultaneously adaptive to the parameters a and s.

Specifically, we consider the minimax risk

$$\inf_{\tilde{\eta}} \sup_{\theta \in \Theta} \frac{1}{s} \mathbf{E}_\theta |\tilde{\eta} - \eta| \tag{3.3}$$

for $\Theta = \Theta_d(s, a)$ and $\Theta = \Theta_d^+(s, a)$, where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors $\tilde{\eta}$. In Section 3.2, for both classes $\Theta = \Theta_d(s, a)$ and $\Theta = \Theta_d^+(s, a)$ we find the upper and lower bounds of the minimax risks and derive minimax selectors for any fixed $d, s, a > 0$ such that $s < d$. For $\Theta = \Theta_d(s, a)$, we also propose another selector attaining the lower bound risk up to the factor 2. Interestingly, the thresholds that correspond to the minimax optimal selectors do not have the classical form $A\sigma\sqrt{\log d}$ for some $A > 0$; the optimal threshold is a function of a and s. Analogous minimax results are obtained for the risk measured by the probability of wrong recovery $\mathbf{P}_\theta(\hat{S} \neq S(\theta))$. Section 3.3 considers extensions of the nonasymptotic minimax theorems of Section 3.2 to settings with non-Gaussian or dependent observations. In Section 3.4, as asymptotic corollaries of these results, we establish sharp conditions under which exact and almost full recovery are achievable. Section 3.5 is devoted to the construction of adaptive selectors that achieve almost full and exact recovery without the knowledge of the parameters a and s. Most of the proofs are given in Appendix 3.6.

Finally, note that quite recently several papers have studied the expected Hamming loss in other problems of variable selection. Asymptotic behavior of the minimax risk analogous to (3.3) for classes Θ different from the sparsity classes that we consider here was analyzed in Butucea and Stepanova (2017) and without the normalizing factor $1/s$ in Ingster and Stepanova (2014). Oracle inequalities for Hamming risks in the problem of multiple classification under sparsity constraints are established in Neuvial and Roquain (2012). The paper Zhang and Zhou (2016) introduces an asymptotically minimax approach based on the Hamming loss in the problem of community detection in networks.

3.2 Nonasymptotic minimax selectors

In what follows, we assume that $s < d$. We first consider minimax variable selection for the class $\Theta_d^+(s,a)$. For this class, we will use a selector $\hat{\eta}^+$ with the components

$$\hat{\eta}_j^+ = I(X_j \geq t), \qquad j = 1, \ldots, d, \tag{3.4}$$

where the threshold is defined by

$$t = \frac{a}{2} + \frac{\sigma^2}{a} \log\left(\frac{d}{s} - 1\right). \tag{3.5}$$

Set

$$\Psi_+(d,s,a) = \left(\frac{d}{s} - 1\right)\Phi\left(-\frac{a}{2\sigma} - \frac{\sigma}{a}\log\left(\frac{d}{s} - 1\right)\right) + \Phi\left(-\frac{a}{2\sigma} + \frac{\sigma}{a}\log\left(\frac{d}{s} - 1\right)\right),$$

where $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function.

Theorem 3.2.1. *For any $a > 0$ and $s \leq d/2$, the selector $\hat{\eta}^+$ in (3.4) with the threshold t defined in (3.5) satisfies*

$$\sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta\left|\hat{\eta}^+ - \eta\right| \leq \Psi_+(d,s,a). \tag{3.6}$$

The proof is given in Appendix 3.6.

A selector $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)$ will be called *separable* if its jth component $\tilde{\eta}_j$ depends only on X_j for all $j = 1, \ldots, d$. We denote by $\mathcal{T}$ the set of all separable selectors.

The next theorem gives a lower bound on the minimax risk showing that the upper bound in Theorem 3.2.1 is tight over separable selectors.

Theorem 3.2.2. *For any $a > 0$ and $s < d$, we have*

$$\inf_{\tilde{\eta} \in \mathcal{T}} \sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta|\tilde{\eta} - \eta| \geq \Psi_+(d,s,a),$$

where $\inf_{\tilde{\eta}\in\mathcal{T}}$ denotes the infimum over all separable selectors $\tilde{\eta}$. Moreover, for any s' in $(0, s]$, we have

$$\inf_{\tilde{\eta}} \sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta|\tilde{\eta} - \eta| \geq \frac{s'}{s}\Psi_+(d,s,a) - \frac{4s'}{s}\exp\left(-\frac{(s-s')^2}{2s}\right),$$

where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors $\tilde{\eta}$.

The proof of the first inequality of Theorem 3.2.2 is given in Appendix 3.6, while the proof of the second inequality is given Appendix 3.7.

As a straightforward corollary of Theorems 3.2.1 and 3.2.2, we obtain that the estimator $\hat{\eta}^+$ is minimax among the separable selectors in the exact sense for the class $\Theta_d^+(s,a)$ and the minimax risk satisfies

$$\inf_{\tilde{\eta}\in\mathcal{T}} \sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta|\tilde{\eta} - \eta| = \sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta\left|\hat{\eta}^+ - \eta\right| = \Psi_+(d,s,a). \tag{3.7}$$

Remarkably, this holds under no assumptions on d, s, a except for, of course, some minimal conditions under which the problem ever makes sense: $a > 0$ and $s \le d/2$. Analogous non-asymptotic minimax result is valid for the class

$$\Theta_d^-(s,a) = \big\{\theta \in \mathbb{R}^d : \text{ there exists a set } S \subseteq \{1,\dots,d\} \text{ with at most } s \text{ elements such that } \theta_j \le -a \text{ for all } j \in S, \text{ and } \theta_j = 0 \text{ for all } j \notin S\big\}.$$

We omit details here.

Next, consider the class $\Theta_d(s,a)$. A direct analog of $\hat\eta^+$ for $\Theta_d(s,a)$ is a selector $\hat\eta$ with the components

$$\hat\eta_j = I\big(|X_j| \ge t\big), \qquad j = 1,\dots,d, \tag{3.8}$$

where the threshold t is defined in (3.5). Set

$$\Psi(d,s,a) = \left(\frac{d}{s} - 1\right)\Phi\left(-\frac{a}{2\sigma} - \frac{\sigma}{a}\log\left(\frac{d}{s} - 1\right)\right) + \Phi\left(-\left(\frac{a}{2\sigma} - \frac{\sigma}{a}\log\left(\frac{d}{s} - 1\right)\right)_+\right),$$

where $x_+ = \max(x, 0)$. Note that

$$\Psi(d,s,a) \le \Psi_+(d,s,a). \tag{3.9}$$

We have the following bound.

Theorem 3.2.3. *For any $a > 0$ and $s \le d/2$, the selector $\hat\eta$ in (3.8) with the threshold t defined in (3.5) satisfies*

$$\sup_{\theta \in \Theta_d(s,a)} \frac{1}{s}\mathbf{E}_\theta|\hat\eta - \eta| \le 2\Psi(d,s,a). \tag{3.10}$$

The proof is given in Appendix 3.6.

For the minimax risk on the class $\Theta_d(s,a)$, we have the following corollary, which is an immediate consequence of Theorems 3.2.2, 3.2.3 and inequality (3.9).

Corollary 3.2.1. *For any $a > 0$ and $s \le d/2$, the selector $\hat\eta$ in (3.8) with the threshold t defined in (3.5) satisfies*

$$\sup_{\theta \in \Theta_d(s,a)} \mathbf{E}_\theta|\hat\eta - \eta| \le 2 \inf_{\tilde\eta \in \mathcal{T}} \sup_{\theta \in \Theta_d(s,a)} \mathbf{E}_\theta|\tilde\eta - \eta|. \tag{3.11}$$

Thus, the risk of the thresholding estimator (3.8) cannot be greater than the minimax risk of separable selectors over the class $\Theta_d(s,a)$ multiplied by 2.

We turn now to exact minimax variable selection over the class $\Theta_d(s,a)$. Consider a selector $\bar\eta = (\bar\eta_1,\dots,\bar\eta_d)$ with the components

$$\bar\eta_j = I\left(\log\left(\cosh\left(\frac{aX_j}{\sigma^2}\right)\right) \ge t\right), \qquad j = 1,\dots,d, \tag{3.12}$$

where the threshold is defined by

$$t = \frac{a^2}{2\sigma^2} + \log\left(\frac{d}{s} - 1\right). \tag{3.13}$$

Set

$$\begin{aligned}\bar{\Psi}(d,s,a) &= \left(\frac{d}{s}-1\right)\mathbf{P}\left(e^{-\frac{a^2}{2\sigma^2}}\cosh\left(\frac{a\xi}{\sigma}\right)\geq\frac{d}{s}-1\right)\\ &\quad+\mathbf{P}\left(e^{-\frac{a^2}{2\sigma^2}}\cosh\left(\frac{a\xi}{\sigma}+\frac{a^2}{\sigma^2}\right)<\frac{d}{s}-1\right),\end{aligned}$$

where ξ denotes a standard Gaussian random variable. Our aim is to show that $\bar{\Psi}(d,s,a)$ is the minimax risk of variable selection under the Hamming loss over the class $\Theta_d(s,a)$ and that it is nearly achieved by the selector in (3.12). We first prove that $\bar{\Psi}(d,s,a)d/(d-s)$ is an upper bound on the maximum risk of the selector (3.12).

Theorem 3.2.4. *For any $a>0$ and $s<d$, the selector $\bar{\eta}$ in (3.12) with the threshold t defined in (3.13) satisfies*

$$\sup_{\theta\in\Theta_d(s,a)}\frac{1}{s}\mathbf{E}_\theta|\bar{\eta}-\eta|\leq\bar{\Psi}(d,s,a)\frac{d}{d-s}.$$

The next theorem establishes the lower bound over separable selectors on the minimax risk associated to the upper bound in Theorem 3.2.4.

Theorem 3.2.5. *For any $a>0$ and $s<d$, we have*

$$\inf_{\tilde{\eta}\in\mathcal{T}}\sup_{\theta\in\Theta_d(s,a)}\frac{1}{s}\mathbf{E}_\theta|\tilde{\eta}-\eta|\geq\bar{\Psi}(d,s,a),$$

where $\inf_{\tilde{\eta}\in\mathcal{T}}$ denotes the infimum over all separable selectors $\tilde{\eta}$.

Finally, we show how the above nonasymptotic minimax results can be extended to the probability of wrong recovery. For any selector $\tilde{\eta}$, we denote by $S_{\tilde{\eta}}$ the selected set of indices: $S_{\tilde{\eta}}=\{j:\tilde{\eta}_j=1\}$.

Theorem 3.2.6. *For any $a>0$ and $s\leq d/2$, the selectors $\hat{\eta}$ in (3.8) and $\hat{\eta}^+$ in (3.4) with the threshold t defined in (3.5), and the selector $\bar{\eta}$ in (3.12) with the threshold t defined in (3.13) satisfy*

$$\sup_{\theta\in\Theta_d^+(s,a)}\mathbf{P}_\theta\big(S_{\hat{\eta}^+}\neq S(\theta)\big)\leq s\Psi_+(d,s,a), \tag{3.14}$$

$$\sup_{\theta\in\Theta_d(s,a)}\mathbf{P}_\theta\big(S_{\bar{\eta}}\neq S(\theta)\big)\leq s\bar{\Psi}(d,s,a)\frac{d}{d-s} \tag{3.15}$$

and

$$\sup_{\theta\in\Theta_d(s,a)}\mathbf{P}_\theta\big(S_{\hat{\eta}}\neq S(\theta)\big)\leq 2s\Psi(d,s,a). \tag{3.16}$$

Furthermore,

$$\inf_{\tilde{\eta}\in\mathcal{T}}\sup_{\theta\in\Theta_d^+(s,a)}\mathbf{P}_\theta\big(S_{\tilde{\eta}}\neq S(\theta)\big)\geq\frac{s\Psi_+(d,s,a)}{1+s\Psi_+(d,s,a)} \tag{3.17}$$

and

$$\inf_{\tilde{\eta}\in\mathcal{T}}\sup_{\theta\in\Theta_d(s,a)}\mathbf{P}_\theta\big(S_{\tilde{\eta}}\neq S(\theta)\big)\geq\frac{s\bar{\Psi}(d,s,a)}{1+s\bar{\Psi}(d,s,a)}. \tag{3.18}$$

The proof is given in Appendix 3.6.

Although Theorem 3.2.6 does not provide the exact minimax solution, it implies sharp minimaxity in asymptotic sense. Indeed, an interesting case is when the minimax risk in Theorem 3.2.6 goes to 0 as $d \to \infty$. Assuming that s and a are functions of d, this corresponds to $s\Psi_+(d,s,a) \to 0$ as $d \to \infty$. In this natural asymptotic setup, the upper and lower bounds of Theorem 3.2.6 for the class $\Theta_d^+(s,a)$ are sharp. The same for the class $\Theta_d(s,a)$, if s and a are such that $s\bar{\Psi}(d,s,a) \to 0$ and that $s/d \to 0$. We discuss this issue in Section 3.4; cf. Theorem 3.4.5.

Remark 3.2.1. *Papers Genovese et al. (2012); Ji and Jin (2012); Jin et al. (2014) use a different Hamming loss defined in terms of vectors of signs. In our setting, this would mean considering not $|\hat{\eta} - \eta|$ but the following loss: $\sum_{j=1}^d I(\mathrm{sign}(\hat{\theta}_j) \neq \mathrm{sign}(\theta_j))$, where $\hat{\theta}_j$ is an estimator of θ_j and $\mathrm{sign}(x) = I(x > 0) - I(x < 0)$. Theorems of this section are easily adapted to such a loss, but in this case the corresponding expressions for the nonasymptotic risk contain additional terms and we do not obtain exact minimax solutions as above. On the other hand, these additional terms are smaller than $\Psi(d,s,a)$ and $\Psi_+(d,s,a)$, and in the asymptotic analysis, such as the one performed in Sections 3.4 and 3.5, can often be neglected. Thus, in many cases, one gets the same asymptotic results for both losses. We do not discuss this issue in more detail here.*

3.3 Generalizations and extensions

Before proceeding to asymptotic corollaries, we discuss some generalizations and extensions of the nonasymptotic results of Section 3.2.

Dependent observations

It is easy to see that Theorems 3.2.1 and 3.2.3 do not use any information on the dependence between the observations, and thus remain valid for dependent X_j. Furthermore, a minimax optimality property within the class of separable selectors holds under dependence as well. To be specific, denote by $\mathcal{N}_d(\theta, \Sigma)$ the d-dimensional Gaussian distribution with mean θ and covariance matrix Σ. Assume that the distribution $\mathbf{P}$ of $(X_1, \dots, X_d)$ belongs to the class

$$\mathcal{P}_d^+\left(s, a, \sigma^2\right) = \left\{\mathcal{N}_d(\theta, \Sigma) : \theta \in \Theta_d^+(s,a), \sigma_{ii} = \sigma^2, \text{ for all } i = 1, \dots, d\right\},$$

where we denote by σ_{ii} the diagonal entries of Σ. Note that, for distributions in this class, Σ can be any covariance matrix with constant diagonal elements.

Theorem 3.3.1. *For any $a > 0$ and $s \le d/2$, and for the selector $\hat{\eta}^+$ in (3.4) with the threshold t defined in (3.5) we have*

$$\inf_{\tilde{\eta} \in \mathcal{T}} \sup_{\mathbf{P} \in \mathcal{P}_d^+(s,a,\sigma^2)} \mathbf{E}_{\mathbf{P}} |\tilde{\eta} - \eta| = \sup_{\mathbf{P} \in \mathcal{P}_d^+(s,a,\sigma^2)} \mathbf{E}_{\mathbf{P}} \left|\hat{\eta}^+ - \eta\right| = s\Psi_+(d,s,a),$$

where $\inf_{\tilde{\eta} \in \mathcal{T}}$ denotes the infimum over all separable selectors $\tilde{\eta}$, and $\mathbf{E}_{\mathbf{P}}$ denotes the expectation with respect to $\mathbf{P}$. Moreover, for any s' in $(0, s]$, we have

$$\inf_{\tilde{\eta}} \sup_{\mathbf{P} \in \mathcal{P}_d^+(s,a,\sigma^2)} \mathbf{E}_{\mathbf{P}} |\tilde{\eta} - \eta| \ge s'\Psi_+(d,s,a) - 4s' \exp\left(-\frac{(s-s')^2}{2s}\right),$$

where $\inf_{\tilde\eta}$ *denotes the infimum over all selectors* $\tilde\eta$.

Proof. The upper bound $\Psi_+(d,s,a)$ on the minimax risk follows from the fact that the proofs of Theorems 3.2.1 and 3.2.3 are not affected by the dependence. Indeed, both the selector and the Hamming loss proceed coordinate-wisely. The lower bound on the minimax risk follows from Theorem 3.2.2 after observing that the maximum over $\mathcal{P}_d^+(s,a,\sigma^2)$ is greater than the maximum over the subfamily of Gaussian vectors with independent entries $\{\mathcal{N}_d(\theta,\sigma^2 I_d) : \theta \in \Theta_d^+(s,a)\}$, where I_d is the $d\times d$ identity matrix. □

An interesting consequence of Theorem 3.3.1 and of (3.7) is that the model with independent X_j is the least favorable model, in the exact nonasymptotic sense, for the problem of variable selection with Hamming loss on the class of vectors $\Theta_d^+(s,a)$.

This fact was also noticed and discussed in Hall and Jin (2010) for the detection problem. That paper considers the Gaussian model with covariance matrix Σ that is not necessarily a diagonal matrix. It is shown that faster detection rates are achieved in the case of dependent observations (under some assumptions) than in the case of independent data. It would be interesting to extend these results to the variable selection problem in hand.

Non-Gaussian models

As a building block for extension to non-Gaussian observations, we first consider the following simple model. We observe independent random variables $X_1,\dots,X_d$ with values in a measurable space $(\mathcal{X},\mathcal{U})$ such that at most s among them are distributed according to the probability distribution P_1 and the other are distributed according to the probability distribution P_0. We assume that $P_0 \neq P_1$. Let f_0 and f_1 be densities of P_0 and P_1 with respect to some dominating measure. Denote by $\eta=(\eta_1,\dots,\eta_d)$ the vector such that $\eta_j=1$ if the distribution of X_j is P_1 and $\eta_j=0$ if it is P_0. Define $\Theta_d(s)$ as the set of all vectors $\eta\in\{0,1\}^d$ with at most s non-zero components. For any fixed η, we denote by $\mathbf{E}_\eta$ the expectation with respect to the distribution of $(X_1,\dots,X_d)$. Consider the selector $\hat\eta=(\hat\eta_1,\dots,\hat\eta_d)$, where

$$\hat\eta_j = I\Big(sf_1(X_j) \ge (d-s)f_0(X_j)\Big), \qquad j=1,\dots,d. \tag{3.19}$$

Theorem 3.3.2. *For any* $s<d$, *the selector* $\hat\eta$ *in* (3.19) *satisfies*

$$\sup_{\eta\in\Theta_d(s)} \mathbf{E}_\eta \frac{1}{s}|\hat\eta-\eta| \le \Psi(d,s)\frac{d}{d-s},$$

and, for any s' *in* $(0,s]$,

$$\inf_{\tilde\eta}\sup_{\eta\in\Theta_d(s)} \frac{1}{s}\mathbf{E}_\eta|\tilde\eta-\eta| \ge \frac{s'}{s}\Psi(d,s) - \frac{4s'}{s}\exp\left(-\frac{(s-s')^2}{2s}\right), \tag{3.20}$$

where $\inf_{\tilde\eta}$ *denotes the infimum over all selectors, and*

$$\begin{aligned}\Psi=\Psi(d,s) = {} & P_1\Big(sf_1(X_1) < (d-s)f_0(X_1)\Big)\\ & + \left(\frac{d}{s}-1\right)P_0\Big(sf_1(X_1)\ge (d-s)f_0(X_1)\Big).\end{aligned} \tag{3.21}$$

The proof is given in Appendix 3.7.

Suppose now that instead of two measures P_0 and P_1 we have a parametric family of probability measures $\{\mathbb{P}_a, a \in \mathcal{U}\}$ where $\mathcal{U} \subseteq \mathbb{R}$. Let f_a be a density of $\mathbb{P}_a$ with respect to some dominating measure. Recall that the family $\{\mathsf{f}_a, a \in \mathcal{U}\}$ is said to have the Monotone Likelihood Ratio (MLR) property if, for all a_0, a_1 in $\mathcal{U}$ such that $a_0 < a_1$, the log-likelihood ratio $\log(\mathsf{f}_{a_1}(X)/\mathsf{f}_{a_0}(X))$ is an increasing function of X. In particular, this implies (cf. Lehmann and Romano (2006), Lemma 3.4.2) that $\{\mathsf{f}_a, a \in \mathcal{U}\}$ is a stochastically ordered family, that is,

$$F_a(x) \geq F_{a'}(x) \qquad \text{for all } x \text{ if } a < a', \tag{3.22}$$

where F_a is the cumulative distribution function corresponding to f_a. Using these facts, we generalize the nonasymptotic results of the previous section in two ways. First, we allow for not necessarily Gaussian distributions and second, instead of the set of parameters $\Theta_d^+(s, a)$, we consider the following set with two restrictions:

$$\begin{aligned}\Theta_d^+(s, a_0, a_1) \;=\; &\big\{\theta \in \mathbb{R}^d : \exists \text{ a set } S \subseteq \{1, \dots, d\} \text{ with at most } s \text{ elements} \\ &\text{such that } \theta_j \geq a_1 \text{ for all } j \in S, \text{ and } \theta_j \leq a_0 \text{ for all } j \notin S\big\},\end{aligned}$$

where $a_0 < a_1$. We assume that X_j is distributed with density f_{θ_j} for $j = 1, \dots, d$, and $X_1, \dots, X_d$ are independent. In the next theorem, $\mathbf{E}_\theta$ is the expectation with respect to the distribution of such $X_1, \dots, X_d$. In what follows, we use the notation $f_j = \mathsf{f}_{a_j}, j = 0, 1$.

Theorem 3.3.3. *Let $\{\mathsf{f}_a, a \in \mathcal{U}\}$ be a family with the MLR property, and let $a_0, a_1 \in \mathcal{U}$ be such that $a_0 < a_1$. Set $f_0 = \mathsf{f}_{a_0}$ and $f_1 = \mathsf{f}_{a_1}$, then, for any $s < d$, the selector $\hat{\eta}$ in (3.19) satisfies*

$$\sup_{\theta \in \Theta_d^+(s, a_0, a_1)} \frac{1}{s} \mathbf{E}_\theta |\hat{\eta} - \eta| \leq \Psi(d, s) \frac{d}{d - s},$$

and, for any s' in $(0, s]$,

$$\inf_{\tilde{\eta}} \sup_{\theta \in \Theta_d^+(s, a_0, a_1)} \frac{1}{s} \mathbf{E}_\theta |\tilde{\eta} - \eta| \geq \frac{s'}{s} \Psi(d, s) - \frac{4s'}{s} \exp\left(-\frac{(s - s')^2}{2s}\right),$$

where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors and Ψ is given in (3.21).

The proof is given in Appendix 3.7.

Example 3.3.1. *Let f_a be the Gaussian $\mathcal{N}(a, \sigma^2)$ density with some $\sigma^2 > 0$, and let $a_0 < a_1$. For $f_1 = \mathsf{f}_{a_1}$ and $f_0 = \mathsf{f}_{a_0}$, the log-likelihood ratio*

$$\log \frac{f_1}{f_0}(X) = X \frac{a_1 - a_0}{\sigma^2} - \frac{a_1^2 - a_0^2}{2\sigma^2}$$

is increasing in X. By Theorem 3.3.3, the selector $\hat{\eta}$ on the class $\Theta_d^+(s, a_0, a_1)$ is a vector with components

$$\hat{\eta}_j = I\big(X_j \geq t(a_0, a_1)\big), \qquad j = 1, \dots, d, \tag{3.23}$$

where

$$t(a_0, a_1) = \frac{a_1 + a_0}{2} + \frac{\sigma^2 \log(d/s - 1)}{a_1 - a_0}.$$

Note that for $a_0 = 0$ it coincides with the selector in (3.4) with $a = a_1$, which is minimax optimal on $\Theta_d^+(s, a_1)$. Moreover, the minimax risk only depends on a_0 and a_1 through the difference $\delta = a_1 - a_0$:

$$\Psi = \Phi\left(-\frac{\delta}{2} + \frac{\sigma^2 \log(d/s - 1)}{\delta}\right) + \left(\frac{d}{s} - 1\right)\Phi\left(-\frac{\delta}{2} + \frac{\sigma^2 \log(d/s - 1)}{\delta}\right).$$

Example 3.3.2. *Let $\mathbb{P}_a$ be the Bernoulli distribution $B(a)$ with parameter $a \in (0, 1)$, and $0 < a_0 < a_1 < 1$. Denoting by f_a the density of $\mathbb{P}_a$ with respect to the counting measure we have, for $f_1 = \mathsf{f}_{a_1}$ and $f_0 = \mathsf{f}_{a_0}$,*

$$\log \frac{f_1}{f_0}(X) = X \log\left(\frac{a_1}{1 - a_1}\frac{1 - a_0}{a_0}\right) + \log \frac{1 - a_1}{1 - a_0}$$

which is increasing in X for $0 < a_0 < a_1 < 1$. The nearly minimax optimal selector $\hat{\eta}$ on the class $\Theta_d^+(s, a_0, a_1)$ is a vector with components $\hat{\eta}_j$ in (3.23) where the threshold $t(a_0, a_1)$ is given by

$$t(a_0, a_1) = \frac{\log(\frac{d}{s} - 1) - \log \frac{1 - a_1}{1 - a_0}}{\log(\frac{a_1}{1 - a_1}\frac{1 - a_0}{a_0})}.$$

Note that the nearly minimax selector $\hat{\eta}_j$ differs from the naive selector $\hat{\eta}_j^n = X_j$. Indeed since $X_j \in \{0, 1\}$ we have $\hat{\eta}_j = 1$ if either $X_j = 1$ or $t(a_0, a_1) \le 0$, and $\hat{\eta}_j = 0$ if either $X_j = 0$ or $t(a_0, a_1) > 1$. The value Ψ in the risk has the form

$$\begin{aligned}
\Psi &= \mathbb{P}_{a_1}\big(X_1 < t(a_0, a_1)\big) + \left(\frac{d}{s} - 1\right)\mathbb{P}_{a_0}\big(X_1 \ge t(a_0, a_1)\big) \\
&= \begin{cases} d/s - 1, & t(a_0, a_1) \le 0, \\ 1 - a_1 + a_0(d/s - 1), & 0 < t(a_0, a_1) < 1, \\ 1, & t(a_0, a_1) \ge 1. \end{cases}
\end{aligned}$$

In the asymptotic regime when $d \to \infty$ and $s \to \infty$, the minimax risk is of order $s\Psi$ and can converge to 0 only when the parameters d, s, a_0, a_1 are kept such that $0 < t(a_0, a_1) < 1$, and in addition $(1 - a_1)s \to 0$, $a_0(d - s) \to 0$. Thus, the risk can converge to 0 only when the Bernoulli probabilities a_1 and a_0 tend sufficiently fast to 1 and to 0, respectively.

Example 3.3.3. *Let $\mathbb{P}_a$ be the Poisson distribution with parameter $a > 0$, and let $a_1 > a_0 > 0$. Denoting by f_a the density of $\mathbb{P}_a$ with respect to the counting measure we have*

$$\log \frac{f_1}{f_0}(X) = X \log\left(\frac{a_1}{a_0}\right) - a_1 + a_0,$$

which is increasing in X. The components of the nearly minimax optimal selector $\hat{\eta}$ are given by (3.23) with

$$t(a_0, a_1) = \frac{\log(d/s - 1) + a_1 - a_0}{\log(a_1/a_0)}.$$

Note that $t(a_0, a_1) > 0$ as soon as $d/s \ge 2$ and $a_1 > a_0 > 0$. The value of Ψ in the risk has the form $\Psi = \mathbb{P}_{a_1}(X_1 < t(a_0, a_1)) + (d/s - 1)\mathbb{P}_{a_0}(X_1 \ge t(a_0, a_1))$.

Crowdsourcing with sparsity constraint

The problem of crowdsourcing with two classes is a clustering problem that can be formalized as follows; cf. Gao et al. (2016). Assume that m workers provide class assignments for d items. The class assignment X_{ij} of the ith worker for the jth item is assumed to have a Bernoulli distribution $B(a_{i0})$ if the jth item belongs to class 0, and a Bernoulli distribution $B(a_{i1})$ if it belongs to class 1. Here, $a_{i0}, a_{i1} \in (0,1)$ and $a_{i0} \neq a_{i1}$ for $i = 1, \dots, m$. The observations $(X_{ij}, i = 1, \dots, m, j = 1, \dots, d)$ are assumed to be jointly independent. Thus, each vector $X_j = (X_{1j}, \dots, X_{mj})$ is distributed according to P_0 or to P_1 where each of these two measures is a product of Bernoulli measures, and $P_0 \neq P_1$. We assume that there are at most s vectors X_j with distribution P_1, and the other vectors X_j with distribution P_0. The aim is to recover the binary vector of class labels $\eta = (\eta_1, \dots, \eta_d)$ based on the observations $\mathbf{X} = (X_1, \dots, X_d)$. Here, $\eta_j \in \{0,1\}$ satisfies $\eta_j = k$ if the jth item belongs to class $k \in \{0,1\}$. Thus, we are in the framework of Theorem 3.3.2 with a particular form of the log-likelihood ratio

$$\log \frac{f_1}{f_0}(X_j) = \sum_{i=1}^{m} \left(X_{ij} \log \left(\frac{a_{i1}}{1-a_{i1}} \frac{1-a_{i0}}{a_{i0}} \right) + \log \frac{1-a_{i1}}{1-a_{i0}} \right), \tag{3.24}$$

where f_k is the density of P_k, $k \in \{0,1\}$. The following corollary is an immediate consequence of Theorem 3.3.2.

Corollary 3.3.1. *Let $s < d$, $a_{i0}, a_{i1} \in (0,1)$ and $a_{i0} \neq a_{i1}$ for $i = 1, \dots, m$. Then the selector $\hat{\eta}$ in (3.19) with $\log \frac{f_1}{f_0}(X_j)$ defined in (3.24) satisfies Theorem 3.3.2.*

For suitable combinations of parameters d, s, a_{i0}, a_{i1}, the exact asymptotic value of the minimax risk Ψ can be further analyzed to obtain asymptotics of interest. Gao et al. Gao et al. (2016) have studied a setting of crowdsourcing problem which is different from the one we consider here. They did not assume sparsity s, and instead of the class $\Theta_d(s, f_0, f_1)$ of at most s-sparse binary sequences, they considered the class of all possible binary sequences $\{0,1\}^d$. For this class, Gao et al. Gao et al. (2016) analyzed specific asymptotics of the minimax risk $\inf_{\tilde{\eta}} \sup_{\eta \in \{0,1\}^d} d^{-1} \mathbf{E} |\tilde{\eta} - \eta|$ in large deviations perspective.

3.4 Asymptotic analysis. Phase transitions

In this section, we conduct the asymptotic analysis of the problem of variable selection. The results are derived as corollaries of the minimax bounds of Section 3.2. We will assume that $d \to \infty$ and that parameters $a = a_d$ and $s = s_d$ depend on d.

The first two asymptotic properties we study here are *exact recovery* and *almost full recovery*. We use this terminology following Genovese et al. (2012); Ji and Jin (2012) but we define these properties in a different way, as asymptotic minimax properties for classes of vectors θ. The papers Genovese et al. (2012); Ji and Jin (2012) considered a Bayesian setup with random θ and studied a linear regression model with i.i.d. Gaussian regressors rather than the sequence model (3.1).

The study of *exact recovery* and *almost full recovery* will be done here only for the classes $\Theta_d(s_d, a_d)$. The corresponding results for the classes $\Theta_d^+(s_d, a_d)$ or $\Theta_d^-(s_d, a_d)$ are completely analogous. We do not state them here for the sake of brevity.

Definition 3.4.1. *Let $(\Theta_d(s_d, a_d))_{d\geq 1}$ be a sequence of classes of sparse vectors:*

- *We say that* exact recovery is possible *for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if there exists a selector $\hat{\eta}$ such that*

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat{\eta} - \eta| = 0. \tag{3.25}$$

  In this case, we say that $\hat{\eta}$ achieves exact recovery.

- *We say that* almost full recovery is possible *for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if there exists a selector $\hat{\eta}$ such that*

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d}\mathbf{E}_\theta|\hat{\eta} - \eta| = 0. \tag{3.26}$$

  In this case, we say that $\hat{\eta}$ achieves almost full recovery.

It is of interest to characterize the sequences $(s_d, a_d)_{d\geq 1}$, for which exact recovery and almost full recovery are possible. To describe the impossibility of exact or almost full recovery, we need the following definition.

Definition 3.4.2. *Let $(\Theta_d(s_d, a_d))_{d\geq 1}$ be a sequence of classes of sparse vectors:*

- *We say that* exact recovery is impossible *for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if*

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\tilde{\eta} - \eta| > 0, \tag{3.27}$$

- *We say that* almost full recovery is impossible *for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if*

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d}\mathbf{E}_\theta|\tilde{\eta} - \eta| > 0, \tag{3.28}$$

  where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors.

The following general characterization theorem is a straightforward corollary of the results of Section 3.2.

Theorem 3.4.1. (i) *Almost full recovery is possible for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if and only if $s_d \to \infty$ and*

$$\Psi_+(d, s_d, a_d) \to 0 \qquad \text{as } d \to \infty. \tag{3.29}$$

In this case, the selector $\hat{\eta}$ defined in (3.8) with threshold (3.5) achieves almost full recovery.

(ii) *Exact recovery is possible for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if and only if*

$$s_d\Psi_+(d, s_d, a_d) \to 0 \qquad \text{as } d \to \infty. \tag{3.30}$$

In this case, the selector $\hat{\eta}$ defined in (3.8) with threshold (3.5) achieves exact recovery.

Although this theorem gives a complete solution to the problem, conditions (3.29) and (3.30) are not quite explicit. Intuitively, we would like to get a "phase transition" values a_d^* such that exact (or almost full) recovery is possible for a_d greater than a_d^* and is impossible for a_d smaller than a_d^*. Our aim now is to find such "phase transition" values. We first do it in the almost full recovery framework.

The following bounds for the tails of Gaussian distribution will be useful:

$$\sqrt{\frac{2}{\pi}}\frac{e^{-y^2/2}}{y+\sqrt{y^2+4}} < \frac{1}{\sqrt{2\pi}}\int_y^\infty e^{-u^2/2}\,du \le \sqrt{\frac{2}{\pi}}\frac{e^{-y^2/2}}{y+\sqrt{y^2+8/\pi}}, \tag{3.31}$$

for all $y \ge 0$. These bounds are an immediate consequence of formula 7.1.13 in Abramowitz and Stegun (1964) with $x = y/\sqrt{2}$.

Furthermore, we will need some nonasymptotic bounds for the expected Hamming loss that will play a key role in the subsequent asymptotic analysis. They are given in the next theorem.

Theorem 3.4.2. *Assume that $s < d/2$.*

(i) If

$$a^2 \ge \sigma^2\big(2\log\big((d-s)/s\big) + W\big) \qquad \text{for some } W > 0, \tag{3.32}$$

then the selector $\hat\eta$ defined in (3.8) with threshold (3.5) satisfies

$$\sup_{\theta\in\Theta_d(s,a)} \mathbf{E}_\theta|\hat\eta - \eta| \le (2+\sqrt{2\pi})s\Phi(-\Delta), \tag{3.33}$$

where Δ is defined by

$$\Delta = \frac{W}{2\sqrt{2\log((d-s)/s) + W}}. \tag{3.34}$$

(ii) If $a > 0$ is such that

$$a^2 \le \sigma^2\big(2\log\big((d-s)/s\big) + W\big) \qquad \text{for some } W > 0, \tag{3.35}$$

then, for any s' in $(0, s]$ we have

$$\inf_{\tilde\eta}\sup_{\theta\in\Theta_d(s,a)} \mathbf{E}_\theta|\tilde\eta - \eta| \ge s'\Phi(-\Delta) - 4s'\exp\left(-\frac{(s-s')^2}{2s}\right), \tag{3.36}$$

where the infimum is taken over all selectors $\tilde\eta$ and $\Delta > 0$ is defined in (3.34).

The proof is given in Appendix 3.6.

The next theorem is an easy consequence of Theorem 3.4.2. It describes a "phase transition" for a_d in the problem of almost full recovery.

Theorem 3.4.3. *Assume that $\limsup_{d\to\infty} s_d/d < 1/2$:*

(i) If, for all d large enough,

$$a_d^2 \ge \sigma^2\big(2\log\big((d-s_d)/s_d\big) + A_d\sqrt{2\log\big((d-s_d)/s_d\big)}\big)$$

for an arbitrary sequence $A_d \to \infty$, as $d \to \infty$, then the selector $\hat\eta$ defined by (3.8) and (3.5) achieves almost full recovery:

$$\lim_{d\to\infty}\sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d}\mathbf{E}_\theta|\hat\eta - \eta| = 0.$$

(ii) Moreover, if there exists $A > 0$ such that for all s and d large enough the reverse inequality holds:

$$a_d^2 \le \sigma^2\big(2\log\big((d-s_d)/s_d\big) + A\sqrt{2\log\big((d-s_d)/s_d\big)}\big) \tag{3.37}$$

then almost full recovery is impossible:

$$\liminf_{d\to\infty} \inf_{\tilde\eta} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d}\mathbf{E}_\theta|\tilde\eta - \eta| > 0.$$

Here, $\inf_{\tilde\eta}$ is the infimum over all selectors $\tilde\eta$.

The proof is given in Appendix 3.6.

Inspection of the proof shows that the lower bound in Theorem 3.4.3 holds true for an arbitrary $s_d \ge 5$ (possibly fixed), if (3.37) is satisfied for some A in (0,1).

Under the sparsity assumption that

$$s_d \to \infty,\ d/s_d \to \infty \qquad \text{as } d\to\infty, \tag{3.38}$$

Theorem 3.4.3 shows that the "phase transition" for almost full recovery occurs at the value $a_d = a_d^*$, where

$$a_d^* = \sigma\sqrt{2\log\big((d-s_d)/s_d\big)}\big(1+o(1)\big). \tag{3.39}$$

Furthermore, Theorem 3.4.3 details the behavior of the $o(1)$ term here.

We now state a corollary of Theorem 3.4.3 under simplified assumptions.

Corollary 3.4.1. *Assume that (3.38) holds and set*

$$a_d = \sigma\sqrt{2(1+\delta)\log(d/s_d)} \qquad \textit{for some } \delta > 0.$$

Then the selector $\hat\eta$ defined by (3.8) with threshold $t = \sigma\sqrt{2(1+\varepsilon(\delta))\log(d/s_d)}$ where $\varepsilon(\delta) > 0$ depends only on δ, achieves almost full recovery.

In the particular case of $s_d = d^{1-\beta}(1+o(1))$ for some $\beta \in (0,1)$, condition (3.38) is satisfied. Then $\log(d/s_d) = \beta(1+o(1))\log d$ and it follows from Corollary 3.4.1 that for $a_d = \sigma\sqrt{2\beta(1+\delta)\log d}$ the selector with components $\hat\eta_j = I(|X_j| > \sigma\sqrt{2\beta(1+\varepsilon)\log d})$ achieves almost full recovery. This is in agreement with the findings of Genovese et al. (2012); Ji and Jin (2012) where an analogous particular case of s_d was considered for a different model and the Bayesian definition of almost full recovery.

We now turn to the problem of exact recovery. First, notice that if

$$\limsup_{d\to\infty} s_d < \infty$$

the properties of exact recovery and almost full recovery are equivalent. Therefore, it suffices to consider exact recovery only when $s_d \to \infty$ as $d\to\infty$. Under this assumption, the "phase transition" for a_d in the problem of exact recovery is described in the next theorem.

Theorem 3.4.4. *Assume that $\lim_{d\to\infty} s_d = \infty$ and $\limsup_{d\to\infty} s_d/d < 1/2$.*

(i) If

$$a_d^2 \geq \sigma^2\bigl(2\log\bigl((d-s_d)/s_d\bigr) + W_d\bigr)$$

for all d large enough, where the sequence W_d is such that

$$\liminf_{d\to\infty} \frac{W_d}{4(\log(s_d) + \sqrt{\log(s_d)\log(d-s_d)})} \geq 1, \tag{3.40}$$

then the selector $\hat\eta$ defined by (3.8) and (3.5) achieves exact recovery:

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat\eta - \eta| = 0. \tag{3.41}$$

(ii) If the complementary condition holds,

$$a_d^2 \leq \sigma^2\bigl(2\log\bigl((d-s_d)/s_d\bigr) + W_d\bigr)$$

for all d large enough, where the sequence W_d is such that

$$\limsup_{d\to\infty} \frac{W_d}{4(\log(s_d) + \sqrt{\log(s_d)\log(d-s_d)})} < 1, \tag{3.42}$$

then exact recovery is impossible, and moreover we have

$$\lim_{d\to\infty} \inf_{\tilde\eta} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\tilde\eta - \eta| = \infty.$$

Here, $\inf_{\tilde\eta}$ is the infimum over all selectors $\tilde\eta$.

The proof is given in Appendix 3.6.

Some remarks are in order here. First of all, Theorem 3.4.4 shows that the "phase transition" for exact recovery occurs at $W_d = 4(\log(s_d) + \sqrt{\log(s_d)\log(d-s_d)})$, which corresponds to the critical value $a_d = a_d^*$ of the form

$$a_d^* = \sigma\bigl(\sqrt{2\log(d-s_d)} + \sqrt{2\log s_d}\bigr). \tag{3.43}$$

This value is greater than the critical value a_d^* for almost full recovery [cf. (3.39)], which is intuitively quite clear. The optimal threshold (3.5) corresponding to (3.43) has a simple form:

$$t_d^* = \frac{a_d^*}{2} + \frac{\sigma^2}{a_d^*}\log\left(\frac{d}{s_d} - 1\right) = \sigma\sqrt{2\log(d-s_d)}.$$

For example, if $s_d = d^{1-\beta}(1+o(1))$ for some $\beta\in(0,1)$, then $a_d^* \sim \sigma(1+\sqrt{1-\beta})\sqrt{2\log d}$. In this particular case, Theorem 3.4.4 implies that if $a_d = \sigma(1+\sqrt{1-\beta})\sqrt{2(1+\delta)\log d}$ for some $\delta > 0$, then exact recovery is possible and the selector with threshold $t = \sigma\sqrt{2(1+\varepsilon)\log d}$ for some $\varepsilon > 0$ achieves exact recovery. This is in agreement with the results of Genovese et al. (2012); Ji and Jin (2012) where an analogous particular case of s_d was considered for a different model and the Bayesian definition of exact recovery. For our model, even a sharper result is true; namely, a simple universal threshold $t = \sigma\sqrt{2\log d}$ guarantees exact recovery adaptively in the parameters a and s. Intuitively, this is suggested by the form of t_d^*. The precise statement is given in Theorem 3.5.1 below.

Finally, we state an asymptotic corollary of Theorem 3.2.6 showing that the selector $\hat\eta$ considered above is sharp in the asymptotically minimax sense with respect to the risk defined as the probability of wrong recovery.

Theorem 3.4.5. *Assume that exact recovery is possible for the classes* $(\Theta_d(s_d, a_d))_{d\ge1}$ *and* $(\Theta_d^+(s_d, a_d))_{d\ge1}$, *that is, condition* (3.30) *holds. Then, for the selectors* $\hat\eta$ *and* $\hat\eta^+$ *defined by* (3.8), (3.4) *and* (3.5), *and for the selector* $\bar\eta$ *defined by* (3.12) *and* (3.13), *we have*

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d^+(s_d,a_d)} \frac{\mathbf{P}_\theta(S_{\hat\eta^+} \ne S(\theta))}{s_d\Psi_+(d, s_d, a_d)} = \lim_{d\to\infty} \inf_{\tilde\eta\in\mathcal{T}} \sup_{\theta\in\Theta_d^+(s_d,a_d)} \frac{\mathbf{P}_\theta(S_{\tilde\eta} \ne S(\theta))}{s_d\Psi_+(d, s_d, a_d)} = 1,$$

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{\mathbf{P}_\theta(S_{\bar\eta} \ne S(\theta))}{s_d\bar\Psi(d, s_d, a_d)} = \lim_{d\to\infty} \inf_{\tilde\eta\in\mathcal{T}} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{\mathbf{P}_\theta(S_{\tilde\eta} \ne S(\theta))}{s_d\bar\Psi(d, s_d, a_d)} = 1$$

and

$$\limsup_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{\mathbf{P}_\theta(S_{\hat\eta} \ne S(\theta))}{s_d\Psi_+(d, s_d, a_d)} \le 2.$$

Note that the threshold (3.5) depends on the parameters s and a, so that the selectors considered in all the results above are not adaptive. In the next section, we propose adaptive selectors that achieve almost full recovery and exact recovery without the knowledge of s and a.

Remark 3.4.1. *Another procedure of variable selection is the exhaustive search estimator of the support* $S(\theta)$ *defined as*

$$\tilde S = \operatorname*{argmax}_{C\subseteq\{1,\dots,d\}:|C|=s} \sum_{j\in C} X_j.$$

This estimator was studied by Butucea et al. Butucea et al. (2015). *The selection procedure can be equivalently stated as choosing the indices* j *corresponding to* s *largest order statistics of the sample* $(X_1, \dots, X_d)$. *In* Butucea et al. (2015), *Theorem 2.5, it was shown that, on the class* $\Theta_d^+(s_d, a_d)$, *the probability of wrong recovery* $\mathbf{P}_\theta(\tilde S \ne S(\theta))$ *tends to 0 as* $d \to \infty$ *under a stronger condition on* (s_d, a_d) *than* (3.30). *The rate of this convergence was not analyzed there. If we denote by* $\eta_{\tilde S}$ *the selector with components* $I(j \in \tilde S)$ *for* j *from 1 to* d, *it can be proved that* $\mathbf{E}|\eta_{\tilde S} - \eta| \le 2\mathbf{E}|\hat\eta^+ - \eta|$, *and thus the risk of* $\eta_{\tilde S}$ *is within at least a factor 2 of the minimax risk over the class* $\Theta_d^+(s, a)$.

3.5 Adaptive selectors

In this section, we consider the asymptotic setup as in Section 3.4 and construct the selectors that provide almost full and exact recovery adaptively, that is, without the knowledge of a and s.

As discussed in Section 3.4, the issue of adaptation for exact recovery is almost trivial. Indeed, the expressions for minimal value a_d^*, for which exact recovery is possible [cf. (3.43)], and for the corresponding optimal threshold t_d^* suggest that taking a selector with the universal threshold $t = \sigma\sqrt{2\log d}$ is enough to achieve exact recovery simultaneously for all values (a_d, s_d), for which the exact recovery is possible. This point is formalized in the next theorem.

Theorem 3.5.1. *Assume that* $s_d \to \infty$ *as* $d \to \infty$ *and that* $\limsup_{d\to\infty} s_d/d < 1/2$. *Let the sequence* $(a_d)_{d\ge1}$ *be above the phase transition level for exact recovery, that is,* $a_d \ge a_d^*$ *for all* d, *where* a_d^* *is defined in* (3.43). *Then the selector* $\hat\eta$ *defined by* (3.8) *with threshold* $t = \sigma\sqrt{2\log d}$ *achieves exact recovery.*

The proof of this theorem is given in Appendix 3.6.

We now turn to the problem of adaption for almost full recovery. Ideally, we would like to construct a selector that achieves almost full recovery for all sequences $(s_d, a_d)_{d\geq 1}$ for which almost full recovery is possible. We have seen in Section 3.4 that this includes a much broader range of values than in case of exact recovery. Thus, using the adaptive selector of Theorem 3.5.1 for almost full recovery does not give a satisfactory result, and we have to take a different approach.

Following Section 3.4, we will use the notation

$$a_0(s, A) \triangleq \sigma\big(2\log\big((d-s)/s\big) + A\sqrt{\log\big((d-s)/s\big)}\big)^{1/2}.$$

As shown in Section 3.4, it makes sense to consider the classes $\Theta_d(s, a)$ only when $a \geq a_0(s, A)$ with some $A > 0$, since for other values of a almost full recovery is impossible. Only such classes will be studied below.

In the asymptotic setup of Section 3.4, we have used the assumption that $d/s_d \to \infty$ (the sparsity assumption), which is now transformed into the condition

$$\begin{aligned} &s_d \in \mathcal{S}_d \triangleq \{1, 2, \ldots, s_d^*\} \\ &\qquad \text{where } s_d^* \text{ is an integer such that } \frac{d}{s_d^*} \to \infty \text{ as } d \to \infty. \end{aligned} \tag{3.44}$$

Assuming s_d to be known, we have shown in Section 3.4 that almost full recovery is achievable for all $a \geq a_0(s_d, A_d)$, where A_d tends to infinity as $d \to \infty$. The rate of growth of A_d was allowed to be arbitrarily slow there; cf. Theorem 3.4.3. However, for adaptive estimation considered in this section we will need the following mild assumption on the growth of A_d:

$$A_d \geq c_0\left(\log\log\left(\frac{d}{s_d^*} - 1\right)\right)^{1/2}, \tag{3.45}$$

where $c_0 > 0$ is an absolute constant. In what follows, we will assume that $s_d^* \leq d/4$, so that the right-hand side of (3.45) is well defined.

Consider a grid of points $\{g_1, \ldots, g_M\}$ on $\mathcal{S}_d$, where $g_j = 2^{j-1}$ and M is the maximal integer such that $g_M \leq s_d^*$. For each g_m, $m = 1, \ldots, M$, we define a selector:

$$\hat{\eta}(g_m) = \big(\hat{\eta}_j(g_m)\big)_{j=1,\ldots,d} \triangleq \big(I\big(|X_j| \geq w(g_m)\big)\big)_{j=1,\ldots,d},$$

where

$$w(s) = \sigma\sqrt{2\log\left(\frac{d}{s} - 1\right)}.$$

Note that $w(s)$ is monotonically decreasing. We now choose the "best" index m, for which g_m is near the true (but unknown) value of s, by the following data-driven procedure:

$$\begin{aligned} \hat{m} = \min\Bigg\{ &m \in \{2, \ldots, M\} : \\ &\sum_{j=1}^{d} I\big(w(g_k) \leq |X_j| < w(g_{k-1})\big) \leq \tau g_k \text{ for all } k \geq m+1\Bigg\}, \end{aligned} \tag{3.46}$$

where

$$\tau = \big(\log\big(d/s_d^* - 1\big)\big)^{-\frac{1}{7}},$$

and we set $\hat{m} = M$ if the set in (3.46) is empty. Finally, we define an adaptive selector as

$$\hat{\eta}^{\mathrm{ad}} = \hat{\eta}(g_{\hat{m}}).$$

This adaptive procedure is quite natural in the sense that it can be related to the Lepski method or to wavelet thresholding that are widely used for adaptive estimation. Indeed, as in wavelet methods, we consider dyadic blocks determined by the grid points g_j. The value $\sum_{j=1}^{d} I(w(g_k) \le |X_j| < w(g_{k-1}))$ is the number of observations within the kth block. If this number is too small (below a suitably chosen threshold), we decide that the block corresponds to pure noise and it is rejected; in other words, this k is not considered as a good candidate for $\hat{m}$. This argument is analogous to wavelet thersholding. We start from the largest k [equivalently, smallest $w(g_k)$] and perform this procedure until we find the first block, which is not rejected. The corresponding value k determines our choice of $\hat{m}$ as defined in (3.46).

Theorem 3.5.2. *Let $c_0 \ge 16$. Then the selector $\hat{\eta}^{\mathrm{ad}}$ adaptively achieves almost full recovery in the following sense:*

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d} \mathbf{E}_\theta \big|\hat{\eta}^{\mathrm{ad}} - \eta\big| = 0 \tag{3.47}$$

for all sequences $(s_d, a_d)_{d\ge1}$ such that (3.44) holds and $a_d \ge a_0(s_d, A_d)$, where A_d satisfies (3.45).

Remark 3.5.1. *Another family of variable selection methods originates from the theory of multiple testing Abramovich and Benjamini (1995); Abramovich et al. (2006). These are, for example, the Benjamini–Hochberg, Benjamini–Yekutieli or SLOPE procedures. We refer to Bogdan et al. (2015) for a recent overview and comparison of these techniques. They have the same structure as the exhaustive search procedure in that they keep only the largest order statistics. The difference is that the value s (which is usually not known in practice) is replaced by an estimator $\hat{s}$ obtained from comparing the ith order statistic of $(|X_1|, \dots, |X_d|)$ with a suitable normal quantile depending on i. The analysis of these methods in the literature is focused on the evaluation of false discovery rate (FDR). Asymptotic power calculations for the Benjamini–Hochberg procedure are given in Arias-Castro and Chen (2017). To the best of our knowledge, the behavior of the risk $\mathbf{P}_\theta(\hat{S} \ne S(\theta))$ and of the Hamming risk, even in a simple consistency perspective, was not studied.*

Remark 3.5.2. *In this chapter, the variance σ was supposed to be known. Extension to the case of unknown σ can be treated as described, for example, in Collier et al. (2018). Namely, we replace σ in the definition of the threshold $w(s)$ by a statistic $\hat{\sigma}$ defined in Collier et al. (2018), Section 3. As shown in Collier et al. (2018), Proposition 1, this statistic is such that $\sigma \le \hat{\sigma} \le C'\sigma$ with high probability provided that $s \le d/2$, and $d \ge d_0$ for some absolute constants $C' > 1, d_0 \ge 1$. Then, replacing σ by $\hat{\sigma}$ in the expression for $w(s)$, one can show that Theorem 3.5.2 remains valid with this choice of $w(s)$ independent of σ, up to a change in numerical constants in the definition of the adaptive procedure. With this modification, we obtain a procedure which is completely*

data-driven and enjoys the property of almost full recovery under the mild conditions given in Theorem 3.5.2. The same modification can be done in Theorem 3.5.1. Namely, under the assumptions of Theorem 3.5.1 and $a_d \geq c' a_d^*$, where $c' \geq 1$ is a numerical constant, the selector $\hat{\eta}$ defined by (3.8) with threshold $t = \hat{\sigma}\sqrt{2\log d}$ achieves exact recovery when σ is unknown.

Remark 3.5.3. *In this section, the problem of adaptive variable selection was considered only for the classes $\Theta_d(s_d, a_d)$. The corresponding results for classes $\Theta_d^+(s_d, a_d)$ and $\Theta_d^-(s_d, a_d)$ are completely analogous. We do not state them here for the sake of brevity.*

3.6 Appendix: Main proofs

Proof of Theorem 3.2.3. We have, for any $t > 0$,

$$\begin{aligned} |\hat{\eta} - \eta| &= \sum_{j:\eta_j=0} \hat{\eta}_j + \sum_{j:\eta_j=1} (1 - \hat{\eta}_j) \\ &= \sum_{j:\eta_j=0} I\big(|\sigma\xi_j| \geq t\big) + \sum_{j:\eta_j=1} I\big(|\sigma\xi_j + \theta_j| < t\big). \end{aligned}$$

Now, for any $\theta \in \Theta_d(s, a)$ and any $t > 0$,

$$\begin{aligned} \mathbf{E}\big(I\big(|\sigma\xi_j + \theta_j| < t\big)\big) &\leq \mathbf{P}\big(|\theta_j| - |\sigma\xi_j| < t\big) \leq \mathbf{P}\big(|\xi| > (a-t)/\sigma\big) \\ &= \mathbf{P}\big(|\xi| > (a-t)_+/\sigma\big), \end{aligned}$$

where ξ denotes a standard Gaussian random variable. Thus, for any $\theta \in \Theta_d(s, a)$,

$$\frac{1}{s}\mathbf{E}_\theta|\hat{\eta} - \eta| \leq \frac{d - |S|}{s}\mathbf{P}\left(|\xi| \geq \frac{t}{\sigma}\right) + \frac{|S|}{s}\mathbf{P}\left(|\xi| > \frac{(a-t)_+}{\sigma}\right) \leq 2\Psi(d, s, a). \tag{3.48}$$

Indeed, for t defined in (3.5), $t \geq (a-t)_+$ given that $s \leq d/2$. Here and in the sequel, $|S|$ denotes the cardinality of $S = S(\theta)$. □

Proof of Theorem 3.2.1. Arguing as in the proof of Theorem 3.2.3, we obtain

$$\big|\hat{\eta}^+ - \eta\big| = \sum_{j:\eta_j=0} I(\xi_j \geq t) + \sum_{j:\eta_j=1} I(\sigma\xi_j + \theta_j < t),$$

and $\mathbf{E}(I(\sigma\xi_j + \theta_j < t)) \leq \mathbf{P}(\xi < (t-a)/\sigma)$. Thus, for any $\theta \in \Theta_d^+(s, a)$,

$$\frac{1}{s}\mathbf{E}_\theta\big|\hat{\eta}^+ - \eta\big| \leq \frac{d - |S|}{s}\mathbf{P}(\xi \geq t/\sigma) + \frac{|S|}{s}\mathbf{P}\big(\xi < (t-a)/\sigma\big) \leq \Psi_+(d, s, a),$$

by the monotonicity of Φ and the condition $s \leq d/2$. □

Proof of Theorem 3.2.2. We prove here the first inequality of Theorem 3.2.2. Since $\tilde{\eta}_j$ depends only on X_j,

$$\mathbf{E}_\theta|\tilde{\eta} - \eta| = \sum_{j=1}^{d} \mathbf{E}_{j,\theta_j}|\tilde{\eta}_j - \eta_j|, \tag{3.49}$$

where $\mathbf{E}_{j,\theta_j}$ is the expectation with respect to the distribution of X_j.

Let Θ' be the set of all θ in $\Theta_d^+(s,a)$ such that s components θ_j of θ are equal to a and the remaining $d-s$ components are 0. Denote by $|\Theta'| = \binom{d}{s}$ the cardinality of Θ'. Then, for any $\tilde{\eta} \in \mathcal{T}$ we have

$$\begin{aligned}
&\sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta|\tilde{\eta}-\eta| \\
&\quad\geq \frac{1}{s|\Theta'|}\sum_{\theta\in\Theta'}\sum_{j=1}^{d}\mathbf{E}_{j,\theta_j}|\tilde{\eta}_j-\eta_j| \\
&\quad= \frac{1}{s|\Theta'|}\sum_{j=1}^{d}\Bigg(\sum_{\theta\in\Theta':\theta_j=0}\mathbf{E}_{j,0}(\tilde{\eta}_j)+\sum_{\theta\in\Theta':\theta_j=a}\mathbf{E}_{j,a}(1-\tilde{\eta}_j)\Bigg) \qquad (3.50)\\
&\quad= \frac{1}{s}\sum_{j=1}^{d}\Bigg(\Big(1-\frac{s}{d}\Big)\mathbf{E}_{j,0}(\tilde{\eta}_j)+\frac{s}{d}\mathbf{E}_{j,a}(1-\tilde{\eta}_j)\Bigg) \\
&\quad\geq \frac{d}{s}\inf_{T\in[0,1]}\Bigg(\Big(1-\frac{s}{d}\Big)\mathbb{E}_0(T)+\frac{s}{d}\mathbb{E}_a(1-T)\Bigg),
\end{aligned}$$

where we have used that $|\{\theta\in\Theta' : \theta_j = a\}| = \binom{d-1}{s-1} = s|\Theta'|/d$. In the last line of display (3.50), $\mathbb{E}_u$ is understood as the expectation with respect to the distribution of $X = u+\sigma\xi$, where $\xi\sim\mathcal{N}(0,1)$ and $\inf_{T\in[0,1]}$ denotes the infimum over all $[0,1]$-valued statistics $T(X)$. Set

$$L^* = \inf_{T\in[0,1]}\Bigg(\Big(1-\frac{s}{d}\Big)\mathbb{E}_0(T)+\frac{s}{d}\mathbb{E}_a(1-T)\Bigg)$$

By the Bayesian version of the Neyman–Pearson lemma, the infimum here is attained for $T = T^*$ given by

$$T^*(X) = I\Bigg(\frac{(s/d)\varphi_\sigma(X-a)}{(1-s/d)\varphi_\sigma(X)} > 1\Bigg)$$

where $\varphi_\sigma(\cdot)$ is the density of an $\mathcal{N}(0,\sigma^2)$ distribution. Thus,

$$L^* \;=\; \Big(1-\frac{s}{d}\Big)\mathbf{P}\Bigg(\frac{\varphi_\sigma(\sigma\xi-a)}{\varphi_\sigma(\sigma\xi)} > \frac{d}{s}-1\Bigg)+\frac{s}{d}\mathbf{P}\Bigg(\frac{\varphi_\sigma(\sigma\xi)}{\varphi_\sigma(\sigma\xi+a)} \leq \frac{d}{s}-1\Bigg).$$

Combining this with (3.49) and (3.50), we get

$$\begin{aligned}
&\inf_{\tilde{\eta}}\sup_{\theta\in\Theta_d^+(s,a)}\frac{1}{s}\mathbf{E}_\theta|\tilde{\eta}-\eta| \\
&\quad\geq \Big(\frac{d}{s}-1\Big)\mathbf{P}\Bigg(\exp\Big(\frac{a\xi}{\sigma}-\frac{a^2}{2\sigma^2}\Big) > \frac{d}{s}-1\Bigg) \\
&\qquad+\mathbf{P}\Bigg(\exp\Big(\frac{a\xi}{\sigma}+\frac{a^2}{2\sigma^2}\Big) \leq \frac{d}{s}-1\Bigg) \\
&\quad= \Big(\frac{d}{s}-1\Big)\mathbf{P}\Bigg(\xi > \frac{a}{2\sigma}+\frac{\sigma}{a}\log\Big(\frac{d}{s}-1\Big)\Bigg) \\
&\qquad+\mathbf{P}\Bigg(\xi \leq -\frac{a}{2\sigma}+\frac{\sigma}{a}\log\Big(\frac{d}{s}-1\Big)\Bigg) \\
&\quad= \Psi_+(d,s,a).
\end{aligned}$$

□

Proof of Theorem 3.2.4. For any $\theta \in \Theta_d(s,a)$, we have

$$\begin{aligned}\mathbf{E}_\theta|\bar\eta-\eta| &= \sum_{j:\theta_j=0}\mathbf{P}_{j,0}(\bar\eta_j=1)+\sum_{j:\theta_j\ge a}\mathbf{P}_{j,\theta_j}(\bar\eta_j=0)\\ &\quad+\sum_{j:\theta_j\le -a}\mathbf{P}_{j,\theta_j}(\bar\eta_j=0)\\ &\le d\mathbf{P}\left(e^{-\frac{a^2}{2\sigma^2}}\cosh\left(\frac{a\xi}{\sigma}\right)>\frac{d}{s}-1\right)\\ &\quad+\sum_{j:\theta_j\ge a}\mathbf{P}_{j,\theta_j}(\bar\eta_j=0)+\sum_{j:\theta_j\le -a}\mathbf{P}_{j,\theta_j}(\bar\eta_j=0),\end{aligned} \tag{3.51}$$

where $\mathbf{P}_{j,\theta_j}$ denotes the distribution of X_j, and ξ is a standard Gaussian random variable. We now bound from above the probabilities $\mathbf{P}_{j,\theta_j}(\bar\eta_j=0)$. Introduce the notation

$$g(x)=\cosh\left(\frac{(x+\sigma\xi)a}{\sigma^2}\right)\qquad \forall x\in\mathbb{R},$$

and

$$u=\exp\left(\frac{a^2}{2\sigma^2}+\log\left(\frac{d}{s}-1\right)\right).$$

We have

$$\mathbf{P}_{j,\theta_j}(\bar\eta_j=0)=\mathbf{P}\big(g(\theta_j)<u\big)=\mathbf{P}(-b-\theta_j<\sigma\xi<b-\theta_j),$$

where $b=(\sigma^2/a)\operatorname{arccosh}(u)>0$. It is easy to check that the function $x\mapsto\mathbf{P}(-b-x<\sigma\xi<b-x)$ is monotonically decreasing on $[0,\infty)$. Therefore, the maximum of $\mathbf{P}(-b-\theta_j<\sigma\xi<b-\theta_j)$ over $\theta_j\ge a$ is attained at $\theta_j=a$. Thus, for any $\theta_j\ge a$ we have

$$\mathbf{P}_{j,\theta_j}(\bar\eta_j=0)\le\mathbf{P}\big(g(a)<u\big)=\mathbf{P}\left(e^{-\frac{a^2}{2\sigma^2}}\cosh\left(\frac{(a+\sigma\xi)a}{\sigma^2}\right)<\frac{d}{s}-1\right). \tag{3.52}$$

Analogously, for any $\theta_j\le -a$,

$$\begin{aligned}\mathbf{P}_{j,\theta_j}(\bar\eta_j=0)&\le\mathbf{P}\left(e^{-\frac{a^2}{2\sigma^2}}\cosh\left(\frac{(-a+\sigma\xi)a}{\sigma^2}\right)<\frac{d}{s}-1\right)\\ &=\mathbf{P}\left(e^{-\frac{a^2}{2\sigma^2}}\cosh\left(\frac{(a+\sigma\xi)a}{\sigma^2}\right)<\frac{d}{s}-1\right),\end{aligned} \tag{3.53}$$

where the last equality follows from the fact that ξ has the same distribution as $-\xi$ and cosh is an even function. Combining (3.51)–(3.53) proves the theorem. □

Proof of Theorem 3.2.5. We follow the lines of the proof of Theorem 3.2.2 with suitable modifications.

Let Θ^+ and Θ^- be the sets of all θ in $\Theta_d(s,a)$ such that $d-s$ components θ_j of θ are equal to 0 and the remaining s components are equal to a (for $\theta\in\Theta^+$) or to $-a$ (for $\theta\in\Theta^-$). For any $\tilde\eta\in\mathcal{T}$, we have

$$\begin{aligned}&\sup_{\theta\in\Theta_d(s,a)}\sum_{j=1}^d\mathbf{E}_{j,\theta_j}|\tilde\eta_j-\eta_j|\\ &\qquad\ge\frac{1}{2}\left\{\sup_{\theta\in\Theta^+}\sum_{j=1}^d\mathbf{E}_{j,\theta_j}|\tilde\eta_j-\eta_j|+\sup_{\theta\in\Theta^-}\sum_{j=1}^d\mathbf{E}_{j,\theta_j}|\tilde\eta_j-\eta_j|\right\}.\end{aligned}$$

As shown in the proof of Theorem 3.2.2, for any $\tilde{\eta} \in \mathcal{T}$,

$$\sup_{\theta\in\Theta^+} \sum_{j=1}^d \mathbf{E}_{j,\theta_j}|\tilde{\eta}_j - \eta_j| \;\geq\; \sum_{j=1}^d \biggl(\Bigl(1 - \frac{s}{d}\Bigr)\mathbf{E}_{j,0}(\tilde{\eta}_j) + \frac{s}{d}\mathbf{E}_{j,a}(1-\tilde{\eta}_j)\biggr).$$

Analogously,

$$\sup_{\theta\in\Theta^-} \sum_{j=1}^d \mathbf{E}_{j,\theta_j}|\tilde{\eta}_j - \eta_j| \;\geq\; \sum_{j=1}^d \biggl(\Bigl(1 - \frac{s}{d}\Bigr)\mathbf{E}_{j,0}(\tilde{\eta}_j) + \frac{s}{d}\mathbf{E}_{j,-a}(1-\tilde{\eta}_j)\biggr).$$

From the last three displays, we obtain

$$\sup_{\theta\in\Theta_d(s,a)} \sum_{j=1}^d \mathbf{E}_{j,\theta_j}|\tilde{\eta}_j - \eta_j| \;\geq\; \sum_{j=1}^d \biggl(\Bigl(1 - \frac{s}{d}\Bigr)\mathbf{E}_{j,0}(\tilde{\eta}_j) + \frac{s}{d}\bar{\mathbf{E}}_{j}(1-\tilde{\eta}_j)\biggr),$$

where $\bar{\mathbf{E}}_j$ is the expectation with respect to the measure $\bar{\mathbf{P}}_j = (\mathbf{P}_{j,a} + \mathbf{P}_{j,-a})/2$. It follows that

$$\sup_{\theta\in\Theta_d(s,a)} \sum_{j=1}^d \mathbf{E}_{j,\theta_j}|\tilde{\eta}_j - \eta_j| \;\geq\; \inf_{T\in[0,1]} \bigl((d-s)\mathbb{E}_0(T) + s\bar{\mathbb{E}}(1-T)\bigr). \tag{3.54}$$

Here, $\mathbb{E}_0$ denotes the expectation with respect to the distribution of X with density $\varphi_\sigma(\cdot)$, $\bar{\mathbb{E}}$ is the expectation with respect to the distribution of X with mixture density $\bar{\varphi}_\sigma(\cdot) = (\varphi_\sigma(\cdot + a) + \varphi_\sigma(\cdot - a))/2$, and $\inf_{T\in[0,1]}$ denotes the infimum over all $[0,1]$-valued statistics $T(X)$. Recall that we denote by $\varphi_\sigma(\cdot)$ is the density of $\mathcal{N}(0,\sigma^2)$ distribution. Set

$$\tilde{L} = \inf_{T\in[0,1]} \biggl(\Bigl(1 - \frac{s}{d}\Bigr)\mathbb{E}_0(T) + \frac{s}{d}\bar{\mathbb{E}}(1-T)\biggr).$$

By the Bayesian version of the Neyman–Pearson lemma, the infimum here is attained for $T = \tilde{T}$ given by

$$\tilde{T}(X) = I\biggl(\frac{(s/d)\bar{\varphi}_\sigma(X)}{(1-s/d)\varphi_\sigma(X)} > 1\biggr).$$

Thus,

$$\begin{aligned}
\tilde{L} &= \Bigl(1 - \frac{s}{d}\Bigr)\mathbf{P}\biggl(\frac{\bar{\varphi}_\sigma(\sigma\xi)}{\varphi_\sigma(\sigma\xi)} > \frac{d}{s} - 1\biggr) + \frac{s}{2d}\mathbb{P}_a\biggl(\frac{\bar{\varphi}_\sigma(X)}{\varphi_\sigma(X)} \leq \frac{d}{s} - 1\biggr) \\
&\quad + \frac{s}{2d}\mathbb{P}_{-a}\biggl(\frac{\bar{\varphi}_\sigma(X)}{\varphi_\sigma(X)} \leq \frac{d}{s} - 1\biggr) \\
&= \Bigl(1 - \frac{s}{d}\Bigr)\mathbf{P}\biggl(e^{-\frac{a^2}{2\sigma^2}}\cosh\Bigl(\frac{a\xi}{\sigma}\Bigr) > \frac{d}{s} - 1\biggr) \\
&\quad + \frac{s}{2d}\mathbb{P}_a\biggl(\frac{\bar{\varphi}_\sigma(X)}{\varphi_\sigma(X)} \leq \frac{d}{s} - 1\biggr) + \frac{s}{2d}\mathbb{P}_{-a}\biggl(\frac{\bar{\varphi}_\sigma(X)}{\varphi_\sigma(X)} \leq \frac{d}{s} - 1\biggr),
\end{aligned} \tag{3.55}$$

where $\mathbb{P}_u$ denotes the probability distribution of X with density $\varphi_\sigma(\cdot - u)$. Note that, for all $x \in \mathbb{R}$,

$$\frac{\bar{\varphi}_\sigma(x)}{\varphi_\sigma(x)} = e^{-\frac{a^2}{2\sigma^2}}\cosh\Bigl(\frac{ax}{\sigma^2}\Bigr).$$

Using this formula with $x = \sigma\xi + a$ and $x = \sigma\xi - a$, and the facts that $\cosh(\cdot)$ is an even function and ξ coincides with $-\xi$ in distribution, we obtain

$$\begin{aligned}\mathbb{P}_a\left(\frac{\bar{\varphi}_\sigma(X)}{\varphi_\sigma(X)} \le \frac{d}{s} - 1\right) &= \mathbb{P}_{-a}\left(\frac{\bar{\varphi}_\sigma(X)}{\varphi_\sigma(X)} \le \frac{d}{s} - 1\right) \\ &= \mathbf{P}\left(e^{-\frac{a^2}{2\sigma^2}} \cosh\left(\frac{a\xi}{\sigma} + \frac{a^2}{\sigma^2}\right) \le \frac{d}{s} - 1\right).\end{aligned}$$

Thus, $\tilde{L} = (s/d)\bar{\Psi}(d, s, a)$. Combining this equality with (3.49) and (3.54) proves the theorem. □

Proof of Theorem 3.2.6. The upper bounds (3.14), (3.15) and (3.16) follow immediately from (3.2) and Theorems 3.2.1, 3.2.4 and 3.2.3, respectively. We now prove the lower bound (3.17). To this end, first note that for any $\theta \in \Theta_d^+(s, a)$ and any $\tilde{\eta} \in \mathcal{T}$ we have

$$\mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) = \mathbf{P}_\theta\left(\bigcup_{j=1}^d \{\tilde{\eta}_j \neq \eta_j\}\right) = 1 - \prod_{j=1}^d p_j(\theta),$$

where $p_j(\theta) \triangleq \mathbf{P}_\theta(\tilde{\eta}_j = \eta_j)$. Hence, for any $\tilde{\eta} \in \mathcal{T}$,

$$\sup_{\theta \in \Theta_d^+(s,a)} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) \ge \max_{\theta \in \Theta'} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) = 1 - p_*, \tag{3.56}$$

where Θ' is the subset of $\Theta_d^+(s, a)$ defined in the proof of Theorem 3.2.2, and $p_* = \min_{\theta \in \Theta'} \prod_{j=1}^d p_j(\theta)$.

Next, for any selector $\tilde{\eta}$ we have $\mathbf{P}_\theta(S_{\tilde{\eta}} \neq S(\theta)) \ge \mathbf{P}_\theta(|\tilde{\eta} - \eta| = 1)$. Therefore,

$$\sup_{\theta \in \Theta_d^+(s,a)} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) \;\ge\; \frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \mathbf{P}_\theta\big(|\tilde{\eta} - \eta| = 1\big). \tag{3.57}$$

Here, $\mathbf{P}_\theta(|\tilde{\eta} - \eta| = 1) = \mathbf{P}_\theta(\bigcup_{j=1}^d B_j)$ with the random events $B_j = \{|\tilde{\eta}_j - \eta_j| = 1, \text{ and } \tilde{\eta}_i = \eta_i, \forall i \neq j\}$. Since the events B_j are disjoint, for any $\tilde{\eta} \in \mathcal{T}$ we get

$$\begin{aligned}&\frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \mathbf{P}_\theta\big(|\tilde{\eta} - \eta| = 1\big) \\ &\quad= \frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \sum_{j=1}^d \mathbf{P}_\theta(B_j) \\ &\quad= \frac{1}{|\Theta'|} \sum_{j=1}^d \Bigg(\sum_{\theta \in \Theta' : \theta_j = 0} \mathbf{P}_{j,0}(\tilde{\eta}_j = 1) \prod_{i \neq j} p_i(\theta) \\ &\qquad+ \sum_{\theta \in \Theta' : \theta_j = a} \mathbf{P}_{j,a}(\tilde{\eta}_j = 0) \prod_{i \neq j} p_i(\theta) \Bigg) \\ &\quad\ge \frac{p_*}{|\Theta'|} \sum_{j=1}^d \Bigg(\sum_{\theta \in \Theta' : \theta_j = 0} \mathbf{P}_{j,0}(\tilde{\eta}_j = 1) + \sum_{\theta \in \Theta' : \theta_j = a} \mathbf{P}_{j,a}(\tilde{\eta}_j = 0) \Bigg) \\ &\quad= \frac{p_*}{|\Theta'|} \sum_{j=1}^d \Bigg(\sum_{\theta \in \Theta' : \theta_j = 0} \mathbf{E}_{j,0}(\tilde{\eta}_j) + \sum_{\theta \in \Theta' : \theta_j = a} \mathbf{E}_{j,a}(1 - \tilde{\eta}_j) \Bigg),\end{aligned} \tag{3.58}$$

where $\mathbf{P}_{j,u}$ denotes the distribution of X_j when $\theta_j = u$. We now bound the right-hand side of (3.58) by following the argument from the last three lines of (3.50) to the end of the proof of Theorem 3.2.2. Applying this argument yields that, for any $\tilde{\eta} \in \mathcal{T}$,

$$\frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \mathbf{P}_\theta\big(|\tilde{\eta} - \eta| = 1\big) \quad \geq \quad p^* d\tilde{L} \geq p^* s\Psi_+(d, s, a). \tag{3.59}$$

Combining (3.56), (3.57) and (3.59), we find that, for any $\tilde{\eta} \in \mathcal{T}$,

$$\begin{aligned} \sup_{\theta \in \Theta_d^+(s,a)} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) &\geq \min_{0 \leq p^* \leq 1} \max\big\{1 - p^*, p^* s\Psi_+(d, s, a)\big\} \\ &= \frac{s\Psi_+(d, s, a)}{1 + s\Psi_+(d, s, a)}. \end{aligned}$$

We now prove the lower bound (3.18). Let the sets Θ^+ and Θ^- and the constants $p_j(\theta)$ be the same as in the proof of Theorem 3.2.5. Then

$$\sup_{\theta \in \Theta_d(s,a)} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) \geq \max_{\theta \in \Theta^+ \cup \Theta^-} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) = 1 - \bar{p},$$

where $\bar{p} = \min_{\theta \in \Theta^+ \cup \Theta^-} \prod_{j=1}^d p_j(\theta)$.

For any selector $\tilde{\eta}$, we use that $\mathbf{P}_\theta(S_{\tilde{\eta}} \neq S(\theta)) \geq \mathbf{P}_\theta(|\tilde{\eta} - \eta| = 1)$ and, therefore,

$$\begin{aligned} \sup_{\theta \in \Theta_d(s,a)} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) \geq \frac{1}{2|\Theta^+|} \sum_{\theta \in \Theta^+} \mathbf{P}_\theta\big(|\tilde{\eta} - \eta| = 1\big) \\ + \frac{1}{2|\Theta^-|} \sum_{\theta \in \Theta^-} \mathbf{P}_\theta\big(|\tilde{\eta} - \eta| = 1\big). \end{aligned}$$

We continue along the same lines as in the proof of (3.58) to get, for any separable selector $\tilde{\eta}$,

$$\begin{aligned} &\sup_{\theta \in \Theta_d(s,a)} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) \\ &\geq \frac{\bar{p}}{2|\Theta^+|} \sum_{j=1}^d \Bigg(\sum_{\theta \in \Theta^+ : \theta_j = 0} \mathbf{E}_{j,0}(\tilde{\eta}_j) + \sum_{\theta \in \Theta^+ : \theta_j = a} \mathbf{E}_{j,a}(1 - \tilde{\eta}_j) \Bigg) \\ &\quad + \frac{\bar{p}}{2|\Theta^-|} \sum_{j=1}^d \Bigg(\sum_{\theta \in \Theta^- : \theta_j = 0} \mathbf{E}_{j,0}(\tilde{\eta}_j) + \sum_{\theta \in \Theta^- : \theta_j = -a} \mathbf{E}_{j,-a}(1 - \tilde{\eta}_j) \Bigg) \\ &\geq \frac{\bar{p}}{2} \sum_{j=1}^d \Bigg(\Big(1 - \frac{s}{d}\Big) \mathbf{E}_{j,0}(\tilde{\eta}_j) + \frac{s}{d} \mathbf{E}_{j,a}(1 - \tilde{\eta}_j) \Bigg) \\ &\quad + \frac{\bar{p}}{2} \sum_{j=1}^d \Bigg(\Big(1 - \frac{s}{d}\Big) \mathbf{E}_{j,0}(\tilde{\eta}_j) + \frac{s}{d} \mathbf{E}_{j,-a}(1 - \tilde{\eta}_j) \Bigg) \\ &= \bar{p} \sum_{j=1}^d \Bigg(\Big(1 - \frac{s}{d}\Big) \mathbf{E}_{j,0}(\bar{\eta}_j) + \frac{s}{d} \bar{\mathbf{E}}_j(1 - \tilde{\eta}_j) \Bigg), \end{aligned}$$

where again $\bar{\mathbf{E}}_j$ denotes the expected value with respect to $\bar{\mathbf{P}}_j = \frac{1}{2}(\mathbf{P}_{j,a} + \mathbf{P}_{j,-a})$. Analogously to the proof of Theorem 3.2.5, the expression in the last display can be further bounded from below by $\bar{p}d\tilde{L} = \bar{p}s\bar{\Psi}(d,s,a)$. Thus,

$$\sup_{\theta\in\Theta_d(s,a)} \mathbf{P}_\theta\big(S_{\tilde{\eta}} \neq S(\theta)\big) \geq \min_{0\leq\bar{p}\leq 1} \max\big\{1-\bar{p}, \bar{p}s\bar{\Psi}(d,s,a)\big\} = \frac{s\bar{\Psi}(d,s,a)}{1+s\bar{\Psi}(d,s,a)}.$$

□

Proof of Theorem 3.4.2. (i) It follows from the second inequality in (3.48) that

$$\sup_{\theta\in\Theta_d(s,a)} \frac{1}{s}\mathbf{E}_\theta|\hat{\eta}-\eta| \leq 2\left(\frac{d}{s}-1\right)\Phi(-t/\sigma) + 2\Phi\big(-(a-t)_+/\sigma\big), \tag{3.60}$$

where $t = \frac{a}{2} + \frac{\sigma^2}{a}\log(\frac{d}{s}-1)$ is the threshold (3.5). Since $a^2 \geq 2\sigma^2\log(d/s-1)$ we get that $a \geq t$ and that $t > a/2$, which is equivalent to $t > a-t$. Furthermore, $(\frac{d}{s}-1)e^{-t^2/(2\sigma^2)} = e^{-(a-t)^2/(2\sigma^2)}$. These remarks and (3.31) imply that

$$\begin{aligned}\left(\frac{d}{s}-1\right)\Phi(-t/\sigma) &\leq \sqrt{\frac{2}{\pi}}\frac{\exp(-(a-t)^2/(2\sigma^2))}{(a-t)/\sigma+\sqrt{(a-t)^2/\sigma^2+8/\pi}} \\ &\leq \frac{\exp(-(a-t)^2/(2\sigma^2))}{(a-t)/\sigma+\sqrt{(a-t)^2/\sigma^2+4}} \\ &\leq \sqrt{\frac{\pi}{2}}\Phi\left(-\frac{a-t}{\sigma}\right).\end{aligned}$$

Combining this with (3.60), we get

$$\sup_{\theta\in\Theta_d(s,a)} \frac{1}{s}\mathbf{E}_\theta|\hat{\eta}-\eta| \leq (2+\sqrt{2\pi})\Phi\left(-\frac{a-t}{\sigma}\right).$$

Now, to prove (3.33) it remains to note that under assumption (3.32),

$$\frac{a-t}{\sigma} = \frac{a}{2\sigma} - \frac{\sigma}{a}\log\left(\frac{d}{s}-1\right) = \frac{a^2-2\sigma^2\log((d-s)/s)}{2a\sigma} \geq \Delta.$$

Indeed, assumption (3.32) states that $a \geq a_0 \triangleq \sigma(2\log((d-s)/s)+W)^{1/2}$, and the function $a \mapsto (a^2-2\sigma^2\log((d-s)/s))/a$ is monotonically increasing in $a>0$. On the other hand,

$$\big(a_0^2 - 2\sigma^2\log\big((d-s)/s\big)\big)/(2a_0\sigma) = \Delta. \tag{3.61}$$

(ii) We now prove (3.36). By Theorem 3.2.2,

$$\inf_{\tilde{\eta}}\sup_{\theta\in\Theta_d(s,a)} \frac{1}{s}\mathbf{E}_\theta|\tilde{\eta}-\eta| \geq \frac{s'}{s}\Phi\left(-\frac{a}{2\sigma}+\frac{\sigma}{a}\log\left(\frac{d}{s}-1\right)\right) - 4\frac{s'}{s}\exp\left(-\frac{(s-s')^2}{2s}\right).$$

Here,

$$-\frac{a}{2\sigma}+\frac{\sigma}{a}\log\left(\frac{d}{s}-1\right) = \frac{2\sigma^2\log((d-s)/s)-a^2}{2\sigma a}.$$

Observe that the function $a \mapsto (2\sigma^2 \log((d-s)/s) - a^2)/a$ is monotonically decreasing in $a > 0$ and that assumption (3.35) states that $a \le a_0$. In view of (3.61), the value of its minimum for $a \le a_0$ is equal to $-\Delta$. The bound (3.36) now follows by the monotonicity of $\Phi(\cdot)$. □

Proof of Theorem 3.4.3. Assume without loss of generality that d is large enough to have $(d-s_d)/s_d > 1$. We apply Theorem 3.4.2 with $W = A\sqrt{2\log((d-s_d)/s_d)}$. Then

$$\Delta^2 = \frac{A^2\sqrt{2\log((d-s_d)/s_d)}}{4(\sqrt{2\log((d-s_d)/s_d)} + A)}.$$

By assumption, there exists $\nu > 0$ such that $(2+\nu)s_d \le d$ for all d large enough. Equivalently, $d/s_d - 1 \ge 1+\nu$ and, therefore, using the monotonicity argument, we find

$$\Delta^2 \ge \frac{A^2\sqrt{2\log(1+\nu)}}{\sqrt{2\log(1+\nu)} + A} \to \infty \qquad \text{as } A \to \infty.$$

This and (3.33) imply part (i) of the theorem.

Part (ii) follows from (3.36) by noticing that $\Delta^2 \le \sup_{x>0} \frac{A^2 x}{4(x+A)} = A^2/4$ for any fixed $A > 0$. Now, for s large enough, let us put $s' = (1-\varepsilon)s$ for some ε in (0,1), fixed. Thus, the lower bound of the risk becomes

$$(1-\varepsilon)\Phi(-\Delta) - 4\exp\left(-\frac{s}{2}(1-\varepsilon)^2\right) > 0,$$

for s large enough. □

Proof of Theorem 3.4.4. Throughout the proof, we assume without loss of generality that d is large enough to have $s_d \ge 2$, and $(d-s_d)/s_d > 1$. Set $W_*(s) \triangleq 4(\log s + \sqrt{\log s \log(d-s)})$, and notice that

$$\frac{W_*(s_d)}{2\sqrt{2\log((d-s_d)/s_d) + W_*(s_d)}} = \sqrt{2\log s_d}, \tag{3.62}$$

$$2\log\big((d-s_d)/s_d\big) + W_*(s_d) = 2\big(\sqrt{\log(d-s_d)} + \sqrt{\log s_d}\big)^2. \tag{3.63}$$

If (3.40) holds, we have $W_d \ge W_*(s_d)$ for all d large enough. By the monotonicity of the quantity Δ defined in (3.34) with respect to W, this implies

$$\begin{aligned} \Delta_d &\triangleq \frac{W_d}{2\sqrt{2\log((d-s_d)/s_d) + W_d}} \\ &\ge \frac{W_*(s_d)}{2\sqrt{2\log((d-s_d)/s_d) + W_*(s_d)}} = \sqrt{2\log s_d}. \end{aligned} \tag{3.64}$$

Now, by Theorem 3.4.2 and using (3.31) we may write

$$\begin{aligned} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat\eta - \eta| &\le (2+\sqrt{2\pi})s_d\Phi(-\Delta_d) \\ &\le 3s_d \min\left\{1, \frac{1}{\Delta_d}\right\}\exp\left(-\frac{\Delta_d^2}{2}\right) \\ &= 3\min\left\{1, \frac{1}{\Delta_d}\right\}\exp\left(-\frac{\Delta_d^2 - 2\log s_d}{2}\right). \end{aligned} \tag{3.65}$$

This and (3.64) imply that, for all d large enough,

$$\sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat\eta-\eta| \le 3\min\left\{1, \frac{1}{\sqrt{2\log s_d}}\right\}.$$

Since $s_d\to\infty$, part (i) of the theorem follows.

We now prove part (ii) of the theorem. It suffices to consider $W_d>0$ for all d large enough since for nonpositive W_d almost full recovery is impossible and the result follows from part (ii) of Theorem 3.4.3. If (3.42) holds, there exists $A<1$ such that $W_d\le AW_*(s_d)$ for all d large enough. By the monotonicity of the quantity Δ defined in (3.34) with respect to W and in view of equation (3.62), this implies

$$\begin{aligned}
&\Delta_d^2 - 2\log s_d \\
&\quad\le \frac{A^2W_*^2(s_d)}{4(2\log((d-s_d)/s_d)+AW_*(s_d))} \\
&\qquad - \frac{W_*^2(s_d)}{4(2\log((d-s_d)/s_d)+W_*(s_d))} \\
&\quad= \frac{(A-1)W_*^2(s_d)(AW_*(s_d)+2(A+1)\log((d-s_d)/s_d))}{4(2\log((d-s_d)/s_d)+AW_*(s_d))(2\log((d-s_d)/s_d)+W_*(s_d))} \qquad (3.66)\\
&\quad\le \frac{(A-1)AW_*^2(s_d)}{4(2\log((d-s_d)/s_d)+W_*(s_d))} \\
&\quad= \frac{2(A-1)A(\log s_d+\sqrt{\log s_d\log(d-s_d)})^2}{(\sqrt{\log(d-s_d)}+\sqrt{\log s_d})^2} = 2(A-1)A\log s_d,
\end{aligned}$$

where we have used the fact that $A<1$ and equations (3.62), (3.63). Next, by Theorem 3.4.2 and using (3.31), we have for $s'=s_d/2$,

$$\inf_{\tilde\eta}\sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\tilde\eta-\eta| \;\ge\; \frac{s_d}{2}\left(\Phi(-\Delta_d)-4\exp\left(-\frac{s_d}{8}\right)\right)$$

and

$$\begin{aligned}
\frac{s_d}{2}\Phi(-\Delta_d) &\ge \frac{s_d}{8}\min\left\{\frac12,\frac{1}{\Delta_d}\right\}\exp\left(-\frac{\Delta_d^2}{2}\right)\\
&= \frac18\min\left\{\frac12,\frac{1}{\Delta_d}\right\}\exp\left(-\frac{\Delta_d^2-2\log s_d}{2}\right).
\end{aligned}$$

Combining this inequality with (3.66), we find that, for all d large enough,

$$\inf_{\tilde\eta}\sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\tilde\eta-\eta| \ge \frac18\min\left\{\frac12,\frac{1}{\Delta_d}\right\}e^{(1-A)A\log s_d} - 2s_de^{-s_d/8}.$$

Since $A<1$ and $\Delta_d\le A\sqrt{2\log s_d}$ by (3.66), the last expression tends to ∞ as $s_d\to\infty$. This proves part (ii) of the theorem. □

Proof of Theorem 3.5.1. By (3.48), for any $\theta\in\Theta_d(s_d,a_d)$, and any $t>0$ we have

$$\mathbf{E}_\theta|\hat\eta-\eta| \le d\mathbf{P}\big(|\xi|\ge t/\sigma\big) + s_d\mathbf{P}\big(|\xi|>(a_d-t)_+/\sigma\big),$$

where ξ is a standard normal random variable. It follows that, for any $a_d \geq a_d^*$, any $\theta \in \Theta_d(s_d, a_d)$, and any $t > 0$,

$$\mathbf{E}_\theta|\hat{\eta} - \eta| \leq d\mathbf{P}\big(|\xi| \geq t/\sigma\big) + s_d\mathbf{P}\big(|\xi| > \big(a_d^* - t\big)_+/\sigma\big).$$

Without loss of generality assume that $d \geq 6$ and $2 \leq s_d \leq d/2$. Then, using the inequality $\sqrt{x} - \sqrt{y} \leq (x-y)/\sqrt{2y}$, $\forall x > y > 0$, we find that, for $t = \sigma\sqrt{2\log d}$,

$$\begin{aligned}
\big(a_d^* - t\big)_+/\sigma &\geq \sqrt{2}\big(\sqrt{\log(d-s_d)} - \sqrt{\log d} + \sqrt{\log(s_d)}\big) \\
&\geq \sqrt{2\log(s_d)} - \log\left(\frac{d}{d-s_d}\right)/\sqrt{\log(d-s_d)} \\
&\geq \sqrt{2\log(s_d)} - (\log 2)/\sqrt{\log(d/2)} > 0.
\end{aligned}$$

From this we also easily deduce that, for $2 \leq s_d \leq d/2$, we have $((a_d^* - t)_+/\sigma)^2/2 \geq \log(s_d) - \sqrt{2}\log 2$. Combining these remarks with (3.31) and (3.43), we find

$$\sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat{\eta} - \eta| \leq \frac{1}{\sqrt{2\log d}} + \frac{s_d \exp(-\log(s_d) + \sqrt{2}\log 2)}{\sqrt{2\log(s_d)}},$$

which immediately implies the theorem by taking the limit as $d \to \infty$. □

Proof of Theorem 3.5.2. Throughout the proof, we will write for brevity $s_d = s, a_d = a, A_d = A$, and set $\sigma = 1$. Since $\Theta_d(s,a) \subseteq \Theta_d(s, a_0(s,A))$ for all $a \geq a_0(s,A)$, it suffices to prove that

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s,a_0(s,A))} \frac{1}{s}\mathbf{E}_\theta\big|\hat{\eta}^{\mathrm{ad}} - \eta\big| = 0. \tag{3.67}$$

Here, $s \leq s_d^*$ and recall that throughout this section we assume that $s_d^* \leq d/4$; since we deal with asymptotics as $d/s_d^* \to \infty$, the latter assumption is without loss of generality in the current proof.

If $s < g_M$, let $m_0 \in \{2, \ldots, M\}$ be the index such that g_{m_0} is the minimal element of the grid, which is greater than the true underlying s. Thus, $g_{m_0}/2 = g_{m_0-1} \leq s < g_{m_0}$. If $s \in [g_M, s_d^*]$, we set $m_0 = M$. In both cases,

$$s \geq g_{m_0}/2. \tag{3.68}$$

We decompose the risk as follows:

$$\frac{1}{s}\mathbf{E}_\theta\big|\hat{\eta}^{\mathrm{ad}} - \eta\big| = I_1 + I_2,$$

where

$$\begin{aligned}
I_1 &= \frac{1}{s}\mathbf{E}_\theta\big(\big|\hat{\eta}(g_{\hat{m}}) - \eta\big|I(\hat{m} \leq m_0)\big), \\
I_2 &= \frac{1}{s}\mathbf{E}_\theta\big(\big|\hat{\eta}(g_{\hat{m}}) - \eta\big|I(\hat{m} \geq m_0 + 1)\big).
\end{aligned}$$

We now evaluate I_1. Using the fact that $\hat{\eta}_j(g_m)$ is monotonically increasing in m and the definition of $\hat{m}$, we obtain that, on the event $\{\hat{m} \le m_0\}$,

$$\begin{aligned}
\big|\hat{\eta}(g_{\hat{m}}) - \hat{\eta}(g_{m_0})\big| &\le \sum_{m=\hat{m}+1}^{m_0} \big|\hat{\eta}(g_m) - \hat{\eta}(g_{m-1})\big| \\
&= \sum_{m=\hat{m}+1}^{m_0} \sum_{j=1}^{d} \big(\hat{\eta}_j(g_m) - \hat{\eta}_j(g_{m-1})\big) \\
&= \sum_{m=\hat{m}+1}^{m_0} \sum_{j=1}^{d} I\big(w(g_m) \le |X_j| < w(g_{m-1})\big) \\
&\le \tau \sum_{m=\hat{m}+1}^{m_0} g_m \le \tau s \sum_{m=2}^{m_0} 2^{m-m_0+1} \le 4\tau s,
\end{aligned}$$

where we have used the equality $g_m = 2^m$ and (3.68). Thus,

$$\begin{aligned}
I_1 &\le \frac{1}{s}\mathbf{E}_\theta\big(\big|\hat{\eta}(g_{\hat{m}}) - \hat{\eta}(g_{m_0})\big| I(\hat{m} \le m_0)\big) + \frac{1}{s}\mathbf{E}_\theta\big|\hat{\eta}(g_{m_0}) - \eta\big| \\
&\le 4\tau + \frac{1}{s}\mathbf{E}_\theta\big|\hat{\eta}(g_{m_0}) - \eta\big|.
\end{aligned} \tag{3.69}$$

Next, note that the first inequality in (3.48) is true for any $t > 0$. Applying it with $t = w(g_{m_0})$, we obtain

$$\begin{aligned}
\frac{1}{s}\mathbf{E}_\theta\big|\hat{\eta}(g_{m_0}) - \eta\big| \le \frac{d}{s}\mathbf{P}\big(|\xi| \ge w(g_{m_0})\big) \\
+ \mathbf{P}\big(|\xi| > \big(a_0(s, A) - w(g_{m_0})\big)_+\big)
\end{aligned} \tag{3.70}$$

where ξ is a standard Gaussian random variable. Using the bound on the Gaussian tail probability and the fact that $g_{m_0} > s \ge g_{m_0}/2$, we get

$$\begin{aligned}
\frac{d}{s}\mathbf{P}\big(|\xi| \ge w(g_{m_0})\big) &\le \frac{d/s}{d/g_{m_0} - 1} \frac{\pi^{-1/2}}{\sqrt{\log(d/g_{m_0} - 1)}} \\
&\le \frac{d}{d - 2s} \frac{2\pi^{-1/2}}{\sqrt{\log(d/s - 1)}} \le \frac{4\pi^{-1/2}}{\sqrt{\log(d/s_d^* - 1)}}.
\end{aligned} \tag{3.71}$$

To bound the second probability on the right-hand side of (3.70), we use the following lemma.

Lemma 3.6.1. *Under the assumptions of Theorem 3.5.2, for any $m \ge m_0$ we have*

$$\mathbf{P}\big(|\xi| > \big(a_0(s, A) - w(g_m)\big)_+\big) \le \big(\log\big(d/s_d^* - 1\big)\big)^{-\frac{1}{2}}. \tag{3.72}$$

Combining (3.70), (3.71) and (3.72) with $m = m_0$, we find

$$\frac{1}{s}\mathbf{E}_\theta\big|\hat{\eta}(g_{m_0}) - \eta\big| \le \frac{4\pi^{-1/2} + 1}{\sqrt{\log(d/s_d^* - 1)}}, \tag{3.73}$$

which together with (3.69) leads to the bound

$$I_1 \;\le\; 4\tau + \frac{4\pi^{-1/2}+1}{\sqrt{\log(d/s_d^*-1)}}. \tag{3.74}$$

We now turn to the evaluation of I_2. It is enough to consider the case $m_0 \le M-1$ since $I_2 = 0$ when $m_0 = M$. We have

$$\begin{aligned} I_2 &= \frac{1}{s}\sum_{m=m_0+1}^{M} \mathbf{E}_\theta\big(|\hat\eta(g_{\hat m}) - \eta| I(\hat m = m)\big) \\ &\le \frac{1}{s}\sum_{m=m_0+1}^{M} \big(\mathbf{E}_\theta|\hat\eta(g_m)-\eta|^2\big)^{1/2}\big(\mathbf{P}_\theta(\hat m = m)\big)^{1/2}. \end{aligned} \tag{3.75}$$

By definition, the event $\{\hat m = m\}$ occurs implies that $\sum_{j=1}^d I(w_m \le |X_j| < w_{m-1}) > \tau g_m \triangleq v_m$, where we set for brevity $w_m = w(g_m)$. Thus,

$$\mathbf{P}_\theta(\hat m = m) \;\le\; \mathbf{P}_\theta\left(\sum_{j=1}^d I\big(w_m \le |X_j| < w_{m-1}\big) > v_m\right). \tag{3.76}$$

By Bernstein's inequality, for any $t > 0$ we have

$$\begin{aligned} &\mathbf{P}_\theta\left(\sum_{j=1}^d I\big(w_m \le |X_j| < w_{m-1}\big) - \mathbf{E}_\theta\left(\sum_{j=1}^d I\big(w_m \le |X_j| < w_{m-1}\big)\right) > t\right) \\ &\quad\le \exp\left(-\frac{t^2/2}{\sum_{j=1}^d \mathbf{E}_\theta(I(w_m \le |X_j| < w_{m-1})) + 2t/3}\right), \end{aligned} \tag{3.77}$$

where we have used that, for random variables with values in $\{0,1\}$, the variance is smaller than the expectation.

Now, similar to (3.48), for any $\theta \in \Theta_d(s, a_0(s,A))$,

$$\begin{aligned} &\mathbf{E}_\theta\left(\sum_{j=1}^d I\big(w_m \le |X_j| < w_{m-1}\big)\right) \\ &\quad\le d\mathbf{P}\big(w_m \le |\xi| < w_{m-1}\big) + \sum_{j:\theta_j\ne 0}\mathbf{P}\big(|\theta_j+\xi| < w_{m-1}\big) \\ &\quad\le d\mathbf{P}\big(|\xi| \ge w_m\big) + s\mathbf{P}\big(|\xi| > -\big(a_0(s,A) - w_{m-1}\big)_+\big), \end{aligned}$$

where ξ is a standard Gaussian random variable. Since $m \ge m_0+1$, from Lemma 3.6.1 we get

$$\mathbf{P}\big(|\xi| > \big(a_0(s,A) - w_{m-1}\big)_+\big) \le \big(\log\big(d/s_d^*-1\big)\big)^{-\frac{1}{2}}. \tag{3.78}$$

Next, using the bound on the Gaussian tail probability and the inequalities $g_m \le s_d^* \le d/4$, we find

$$d\mathbf{P}\big(|\xi| \ge w_m\big) \le \frac{d}{d/g_m - 1}\frac{\pi^{-1/2}}{\sqrt{\log(d/g_m-1)}} \le \frac{(4/3)\pi^{-1/2} g_m}{\sqrt{\log(d/s_d^*-1)}}. \tag{3.79}$$

We now deduce from (3.78) and (3.79), and the inequality $s \leq g_m$ for $m \geq m_0 + 1$, that

$$\mathbf{E}_\theta\left(\sum_{j=1}^{d} I\big(w_m \leq |X_j| < w_{m-1}\big)\right) \leq \frac{((4/3)\pi^{-1/2}+1)g_m}{\sqrt{\log(d/s_d^*-1)}} \leq 2\tau g_m. \tag{3.80}$$

Taking in (3.77) $t = 3\tau g_m = 3v_m$ and using (3.80), we find

$$\mathbf{P}_\theta\left(\sum_{j=1}^{d} I\big(w_m \leq |X_j| < w_{m-1}\big) > v_m\right) \leq \exp(-C_1 v_m) = \exp\big(-C_1 2^m \tau\big),$$

for some absolute constant $C_1 > 0$. This implies

$$\mathbf{P}_\theta(\hat{m} = m) \leq \exp\big(-C_1 2^m \tau\big). \tag{3.81}$$

On the other hand, notice that the bounds (3.70), and (3.71) are valid not only for g_{m_0} but also for any g_m with $m \geq m_0 + 1$. Using this observation and Lemma 3.6.1 we get that, for any $\theta \in \Theta_d(s, a_0(s, A))$ and any $m \geq m_0 + 1$,

$$\begin{aligned}\mathbf{E}_\theta\big|\hat{\eta}(g_m) - \eta\big| &\leq s\left[\frac{d/s}{d/g_m - 1}\frac{\pi^{-1/2}}{\sqrt{\log(d/g_m - 1)}} + \big(\log\big(d/s_d^* - 1\big)\big)^{-\frac{1}{2}}\right] \\ &\leq \frac{((4/3)\pi^{-1/2}+1)g_m}{\sqrt{\log(d/s_d^*-1)}} \triangleq \tau' g_m = \tau' 2^m,\end{aligned} \tag{3.82}$$

where the last inequality follows from the same argument as in (3.79). We denote by $Var_\theta\left(\big|\hat{\eta}(g_m) - \eta\big|\right)$ the variance of $\big|\hat{\eta}(g_m) - \eta\big|$. Observing that $\big|\hat{\eta}(g_m) - \eta\big|$ is a sum of independent Bernoulli random variables, we get

$$\begin{aligned}\mathbf{E}_\theta\big|\hat{\eta}(g_m) - \eta\big|^2 &= Var_\theta\left(\big|\hat{\eta}(g_m) - \eta\big|\right) + \big(\mathbf{E}_\theta\big|\hat{\eta}(g_m) - \eta\big|\big)^2 \\ &\leq \mathbf{E}_\theta\big|\hat{\eta}(g_m) - \eta\big| + \big(\mathbf{E}_\theta\big|\hat{\eta}(g_m) - \eta\big|\big)^2.\end{aligned}$$

Using (3.82) and the fact that τ' is bounded, we get that

$$\mathbf{E}_\theta\big|\hat{\eta}(g_m) - \eta\big|^2 \leq C_2 \tau' 2^{2m}, \tag{3.83}$$

for some absolute constant $C_2 > 0$.

Now, we plug (3.81) and (3.83) in (3.75) to obtain

$$\begin{aligned}I_2 &\leq \frac{(C_2\tau')^{1/2}}{s}\sum_{m=m_0+1}^{M} 2^m \exp\big(-C_1 2^{m-1}\tau\big) \\ &\leq C_3\big(\tau'\big)^{1/2}\tau^{-1}\exp\big(-C_1 2^{m_0-1}\tau\big) \leq C_3\big(\tau'\big)^{1/2}\tau^{-1}\end{aligned}$$

for some absolute constant $C_3 > 0$. Notice that $(\tau')^{1/2} = O((\log(d/s_d^* - 1))^{-\frac{1}{4}})$ as $d/s_d^* \to \infty$ while $\tau^{-1} = O((\log(d/s_d^* - 1))^{\frac{1}{7}})$. Thus, $I_2 = o(1)$ as $d \to \infty$. Since from (3.74) we also get that $I_1 = o(1)$ as $d \to \infty$, the proof is complete. □

Proof of Lemma 3.6.1. Let first $s < g_M$. Then, by definition of m_0, we have $s < g_{m_0}$. Therefore, $s < g_m$ for $m \ge m_0$, and we have $w(g_m) < w(s)$. It follows that

$$a_0(s,A) - w(g_m) \ge a_0(s,A) - w(s) \ge \frac{\sqrt{A}}{2\sqrt{2}} \min\biggl(\frac{\sqrt{A}}{\sqrt{2}}, \log^{1/4}(d/s-1)\biggr),$$

where we have used the elementary inequalities

$$\sqrt{x+y} - \sqrt{x} \ge y/(2\sqrt{x+y}) \ge (2\sqrt{2})^{-1} \min(y/\sqrt{x}, \sqrt{y})$$

with $x = 2\log(d/s-1)$ and $y = A\sqrt{\log(d/s-1)}$. By assumption, $A \ge 16\sqrt{\log\log(d/s_d^*-1)}$, so that we get

$$a_0(s,A) - w(g_m) \ge a_0(s,A) - w(s) \ge 4\biggl(\log\log\biggl(\frac{d}{s_d^*}-1\biggr)\biggr)^{1/2}. \tag{3.84}$$

This and the standard bound on the Gaussian tail probability imply

$$\begin{aligned}\mathbf{P}\bigl(|\xi| > \bigl(a_0(s,A) - w(g_m)\bigr)_+\bigr) &\le \exp\bigl(-\bigl(a_0(s,A) - w(g_m)\bigr)^2/2\bigr)\\ &\le \bigl(\log\bigl(d/s_d^*-1\bigr)\bigr)^{-\frac{1}{2}}.\end{aligned} \tag{3.85}$$

Let now $s \in [g_M, s_d^*]$. Then $m_0 = M$ and we need to prove the result only for $m = M$. By definition of M, we have $s_d^* \le 2g_M$. This and (3.84) imply

$$\begin{aligned}a_0(s,A) - w(g_M) &\ge a_0(s,A) - w(s) - \bigl(w\bigl(s_d^*/2\bigr) - w\bigl(s_d^*\bigr)\bigr)\\ &\ge 4\biggl(\log\log\biggl(\frac{d}{s_d^*}-1\biggr)\biggr)^{1/2} - \bigl(w\bigl(s_d^*/2\bigr) - w\bigl(s_d^*\bigr)\bigr).\end{aligned}$$

Now, using the elementary inequality $\sqrt{\log(x+y)} - \sqrt{\log(x)} \le y/(2x\sqrt{\log(x)})$ with $x = d/s_d^* - 1$ and $y = d/s_d^*$, and the fact that $s_d^* \le d/4$ we find

$$\begin{aligned}w\bigl(s_d^*/2\bigr) - w\bigl(s_d^*\bigr) &\le \frac{1}{\sqrt{2\log(d/s_d^*-1)}} \frac{d}{d-s_d^*} \le \frac{2\sqrt{2}}{3\sqrt{\log(d/s_d^*-1)}}\\ &\le 3\biggl(\log\log\biggl(\frac{d}{s_d^*}-1\biggr)\biggr)^{1/2}.\end{aligned}$$

The last two displays yield $a_0(s,A) - w(g_M) \ge (\log\log(\frac{d}{s_d^*}-1))^{1/2}$, and we conclude as in (3.85). □

3.7 Appendix: More proofs of lower bounds

In this section, we derive a general lower bound for the minimax risk over all selectors on the class of at most s-sparse vectors. The main term of this bound is a Bayes risk with arbitrary prior and the non-asymptotic remainder term is given explicitly.

Nonasymptotic lower bound of the minimax risk

In the next theorem, we reduce the minimax risk over all selectors to a Bayes risk with arbitrary prior measure π on $\{0,1\}^d$ and give a bound on the difference between the two risks. This result is true in a general setup, non necessarily for Gaussian models. For a particular choice of measure π, we provide an explicit bound on the remainder term.

Consider the set of binary vectors

$$\Theta_s = \left\{\eta \in \{0,1\}^d : \ |\eta|_0 \le s\right\}, \ \text{where } |\eta|_0 = \sum_{j=1}^d I(\eta_j \neq 0),$$

and assume that we are given a family $\{P_\eta, \eta \in \Theta_s\}$ where each P_η is a probability distribution on a measurable space $(\mathcal{X}, \mathcal{U})$. We observe X drawn from P_η with some unknown $\eta \in \Theta_s$ and we consider the Hamming risk of a selector $\hat\eta = \hat\eta(X)$:

$$\sup_{\eta\in\Theta_s} \mathbf{E}_\eta |\hat\eta - \eta|$$

where $\mathbf{E}_\eta$ is the expectation with respect to P_η. Here and in what follows we denote by $|\eta - \eta'|$ the Hamming distance between two binary sequences $\eta, \eta' \in \{0,1\}^d$, and we call the selector any estimator with values in $\{0,1\}^d$. Let π be a probability measure on $\{0,1\}^d$ (a prior on η). We denote by $\mathbb{E}_\pi$ the expectation with respect to π.

Theorem 3.7.1. *For any $s < d$ and any probability measure π on $\{0,1\}^d$, we have*

$$\inf_{\hat\eta} \sup_{\eta\in\Theta_s} \mathbf{E}_\eta |\hat\eta - \eta| \ge \inf_{\hat T\in[0,1]^d} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^d |\hat T_j(X) - \eta_j| - 4\,\mathbb{E}_\pi\left[|\eta|_0 I(|\eta|_0 \ge s+1)\right], \quad (3.86)$$

where $\inf_{\hat\eta}$ is the infimum over all selectors and $\inf_{\hat T\in[0,1]^d}$ is the infimum over all estimators $\hat T(X) = (\hat T_1(X), \dots, \hat T_d(X))$ with values in $[0,1]^d$.

In particular, if π is a product of d Bernoulli distributions with parameters d and s'/d where $s' \in (0, s]$, we have

$$\inf_{\hat\eta} \sup_{\eta\in\Theta_s} \mathbf{E}_\eta |\hat\eta - \eta| \ge \inf_{\hat T\in[0,1]^d} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^d |\hat T_j - \eta_j| - 4s' \exp\Big(-\frac{(s-s')^2}{2s}\Big). \quad (3.87)$$

Proof of Theorem 3.7.1. Throughout the proof, we write for brevity $A = \Theta_s$. Set $\eta^A = \eta I(\eta \in A)$ and denote by π_A the probability measure π conditioned by the event $\{\eta \in A\}$, that is, for any $C \subseteq \{0,1\}^d$,

$$\pi_A(C) = \frac{\pi(C \cap \{\eta \in A\})}{\pi(\eta \in A)}\,.$$

The measure π_A is supported on A and we have

$$\begin{aligned}\inf_{\hat\eta} \sup_{\eta\in A} \mathbf{E}_\eta |\hat\eta - \eta| &\ge \inf_{\hat\eta} \mathbb{E}_{\pi_A} \mathbf{E}_\eta |\hat\eta - \eta| = \inf_{\hat\eta} \mathbb{E}_{\pi_A} \mathbf{E}_\eta |\hat\eta - \eta^A| \\ &\ge \sum_{j=1}^d \inf_{\hat T_j} \mathbb{E}_{\pi_A} \mathbf{E}_\eta |\hat T_j - \eta_j^A|\end{aligned}$$

where $\inf_{\hat{T}_j}$ is the infimum over all estimators $\hat{T}_j = \hat{T}_j(X)$ with values in $\mathbb{R}$. According to Theorem 1.1 and Corollary 1.2 on page 228 in Lehmann and Casella (2006), there exists a Bayes estimator $B_j^A = B_j^A(X)$ such that

$$\inf_{\hat{T}_j} \mathbb{E}_{\pi_A} \mathbf{E}_\eta |\hat{T}_j - \eta_j^A| = \mathbb{E}_{\pi_A} \mathbf{E}_\eta |B_j^A - \eta_j^A|,$$

and this estimator is a conditional median of η_j^A given X; in particular, for any estimator $\hat{T}_j(X)$ we have

$$\mathbb{E}^A\big(|B_j^A(X) - \eta_j^A| \big| X\big) \le \mathbb{E}^A\big(|\hat{T}_j(X) - \eta_j^A| \big| X\big) \tag{3.88}$$

almost surely. Here, the superscript A indicates that the conditional expectation $\mathbb{E}^A(\cdot|X)$ is taken when η is distributed according to π_A. Therefore,

$$\inf_{\hat{\eta}} \sup_{\eta \in A} \mathbf{E}_\eta |\hat{\eta} - \eta| \ge \mathbb{E}_{\pi_A} \mathbf{E}_\eta \sum_{j=1}^d |B_j^A - \eta_j^A|. \tag{3.89}$$

Note that $B_j^A \in [0,1]$ since η_j^A takes its values in $[0,1]$. Using this, we obtain

$$\begin{aligned}
\inf_{\hat{T} \in [0,1]^p} \mathbb{E}_\pi \mathbf{E}_\eta |\hat{T} - \eta| &\le \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^d |B_j^A - \eta_j| \\
&= \mathbb{E}_\pi \mathbf{E}_\eta \Big(\sum_{j=1}^d |B_j^A - \eta_j| I(\eta \in A) \Big) + \mathbb{E}_\pi \mathbf{E}_\eta \Big(\sum_{j=1}^d |B_j^A - \eta_j| I(\eta \in A^c) \Big) \\
&= \mathbb{E}_{\pi_A} \mathbf{E}_\eta \sum_{j=1}^d |B_j^A - \eta_j^A| + \mathbb{E}_\pi \mathbf{E}_\eta \Big(\sum_{j=1}^d |B_j^A - \eta_j| I(\eta \in A^c) \Big) \\
&\le \mathbb{E}_{\pi_A} \mathbf{E}_\eta \sum_{j=1}^d |B_j^A - \eta_j^A| + \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^d B_j^A I(\eta \in A^c) + \mathbb{E}_\pi \sum_{j=1}^d \eta_j I(\eta \in A^c).
\end{aligned} \tag{3.90}$$

Our next step is to bound the term

$$\mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^d B_j^A I(\eta \in A^c).$$

For this purpose, we first note that inequality (3.88) with $\hat{T}_j(X) = \mathbb{E}^A(\eta_j^A|X)$ implies that

$$B_j^A(X) = \mathbb{E}^A(B_j^A(X)|X) \le \mathbb{E}^A\big(|\mathbb{E}^A(\eta_j^A|X)| \big| X\big) + 2\mathbb{E}^A\big(|\eta_j^A| \big| X\big) = 3\mathbb{E}^A(\eta_j^A|X)$$

where we have used the fact that $\eta_j^A \in [0,1]$. Since $\sum_{j=1}^d \eta_j^A \le s$ (cf. definition of η_j^A), we find that $\sum_{j=1}^d B_j^A \le 3s$. Finally, as $\sum_{j=1}^d \eta_j > s$ on A^c we get $\sum_{j=1}^d B_j^A I(\eta \in A^c) \le 3 \sum_{j=1}^d \eta_j I(\eta \in A^c)$, and thus

$$\mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^d B_j^A I(\eta \in A^c) \le 3\, \mathbb{E}_\pi \sum_{j=1}^d \eta_j I(\eta \in A^c).$$

Combining this inequality with (3.89) and (3.90) yields (3.86).

We now prove inequality (3.87). In this case, $\sum_{j=1}^{d} \eta_j := \zeta$ has the binomial distribution $\mathcal{B}(d, q)$ with parameters d and $q = s'/d$. Then,

$$\begin{aligned}
\mathbf{E}(\zeta I(\zeta \geq s+1)) &= \sum_{k=s+1}^{d} k \binom{d}{k} q^k (1-q)^{d-k} \\
&= \sum_{k=s+1}^{d} \frac{d(d-1)!}{(k-1)!(d-k)!} q^k (1-q)^{d-k} \\
&= dq \sum_{k=s+1}^{d} \binom{d-1}{k-1} q^{k-1} (1-q)^{d-k} \\
&= dq \sum_{m=s}^{d-1} \binom{d-1}{m} q^m (1-q)^{(d-1)-m} \\
&= s' \mathbf{P}(\mathcal{B}(d-1, s'/d) \geq s) \leq s' \mathbf{P}(\mathcal{B}(d, s'/d) \geq s).
\end{aligned}$$

Thus, to complete the proof of (3.87), it is enough to bound the probability $\mathbf{P}(\mathcal{B}(d, s'/d) \geq s)$. To this end, we use the following lemma, which is a combination of formulas (3) and (10) on pages 440–441 in Shorack and Wellner (2009).

Lemma 3.7.1. *Let $\mathcal{B}(d, q)$ be the binomial random variable with parameters d and $q \in (0, 1)$. Then, for any $\lambda > 0$,*

$$\mathbf{P}\big(\mathcal{B}(d, q) \geq \lambda\sqrt{d} + dq\big) \leq \exp\left(- \frac{\lambda^2}{2q(1-q)\big(1 + \frac{\lambda}{3q\sqrt{d}}\big)} \right). \tag{3.91}$$

Applying this lemma with $q = s'/d$ and $\lambda = (s - s')/\sqrt{d}$ we find that

$$\mathbf{P}(\mathcal{B}(d, s'/d) \geq s) \leq \exp\left(- \frac{(s-s')^2}{2s} \right).$$

Thus, (3.87) follows.

□

Remark. If we take $s' = s - (3/2)\sqrt{s \log(s)}$ (which is possible since for all $s \geq 2$ we have $s' > 0$) inequality (3.87) implies that, for all $s \geq 2$,

$$\inf_{\hat{\eta}} \sup_{\eta \in A_s} \mathbf{E}_\eta |\hat{\eta} - \eta| \geq \inf_{\hat{T} \in [0,1]^d} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^{d} |\hat{T}_j - \eta_j| - 4s^{-1/8}.$$

Proof of the second lower bound in Theorem 3.2.2. This bound follows directly from the lower bounds of Theorems 3.2 and 3.3 that hold for general distributions. □

Proof of Theorem 3.3.2. The upper bound $\sup_{\eta \in \Theta_s} \mathbf{E}_\eta |\hat{\eta} - \eta| \leq \Psi(d, s) sd/(d - s)$ is straightforward in view of the definition of $\hat{\eta}$.

We now prove the lower bound of Theorem 3.2. First note that, in view of (3.87), the proof is reduced to showing that

$$\inf_{\hat{T}\in[0,1]^d} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^{d} |\hat{T}_j - \eta_j| \geq s'\Psi(d,s), \tag{3.92}$$

where π is a product on d Bernoulli distributions with parameter s'/d and $s' \in (0,s]$. We have

$$\begin{aligned}
\inf_{\hat{T}\in[0,1]^d} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^{d} |\hat{T}_j - \eta_j| &\geq \sum_{j=1}^{d} \inf_{\hat{T}_j\in[0,1]} \mathbb{E}_\pi \mathbf{E}_\eta |\hat{T}_j(X) - \eta_j| \\
&\geq \sum_{j=1}^{d} \mathbb{E}_\pi \mathbf{E}_\eta \Big(\inf_{\hat{T}_j\in[0,1]} \mathbb{E}_\pi \mathbf{E}_\eta \big(|\hat{T}_j(X) - \eta_j| \,\big|\, \{(X_j,\eta_j), j\neq i\}\big)\Big).
\end{aligned}$$

Since the components X_j of $X = (X_1,\ldots,X_d)$ are independent, the jth conditional expectation in the last expression reduces to the unconditional expectation over (X_j,η_j), which is bounded from below by

$$\inf_{T\in[0,1]} \left(\left(1 - \frac{s'}{d}\right) E_0(T) + \frac{s'}{d} E_1(1-T) \right) = s'\Psi(d,s')/d.$$

Here, E_i is the expectation with respect to P_i, $i = 0,1$. Thus,

$$\inf_{\hat{T}\in[0,1]^d} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{j=1}^{d} |\hat{T}_j - \eta_j| \geq s'\Psi(d,s').$$

To finish the proof of (3.92), it remains to show that $\Psi(d,s') \geq \Psi(d,s)$ for all $s' \leq s$. For this purpose, we extend $\Psi(d,\cdot)$ to $\mathbb{R}_+$ by defining, for all $u > 0$,

$$\begin{aligned}
\Psi(d,u) &= P_1\left(uf_1(X_1) - (d-u)f_0(X_1) < 0\right) + \left(\frac{d}{u} - 1\right) P_0\left(uf_1(X_1) - (d-u)f_0(X_1) \geq 0\right) \\
&= 1 - E_1[I(Y(u) \geq 0)] + \left(\frac{d}{u} - 1\right) E_0[I(Y(u) \geq 0)],
\end{aligned}$$

where $Y(u) = uf_1(X_1) - (d-u)f_0(X_1)$. For $\epsilon > 0$, we define the function $g_\epsilon : \mathbb{R} \to \mathbb{R}_+$ by

$$g_\epsilon(u) = \frac{2u^2}{\epsilon^2} I\left(0 \leq u < \frac{\epsilon}{2}\right) + \left(1 - \frac{2(\epsilon-u)^2}{\epsilon^2}\right) I\left(\frac{\epsilon}{2} \leq u < \epsilon\right) + I\left(u \geq \epsilon\right).$$

It is easy to check that g_ϵ is continuously differentiable on $\mathbb{R}$ and that, for all u in $\mathbb{R}$,

$$\lim_{\epsilon\to 0} g_\epsilon(u) = I(u \geq 0) \text{ and } u g_\epsilon'(u) \geq 0.$$

Finally, for $\epsilon > 0$ and $u > 0$, we define $\Psi_\epsilon(d,u)$ by the formula

$$\Psi_\epsilon(d,u) = 1 - E_1\left[g_\epsilon(Y(u))\right] + \left(\frac{d}{u} - 1\right) E_0\left[g_\epsilon(Y(u))\right].$$

An application of the dominated convergence theorem proves that $\lim_{\epsilon\to 0}\Psi_\epsilon(d,u)=\Psi(d,u)$ for all $u>0$, and that one can differentiate in the expression for $\Psi_\epsilon(d,\cdot)$ under the expectation signs on $\mathbb{R}_+$. We also note that $\Psi_\epsilon(d,\cdot)$ is decreasing on $\mathbb{R}_+$. Indeed, for any $u>0$,

$$\begin{aligned}\frac{\partial}{\partial u}\Psi_\epsilon(d,u) &= -\frac{d}{u^2}E_0\left[g_\epsilon(uf_1(X_1)-(d-u)f_0(X_1))\right]\\ &\quad -\sum_{i=0,1}E_i\Big[g'_\epsilon(uf_1(X_1)-(d-u)f_0(X_1))\frac{uf_1(X_1)-(d-u)f_0(X_1)}{u}\Big].\end{aligned}$$

Using the inequality $w\,g'_\epsilon(w)\ge 0$ we get that $\Psi_\epsilon(d,\cdot)$ is decreasing on $\mathbb{R}_+$. Finally, pointwise convergence of $\Psi_\epsilon(d,\cdot)$ to $\Psi(d,\cdot)$ implies that $\Psi(d,\cdot)$ is also decreasing on $\mathbb{R}_+$. □

Proof of Theorem 3.3.3. We have

$$\begin{aligned}\sup_{\theta\in\Theta_d^+(s,a_0,a_1)}\frac{1}{s}\mathbf{E}_\theta|\hat\eta-\eta| &= \sup_{a\ge a_1}\frac{|S|}{s}\mathbf{P}_a\Big(\log\frac{f_1}{f_0}(X_1)<\log\Big(\frac{d}{s}-1\Big)\Big)\\ &\quad+\sup_{a\le a_0}\Big(\frac{d-|S|}{s}\Big)\mathbf{P}_a\Big(\log\frac{f_1}{f_0}(X_1)\ge\log\Big(\frac{d}{s}-1\Big)\Big)\\ &= \frac{|S|}{s}\mathbf{P}_{a_1}\Big(\log\frac{f_1}{f_0}(X_1)<\log\Big(\frac{d}{s}-1\Big)\Big)\\ &\quad+\Big(\frac{d-|S|}{s}\Big)\mathbf{P}_{a_0}\Big(\log\frac{f_1}{f_0}(X_1)\ge\log\Big(\frac{d}{s}-1\Big)\Big)\\ &\le \Psi(d,s)\frac{d}{d-s},\end{aligned}$$

where the last equality is due to the monotonicity of $\log\frac{f_1}{f_0}(X)$ and to the stochastic order of the family $\{\mathsf{f}_a,\ a\in\mathcal{U}\}$.

The lower bound on the minimax risk

$$\inf_{\tilde\eta}\sup_{\theta\in\Theta_d^+(s,a_0,a_1)}\frac{1}{s}\mathbf{E}_\theta|\tilde\eta-\eta|$$

follows from the lower bound of Theorem 3.2 by taking there $f_0=\mathsf{f}_{a_0}$ and $f_1=\mathsf{f}_{a_1}$. □

Chapter 4

Optimal variable selection and adaptive noisy Compressed Sensing

For high-dimensional linear regression model, we propose an algorithm of exact support recovery in the setting of noisy compressed sensing where all entries of the design matrix are i.i.d standard Gaussian. This algorithm achieves the same conditions of exact recovery as the exhaustive search (maximal likelihood) decoder, and has an advantage over the latter of being adaptive to all parameters of the problem and computable in polynomial time. The core of our analysis consists in the study of the non-asymptotic minimax Hamming risk of variable selection. This allows us to derive a procedure, which is nearly optimal in a non-asymptotic minimax sense. Then, we develop its adaptive version, and propose a robust variant of the method to handle datasets with outliers and heavy-tailed distributions of observations. The resulting polynomial time procedure is near optimal, adaptive to all parameters of the problem and also robust.

Based on Ndaoud and Tsybakov (2018): Ndaoud, M. and Tsybakov, A. B. (2018). Optimal variable selection and adaptive noisy compressed sensing. *arXiv preprint arXiv:1809.03145*.

4.1 Introduction

Statement of the problem

Assume that we have the vector of measurements $Y \in \mathbb{R}^n$ satisfying

$$Y = X\beta + \sigma\xi \tag{4.1}$$

where $X \in \mathbb{R}^{n\times p}$ is a given design or sensing matrix, $\beta \in \mathbb{R}^p$ is the unknown signal, and $\sigma > 0$. In this chapter, we mostly focus on the setting where all entries of X are i.i.d. standard Gaussian random variables and the noise $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$ is a standard Gaussian vector independent of X. Here, $\mathbb{I}_n$ denotes the $n \times n$ identity matrix. This setting is typical for noisy compressed sensing, cf. references below. We will also consider extensions to sub-Gaussian design X and to noise ξ with heavy-tailed distribution.

In this chapter, one of the main problems that we are interested in consists in recovering the support of β, that is the set S_β of non-zero components of β. For an integer $s \leq p$, we assume that β is s-sparse, that is it has at most s non-zero components. We

also assume that these components cannot be arbitrarily small. This motivates us to define the following set $\Omega^p_{s,a}$ of s-sparse vectors:

$$\Omega^p_{s,a} = \{\beta \in \mathbb{R}^p : \quad |\beta|_0 \le s \quad \text{and} \quad |\beta_i| \ge a, \quad \forall i \in S_\beta\},$$

where $a > 0$, β_i are the components of β for $i = 1, \dots, p$, and $|\beta|_0$ denotes the number of non-zero components of β. We consider the problem of variable selection stated as follows: Given the observations (X, Y), estimate the binary vector

$$\eta_\beta = (\mathbf{1}\{\beta_1 \neq 0\}, \dots, \mathbf{1}\{\beta_p \neq 0\}),$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. In order to estimate η_β (and thus the support S_β), we define a decoder (selector) $\hat{\eta} = \hat{\eta}(X, Y)$ as a measurable function of the observations (X, Y) with values in $\{0, 1\}^p$. The performance of selector $\hat{\eta}$ is measured by the maximal risks

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{P}_\beta (\hat{\eta} \neq \eta_\beta) \qquad \text{and} \qquad \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta |\hat{\eta} - \eta_\beta|$$

where $|\hat{\eta} - \eta_\beta|$ stands for the Hamming distance between $\hat{\eta}$ and η_β, $\mathbf{P}_\beta$ denotes the joint distribution of (X, Y) satisfying (4.1), and $\mathbf{E}_\beta$ denotes the corresponding expectation. We say that a selector $\hat{\eta}$ achieves *exact support recovery* with respect to one of the above two risks if

$$\lim_{p \to \infty} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{P}_\beta (\hat{\eta} \neq \eta_\beta) = 0, \tag{4.2}$$

or

$$\lim_{p \to \infty} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta |\hat{\eta} - \eta_\beta| = 0, \tag{4.3}$$

where the asymptotics are considered as $p \to \infty$ when all other parameters of the problem (namely, n, s, a, σ) depend on p in such a way that $n = n(p) \to \infty$. In particular, the high-dimensional setting with $p \ge n$ is covered. For brevity, the dependence of these four parameters on p will be further omitted in the notation. Since

$$\mathbf{P}_\beta (\hat{\eta} \neq \eta_\beta) \le \mathbf{E}_\beta |\hat{\eta} - \eta_\beta|,$$

the property (4.3) implies (4.2). Therefore, we will mainly study the Hamming distance risk.

Notation. In the rest of this paper we use the following notation. For given sequences a_n and b_n, we say that $a_n = \mathcal{O}(b_n)$ (resp $a_n = \Omega(b_n)$) when $a_n \le c b_n$ (resp $a_n \ge c b_n$) for some absolute constant $c > 0$. We write $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, $\|\mathbf{x}\|$ is the Euclidean norm of $\mathbf{x}$, and $\mathbf{x}^\top \mathbf{y}$ the corresponding inner product. For a matrix X, we denote by X_j its jth column. For $x, y \in \mathbb{R}$, we denote by $x \vee y$ the maximum of x and y, by $\lfloor x \rfloor$ the maximal integer less than x and we set $x_+ = x \vee 0$. The notation $\mathbf{1}\{\cdot\}$ stands for the indicator function, and $|A|$ for the cardinality of a finite set A. We denote by C and c positive constants that can differ on different occurences.

Related literature

The literature on support recovery in high-dimensional linear model under sparsity is very rich and its complete overview falls beyond the format of this paper. Some important common features of the obtained results are as follows.

- The existing selectors (decoders) can be split into two main families. The first family consists of polynomial time algorithms, such as selectors based on the Lasso Zhao and Yu (2006); Wainwright (2009b), the orthogonal matching pursuit Tropp and Gilbert (2007); Zhang (2011b); Cai and Wang (2011) or thresholding Fletcher et al. (2009); Joseph (2013). The second contains exhaustive search methods, for instance, the Maximum Likelihood (ML) decoder; they are not realizable in polynomial time. The ML decoder outputs the support $S_{\hat{\beta}}$ of the least squares solution
$$\hat{\beta} \in \arg\min_{\theta:\, |\theta|_0 = s} \|Y - X\theta\|,$$
which is the ML estimator of β on the set $\{\beta : |\beta|_0 = s\}$ when the noise is Gaussian.

- The available results are almost exclusively of the form (4.2), where the asymptotics is considered under various additional restrictions on the behavior of (n, s, a, σ) as $p \to \infty$. One of the main restrictions concerns the asymptotic behavior of the signal-to-noise ratio (SNR). For $\sigma \asymp 1$, the noise and the entries of the sensing matrix X are of the same order as in Fletcher et al. (2009) and Wainwright (2009a), while in Aeron et al. (2010), $\sigma \asymp \sqrt{n}$, and hence the noise scales largely compared to the signal.

We now briefly overview results for specific asymptotics, with the emphasis on the *phase transition*, that is on the necessary and sufficient conditions of exact recovery. To the best of our knowledge, they cover only the exact recovery of the type (4.2).

In the strong noise regime $\sigma \asymp \sqrt{n}$, Aeron et al. (2010) show that necessary and sufficient conditions for (4.2) are given by $n = \Omega\left(s \log(\frac{p}{s})\right)$, and $a^2 = \Omega\left(\log(p-s)\right)$, and the ML decoder is optimal in the sense that it achieves exact recovery under these conditions. In the same regime $\sigma \asymp \sqrt{n}$, Saligrama and Zhao (2011) present a polynomial time procedure achieving (4.2) under sub-optimal sufficient conditions $n = \Omega\left(s \log(\frac{p}{s})\right)$, and $a^2 = \Omega\left((\log p)^3\right)$. This procedure requires a prior knowledge of the threshold a.

For $\sigma \asymp 1$, which is in fact the general case (equivalent to fixed σ), the results are different. First, the following necessary condition for exact recovery (in the sense (4.2)) for any decoder is obtained in Wang et al. (2010):

$$n = \Omega\left(\frac{s \log(\frac{p}{s})}{\log(1 + s\frac{a^2}{\sigma^2})} \vee \frac{\log(p-s)}{\log(1+\frac{a^2}{\sigma^2})}\right). \tag{4.4}$$

In Rad (2011), it is shown that, under the restrictions $a/\sigma = \mathcal{O}(1)$ and $a/\sigma = \Omega(1/\sqrt{s})$ on the signal-to-noise ratio a/σ, the ML decoder is optimal in the sense that it achieves (4.2) under the necessary condition (4.4). Note that the second term in (4.4) satisfies

$$\frac{\log(p-s)}{\log(1+\frac{a^2}{\sigma^2})} \asymp \frac{\sigma^2 \log(p-s)}{a^2} \quad \text{for } a/\sigma = \mathcal{O}(1). \tag{4.5}$$

In the general case, that is with no restrictions on the joint behavior of s, σ and a, the following sufficient condition for the ML decoder to achieve exact recovery (4.2) is given in Wainwright (2009a):

$$n = \Omega\left(s\log\left(\frac{p}{s}\right) \vee \frac{\sigma^2\log(p-s)}{a^2}\right). \tag{4.6}$$

One can check that, for $a/\sigma = \mathcal{O}(1/\sqrt{s})$, the second terms in (4.4) and in (4.6) are dominant, while for $a/\sigma = \Omega(1)$, the first terms are dominant. These remarks and (4.4) - (4.6) lead us to the following table of phase transitions for exact recovery in the sense of (4.2). We recall that this table, as well as the whole discussion in this subsection, deal only with the setting where both X and ξ are Gaussian.

SNR	Upper bound for ML	Lower bound
$a/\sigma = \mathcal{O}(1/\sqrt{s})$	$\frac{\sigma^2\log(p-s)}{a^2}$	
$a/\sigma = \mathcal{O}(1)$ and $a/\sigma = \Omega(1/\sqrt{s})$	$\frac{s\log(\frac{p}{s})}{\log(1+s\frac{a^2}{\sigma^2})} \vee \frac{\log(p-s)}{\log(1+\frac{a^2}{\sigma^2})}$	
$a/\sigma = \Omega(1)$	$s\log(\frac{p}{s})$	$\frac{s\log(p/s)}{\log(1+sa^2/\sigma^2)}$

Table 4.1: Phase transitions in Gaussian setting.

It remains an open question what is the exact phase transition for $a/\sigma = \Omega(1)$. We also note that, in the zone $a/\sigma = \mathcal{O}(1)$, the exact phase transitions in this table are attained by the ML decoder, which is not computable in polynomial time and requires the knowledge of s. Known polynomial time algorithms are shown to be optimal only in the regime $a/\sigma = \mathcal{O}(1/\sqrt{s})$. In Fletcher et al. (2009), it is shown that Lasso is sub-optimal compared to the ML decoder. For the regime $a^2/\sigma^2 = \mathcal{O}(\frac{\log(s)}{s})$ and $s \asymp p$, the ML decoder requires $n = \Omega(p)$ observations to achieve exact recovery, while polynomial time algorithms require $n = \Omega(p\log(p))$. In this regime, the ML decoder is optimal, cf. Table 1. In the regime of $a/\sigma = \Omega(1)$, there exists an algorithmic gap making the problem of exact recovery hard whenever the sample size satisfies $n \leq 2\sigma^2 s\log(p)$ Gamarnik and Zadik (2017). This implies that the sample size of the order $\frac{s\log(p/s)}{\log(1+sa^2/\sigma^2)}$ (the necessary condition 4.4 in the regime $a/\sigma = \Omega(1)$) is not sufficient for exact recovery via a computationally tractable method. Variable selection algorithms based on techniques from sparse graphs theory such as sparsification of the Gram matrix $X^\top X$ are suggested in Ji and Jin (2012), Jin et al. (2014) and Ke et al. (2014). In those papers, phase transitions are derived for the asymptotics where the sparsity s and the sample size n scale as power functions of the dimension p. In general, sufficient conditions for the ML decoder are less restrictive than conditions obtained for known polynomial time algorithms. A more complete overview of necessary and sufficient conditions for exact recovery defined in the form (4.2) for different models can be found in Aksoylar et al. (2017).

Contributions

The main contribution of this paper is a polynomial time algorithm that achieves exact recovery with respect to both criteria (4.2) and (4.3) under the same sufficient conditions

(4.6) as the ML decoder. An open question stated in Fletcher et al. (2009) is whether any computationally tractable algorithm can achieve a scaling similar to the ML decoder. This paper answers the question positively under rather general conditions. In Fletcher et al. (2009), a sufficient condition for exact recovery by a thresholding procedure is obtained in the form

$$n > \frac{8\sqrt{\|\beta\|_2^2 + \sigma^2}}{a^2} \log(p - s).$$

This condition cannot be satisfied uniformly on the class $\Omega_{s,a}^p$ since $\|\beta\|_2^2$ can be arbitrarily large. The selector $\hat{\eta}$ that we suggest here is defined by a two step algorithm, which computes at the first step the Square-Root SLOPE estimator $\hat{\beta}$ of β. At the second step, the components of $\hat{\eta}$ are obtained by thresholding of debiased estimators of the components of β based on the preliminary estimator $\hat{\beta}$.

We now proceed to the formal definition of this selection procedure. Split the sample (X_i, Y_i), $i = 1, \dots, n$, into two subsamples $\mathcal{D}_1$ and $\mathcal{D}_2$ with respective sizes n_1 and n_2, such that $n = n_1 + n_2$. For $k = 1, 2$, denote by $(X^{(k)}, Y^{(k)})$ the corresponding submatrices $X^{(k)} \in \mathbb{R}^{n_k \times p}$ and subvectors $Y^{(k)} \in \mathbb{R}^{n_k}$. The Square-Root SLOPE estimator based on the first subsample $(X^{(1)}, Y^{(1)})$ is defined as follows. Let $\lambda \in \mathbb{R}^p$ be a vector of tuning parameters

$$\lambda_j = A\sqrt{\frac{\log(\frac{2p}{j})}{n}}, \qquad j = 1, \dots, p,$$

for large enough constant $A > 0$. For any $\beta \in \mathbb{R}^p$, let $(\beta_1^*, \dots, \beta_p^*)$ be a non-increasing rearrangement of $|\beta_1|, \dots, |\beta_p|$. Consider

$$|\beta|_* = \sum_{j=1}^p \lambda_j \beta_j^*, \quad \beta \in \mathbb{R}^p,$$

which is a norm on $\mathbb{R}^p$, cf., e.g., Bogdan et al. (2015). The Square-Root SLOPE estimator is a solution of the convex minimization problem

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\frac{\|Y^{(1)} - X^{(1)}\beta\|}{\sqrt{n_1}} + 2|\beta|_* \right). \tag{4.7}$$

Note that this estimator does not depend on the parameters s, σ, and a. Details about the computational aspects and statistical properties of the Square-Root SLOPE estimator can be found in Derumigny (2018).

The suggested selector is defined as a binary vector

$$\hat{\eta}(X, Y) = (\hat{\eta}_1(X, Y), \dots, \hat{\eta}_p(X, Y)) \tag{4.8}$$

with components

$$\hat{\eta}_i(X, Y) = \mathbf{1}\left\{ \left| X_i^{(2)\top} \left(Y^{(2)} - \sum_{j \neq i} X_j^{(2)} \hat{\beta}_j \right) \right| > t(X_i^{(2)}) \|X_i^{(2)}\| \right\} \tag{4.9}$$

for $i = 1, \dots, p$, where $X_i^{(2)}$ denotes the ith column of matrix $X^{(2)}$. The threshold $t(\cdot)$ in (4.9) will be defined by different expressions, with a basic prototype of the form

$$t(u) = t_\sigma(u) = \frac{a\|u\|}{2} + \frac{\sigma^2 \log(\frac{p}{s} - 1)}{a\|u\|}, \qquad \forall u \in \mathbb{R}^{n_2}. \tag{4.10}$$

The decoder (4.8) - (4.9) is the core procedure of this paper. We show that it improves known sufficient conditions of exact recovery for methods realizable in polynomial time. We also show that it can be turned into a completely adaptive procedure (once the sufficient conditions are fulfilled) by suitably modifying the definition (4.10) of the threshold. Another advantage is that the decoder (4.8) - (4.9) can be generalized to sub-Gaussian design matrices X and to heavy-tailed noise. In Section 4.2, we study the non-asymptotic minimax Hamming distance risk, we derive a lower bound and prove that the decoder (4.8) - (4.9) is nearly optimal. Section 4.3 gives sufficient conditions for exact support recovery for the method (4.8) - (4.9). Section 4.4 is devoted to adaptivity to all parameters of the setting, while, in Section 4.5, we show how to extend all previous results to sub-Gaussian $X$and ξ. Finally, in Section 4.6, we give a robust version of our procedure when the noise ξ is heavy-tailed and the data are corrupted by arbitrary outliers.

4.2 Non-asymptotic bounds on the minimax risk

Here, as well as in Sections 4.3 and 4.4, we assume that all entries of X are i.i.d. standard Gaussian random variables and the noise $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$ is a standard Gaussian vector independent of X.

In this section, we present a non-asymptotic minimax lower bound on the Hamming risk of any selectors as well as non-asymptotic upper bounds for the two risks of selector (4.8) - (4.9). In several papers, lower bounds are derived using the Fano lemma in order to get necessary conditions of exact support recovery, i.e., the convergence of the minimax risk to 0. However, they do not give information about the rate of convergence. Our first aim in this section is to obtain an accurate enough lower bound characterizing the rate. The Fano lemma is too rough for this purpose and we use instead more refined techniques based on explicit Bayes risk calculation. Set

$$\psi_+(n,p,s,a,\sigma) = (p-s)\,\mathbf{P}\left(\sigma\varepsilon > t(\zeta)\right) + s\mathbf{P}\left(\sigma\varepsilon \geq a\|\zeta\| - t(\zeta)\right),$$

where ε is a standard Gaussian random variable, $\zeta \sim \mathcal{N}(0, \mathbb{I}_n)$ is a standard Gaussian random vector in $\mathbb{R}^n$ independent of ε, and $t(\cdot)$ is defined in (4.10).

The following minimax lower bound holds.

Theorem 4.2.1. *For any $a > 0$, $\sigma > 0$ and any integers n, p, s such that $s < p$ we have*

$$\forall s' \in (0, s], \quad \inf_{\tilde\eta} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta |\tilde\eta - \eta_\beta| \geq \frac{s'}{s}\left(\psi_+(n,p,s,a,\sigma) - 4se^{-\frac{(s-s')^2}{2s}}\right),$$

where $\inf_{\tilde\eta}$ denotes the infimum over all selectors $\tilde\eta$.

The proof of this theorem is given in Appendix 4.8. The idea is in reduction to considering the component-wise Bayes risk. Achieving the minimal value of the risk for each component leads to an equivalent of the oracle (non-realizable) decoder η^* with components

$$\eta_i^*(X,Y) = \mathbf{1}\left\{X_i^\top\left(Y - \sum_{j\neq i} X_j\beta_j\right) > t(X_i)\|X_i\|\right\}, \quad i = 1,\dots,p, \tag{4.11}$$

where $t(\cdot)$ is the threshold defined in (4.10). This decoder has a structure similar to (4.9). It selects the components by thresholding the random variables

$$\frac{X_i^\top \left(Y - \sum_{j\neq i} X_j \beta_j\right)}{\|X_i\|}. \tag{4.12}$$

Note that, under the model (4.1), the random variable (4.12) has the same distribution as

$$\beta_i \|X_i\| + \sigma \varepsilon_i,$$

where ε_i is a standard Gaussian random variable independent of $\|X_i\|$. Thus, conditionally on the design X, we are in the framework of variable selection in the normal means model, where the lower bound techniques developed in Butucea et al. (2018) can be applied to obtain the result.

Clearly, the oracle decoder η^* is not realizable since it depends on the unknown β. We do not know the rest of the components of β when we try to recover its ith component. Since the sensing matrix X is assumed Gaussian with i.i.d entries, it is straightforward to see that $\sum_{j\neq i} X_j\beta_j$ is a zero-mean Gaussian variable with variance not greater than $\|\beta\|^2$. Hence we can consider this term as an additive noise, but the fact that we cannot control $\|\beta\|$ means that the variance of the noise is not controlled neither. In order to get around this drawback, we plug in an estimator $\hat\beta$ instead of β in the oracle expression. This motivates the two-step selector defined in (4.8) - (4.9). At the first step, we use the Square-Root SLOPE estimator $\hat\beta$ based on the subsample $\mathcal{D}_1$. We have the following bound on the ℓ_2 error of the Square-Root SLOPE estimator.

Proposition 4.2.1. *Let $\hat\beta$ be the Square-Root SLOPE estimator defined in Section 4.1 with large enough $A > 0$. There exist positive constants C_0, C_1 and C_2 such that for all $\delta \in (0,1]$ and $n_1 > \frac{C_0}{\delta^2} s \log\left(\frac{ep}{s}\right)$ we have*

$$\sup_{|\beta|_0 \leq s} \mathbf{P}_\beta \left(\|\hat\beta - \beta\| \geq \delta\sigma \right) \leq C_1 \Big(\frac{s}{2p}\Big)^{C_2 s}.$$

This proposition is a special case of Proposition 4.5.1 below.

In what follows, for the sake of readability, we will write X and Y instead of $X^{(2)}$ and $Y^{(2)}$ since we will condition on the first subsample $\mathcal{D}_1$ and only use the second subsample $\mathcal{D}_2$ in our argument. We only need to remember that $\hat\beta$ is independent from the second sample of size n_2. With this convention, definition (4.9) involves now the random variables

$$\alpha_i := \frac{X_i^\top \left(Y - \sum_{j\neq i} X_j \hat\beta_j\right)}{\|X_i\|} = \beta_i \|X_i\| + \frac{1}{\|X_i\|} X_i^\top \Big(\sum_{j\neq i} X_j(\beta_j - \hat\beta_j) + \sigma\xi \Big) \tag{4.13}$$

for $i = 1, \dots, p$. Conditionally on $\hat\beta$ and X_i, the variable α_i has the same distribution as

$$\beta_i \|X_i\| + \Big(\sigma^2 + \sum_{j\neq i} |\beta_j - \hat\beta_j|^2\Big)^{\frac{1}{2}} \varepsilon, \tag{4.14}$$

where ε is a standard Gaussian random variable. Hence, considering α_i as new observations, we have a conditional normal means model, for which a natural procedure to

detect the non-zero components consists in comparing α_i to a threshold. Choosing the same threshold $t(\cdot)$ as in the lower bound of Theorem 4.2.1 leads to the selector (4.8) - (4.9).

Consider now a quantity close to ψ_+ given by the formula

$$\psi\left(n,p,s,a,\sigma\right)=\left(p-s\right)\mathbf{P}\left(\sigma\varepsilon>t\left(\zeta\right)\right)+s\mathbf{P}\left(\sigma\varepsilon>\left(a\|\zeta\|-t\left(\zeta\right)\right)_{+}\right)$$

where $t(u)=t_\sigma(u)$ is defined in (4.10). Note that

$$\psi\left(n,p,s,a,\sigma\right)\leq\psi_{+}\left(n,p,s,a,\sigma\right).$$

We have the following upper bound for the minimax risks of the selector (4.8) - (4.9).

Theorem 4.2.2. *Let n,p,a,σ be as in Theorem 4.2.1, let s be an integer such that $s\leq p/2$, and let $\hat{\eta}$ be the selector (4.8) - (4.9) with the threshold $t(\cdot)=t_{\sigma\sqrt{1+\delta^2}}(\cdot)$ defined in (4.10), with some $\delta\in(0,1]$. Then there exists a constant $C_0>0$ such that for all $n_1>\frac{C_0}{\delta^2}s\log\left(\frac{ep}{s}\right)$ we have*

$$\sup_{\beta\in\Omega_{s,a}^{p}}\mathbf{E}_\beta|\hat{\eta}-\eta_\beta|\leq 2\psi(n_2,p,s,a,\sigma\sqrt{1+\delta^2})+C_1p\Big(\frac{s}{2p}\Big)^{C_2s},$$

and

$$\sup_{\beta\in\Omega_{s,a}^{p}}\mathbf{P}_\beta\left(\hat{\eta}\neq\eta_\beta\right)\leq 2\psi(n_2,p,s,a,\sigma\sqrt{1+\delta^2})+C_1\Big(\frac{s}{2p}\Big)^{C_2s}.$$

Proof. Define the random event $\mathbb{A}=\{\|\hat{\beta}-\beta\|\leq\delta\sigma\}$, where $\hat{\beta}$ is based on the subsample $\mathcal{D}_1$. For any $\beta\in\Omega_{s,a}^{p}$, we have

$$\begin{aligned}\mathbf{E}_\beta\left[|\hat{\eta}-\eta_\beta|\ |\mathcal{D}_1\right] &= \sum_{i:\beta_i=0}\mathbf{E}_\beta\left[\hat{\eta}_i|\mathcal{D}_1\right]+\sum_{i:\beta_i\neq0}\mathbf{E}_\beta\left[1-\hat{\eta}_i|\mathcal{D}_1\right]\\ &= \sum_{i:\beta_i=0}\mathbf{P}_\beta(|\alpha_i|>t(X_i)|\mathcal{D}_1)+\sum_{i:\beta_i\neq0}\mathbf{P}_\beta(|\alpha_i|\leq t(X_i)|\mathcal{D}_1).\end{aligned}$$

Here, $t(X_i)\geq 0$ since $s\leq p/2$. Using the fact that, conditionally on $\hat{\beta}$ and X_i, the variable α_i has the same distribution as (4.14) we find that, for all i such that $\beta_i=0$,

$$\mathbf{P}_\beta(|\alpha_i|>t(X_i)|\mathcal{D}_1)\leq\mathbf{P}(\sigma_*|\varepsilon|>t(X_i)|\mathcal{D}_1)=2\mathbf{P}(\sigma_*\varepsilon>t(X_i)|\mathcal{D}_1)$$

where $\sigma_*=(\sigma^2+\|\hat{\beta}-\beta\|^2)^{1/2}$ and ε is a standard Gaussian random variable independent of $\|X_i\|$. Analogous argument and the fact that $|\beta_i|\geq a$ for all non-zero β_i lead to the bound

$$\mathbf{P}_\beta(|\alpha_i|\leq t(X_i)|\mathcal{D}_1)\leq\mathbf{P}(\sigma_*|\varepsilon|\geq a\|X_i\|-t(X_i)|\mathcal{D}_1)=2\mathbf{P}(\sigma_*\varepsilon\geq(a\|X_i\|-t(X_i))_+|\mathcal{D}_1)$$

valid for all i such that $\beta_i\neq0$. Therefore,

$$\mathbf{E}_\beta\left[|\hat{\eta}-\eta_\beta|\ |\mathcal{D}_1\right]\leq 2(p-s)\mathbf{P}(\sigma_*\varepsilon>t(\zeta)|\mathcal{D}_1)+2s\mathbf{P}(\sigma_*\varepsilon\geq(a\|\zeta\|-t(\zeta))_+|\mathcal{D}_1),\quad(4.15)$$

where $\zeta\sim\mathcal{N}\left(0,\mathbb{I}_{n_2}\right)$ is a standard Gaussian random vector in $\mathbb{R}^{n_2}$ independent of ε. Using this bound on the event $\mathbb{A}$ and taking expectations with respect to $\mathcal{D}_1$ yields

$$\mathbf{E}_\beta|\hat{\eta}-\eta_\beta|\leq 2\psi(n_2,p,s,a,\sigma\sqrt{1+\delta^2})+2p\mathbb{P}\left(\mathbb{A}^c\right).$$

For $\mathbf{P}_\beta(\hat{\eta} \neq \eta_\beta)$, we have an analogous bound where the factor p in the second term disappears since we intersect only once with the event $\mathbb{A}$ in the probability of wrong recovery, instead of doing it for p components. The theorem follows by applying Proposition 4.2.1. □

Remark 4.2.1. *As we will see in the next section, the term $p\left(\frac{s}{2p}\right)^{C_2 s}$ is small compared to ψ for large p. Hence, ψ, or the close quantity ψ_+, characterize the main term of the optimal rate of convergence. Uniformly on $\Omega^p_{s,a}$, no selector can reach a better rate of the minimax risk in asymptotical regime. The discrepancy between the upper and lower bounds comes from increasing the sample size by n_1, in order to estimate β (in the upper bound, the first argument of ψ_+ is the smaller sample size $n_2 < n$, which makes ψ_+ greater), and a higher variance $\sigma^2(1+\delta^2)$, even if we can make it very close to σ^2 by choosing δ.*

Remark 4.2.2. *Our choice of Square-Root SLOPE estimator $\hat{\beta}$ is motivated by the fact that it achieves the optimal rate of ℓ_2 estimation adaptively to s and σ, which will be useful in Section 4.4. Since in this section we do not consider adaptivity issues, we can also use as $\hat{\beta}$ the LASSO estimator with regularization parameter depending on both s and σ or the SLOPE estimator, for which the regularization parameter depends σ but not on s. Indeed, it follows from Bellec et al. (2018) that Proposition 4.2.1 holds when $\hat{\beta}$ is such a LASSO or a SLOPE estimator. Thus, Theorem 4.2.2 remains valid for these two estimators as well.*

The values α_i can be viewed as "de-biased" observations in high-dimensional regression. Other de-biasing schemes can be used, for example, the method considered in Section 4.6. The most popular de-biasing technique is based on the LASSO and would consider in our context $\alpha_i = \hat{\beta}^d_i$ where $\hat{\beta}^d_i$ are the components of the vector

$$\hat{\beta}^d = \hat{\beta}^L + \frac{1}{n} X^\top \left(Y - X\hat{\beta}^L\right).$$

and $\hat{\beta}^L$ is the LASSO estimator (see, for example, Javanmard and Montanari (2018) and the references therein). As in our case, this reduces the initial regression model to the mean estimation model (conditionally on $\hat{\beta}^L$), which is not exactly the normal means model but rather its approximation. Indeed, we may equivalently write

$$\hat{\beta}^d_i = \beta_i + \frac{X_i^\top}{n}\Big(\sum_{j\neq i} X_j(\beta_j - \hat{\beta}^L_j) + \sigma\xi\Big) + \left(1 - \frac{\|X_i\|^2}{n}\right)(\hat{\beta}^L_i - \beta_i).$$

The difference from (4.13) is in the fact that, conditionally on $\hat{\beta}^L$ and X_i, we have here a bias $\left(1 - \frac{\|X_i\|^2}{n}\right)(\hat{\beta}^L_i - \beta_i)$, and that there is no scaling by the norm of X_i. Note that scaling by the norm $\|X_i\|$ instead of n is crucial in our construction. It allows us to obtain in Theorem 4.2.2 the expression for the risk analogous to the lower bound of Theorem 4.2.1.

Finally, note that in parallel to our work, a study of a specific type of two-stage algorithms for variable selection in linear models is developed in Wang et al. (2017). Our results and the questions that we address here are significantly different.

4.3 Phase transition

Using the upper and lower bounds of Section 4.2, we can now study the phase transition, i.e., the necessary and sufficient conditions on the sample size to achieve exact recovery under the Hamming risk. A first lower bound is given by the following result.

Proposition 4.3.1. *Let n, p, s, a, σ be as in Theorem 4.2.1. If also $s \geq 6$ and $n \leq \frac{2\sigma^2 \log\left(\frac{p}{s}-1\right)}{a^2}$, there exists an absolute constant $c > 0$ such that*

$$\inf_{\tilde{\eta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\tilde{\eta} - \eta_\beta\right| \geq \left(c \vee \frac{s}{8}\left(1 - 16e^{-\frac{s}{8}}\right)\right),$$

where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors $\tilde{\eta}$.

Proof. We start by proving a lower bound on the function ψ_+. We have

$$\psi_+(n,p,s,a,\sigma) \geq s\mathbf{P}\left(\sigma\varepsilon \geq a\|\zeta\| - t(\zeta)\right) \geq s\mathbf{P}(\sigma\varepsilon \geq 0)\mathbf{P}\left(\mathbb{B}\right) = \frac{s}{2}\mathbf{P}\left(\mathbb{B}\right).$$

where $\mathbb{B} = \{a\|\zeta\| \leq t(\zeta)\}$. Since a chi-squared random variable with n degrees of freedom has a median smaller than n, we get under the conditions stated above that

$$\mathbf{P}\left(\mathbb{B}\right) = \mathbf{P}\left(\|\zeta\|^2 \leq \frac{2\sigma^2 \log(\frac{p}{s}-1)}{a^2}\right) \geq \frac{1}{2}.$$

Therefore, using Theorem 4.2.1 we get

$$\forall s' \in (0,s), \quad \inf_{\tilde{\eta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\tilde{\eta} - \eta_\beta\right| \geq s'\left(\frac{1}{4} - 4e^{-\frac{(s-s')^2}{2s}}\right).$$

Since $s \geq 6$ we have $4e^{-s/2} < \frac{1}{4}$. Hence,

$$\lim_{s' \to 0^+}\left(\frac{1}{4} - 4e^{-\frac{(s-s')^2}{2s}}\right) = \frac{1}{4} - 4e^{-s/2} > 0.$$

Thus, there exists $c > 0$ such that

$$\inf_{\tilde{\eta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\tilde{\eta} - \eta_\beta\right| \geq c.$$

By setting $s' = s/2$, we also get

$$\inf_{\tilde{\eta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\tilde{\eta} - \eta_\beta\right| \geq \frac{s}{8}\left(1 - 16e^{-\frac{s}{8}}\right).$$

The proposition follows. □

Proposition 4.3.1 implies that the condition $n \geq \frac{2\sigma^2 \log(\frac{p}{s}-1)}{a^2}$ is necessary to achieve exact recovery for the Hamming risk. We give now a more accurate necessary condition for the regime $a = \mathcal{O}(\sigma)$. This regime is the most interesting when we consider the asymptotic setting where a is decreasing.

Theorem 4.3.1. *Let n, p, s, a, σ be as in Theorem 4.2.1. If also $n > \frac{2\sigma^2 \log\left(\frac{p}{s}-1\right)}{a^2}$, $a < \sqrt{2}\sigma$, and $s < p/2$, then there exists an absolute constant $c > 0$ such that*

$$\inf_{\tilde{\eta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\tilde{\eta} - \eta_\beta\right| \geq c \sqrt{\frac{s^{7/4}(p-s)^{1/4}}{n \log\left(1 + \frac{a^2}{4\sigma^2}\right)}} \exp\left(-\frac{n}{2} \log\left(1 + \frac{a^2}{4\sigma^2}\right)\right) - 2se^{-\frac{s}{8}},$$

where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors $\tilde{\eta}$.

The proof of Theorem 4.3.1 is given in Appendix 4.8.

Corollary 4.3.1. *Let $s \geq 6$, $a < \sqrt{2}\sigma$, and let*

$$n < (1-\epsilon) \frac{\log(p-s) + 7\log(s)}{4 \log\left(1 + \frac{a^2}{4\sigma^2}\right)},$$

for some $\epsilon \in (0,1)$. Then, there exists $c > 0$ such that

$$\liminf_{p \to \infty} \inf_{\tilde{\eta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\tilde{\eta} - \eta_\beta\right| \geq c.$$

Proof. If $n \leq \frac{2\sigma^2 \log\left(\frac{p}{s}-1\right)}{a^2}$, then the result follows from Proposition 4.3.1. Now if $n > \frac{2\sigma^2 \log\left(\frac{p}{s}-1\right)}{a^2}$, then Theorem 4.3.1 yields

$$\inf_{\tilde{\eta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\tilde{\eta} - \eta_\beta\right| \geq c \left(\frac{\left(s^{7/4}(p-s)^{1/4}\right)^\epsilon}{(1-\epsilon)\log\left(s^{7/4}(p-s)^{1/4}\right)}\right)^{\frac{1}{2}} - 2se^{-\frac{s}{8}}.$$

As $1 \leq s < p$, we have $\lim_{p \to \infty} s^{7/4}(p-s)^{1/4} = \infty$. The result follows. □

Corollary 4.3.1 implies the following necessary condition for exact recovery under the Hamming risk:

$$n \geq \frac{\log(p-s) + 7\log(s)}{4 \log\left(1 + \frac{a^2}{4\sigma^2}\right)}.$$

We will show now that the upper bound on the minimax risk decreases exponentially with the sample size. This will allow us to show that the decoder (4.8) - (4.9) achieves exact recovery under the same conditions as the ML decoder.

Theorem 4.3.2. *Let $s \leq p/2$, n, p, a, σ be as in Theorem 4.2.1, and $a \leq \sigma$. Assume that for some $\delta \in (0,1]$ the following inequalities hold*

$$n_1 > \frac{C_0}{\delta^2} s \log\left(\frac{ep}{s}\right) \qquad \text{and} \qquad n_2 \geq \frac{4\sigma^2 \log\left(\frac{p}{s}-1\right)}{a^2}.$$

Let $\hat{\eta}$ be the selector as in Theorem 4.2.2. Then,

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left|\hat{\eta} - \eta_\beta\right| \leq 2\sqrt{s(p-s)} \exp\left(-\frac{n_2}{2} \log\left(1 + \frac{a^2}{4\sigma^2(1+\delta^2)}\right)\right) + se^{-\frac{n_2}{24}} + C_1 p \left(\frac{s}{2p}\right)^{C_2 s},$$

and

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{P}_\beta \left(\hat{\eta} \neq \eta_\beta\right) \leq 2\sqrt{s(p-s)} \exp\left(-\frac{n_2}{2} \log\left(1 + \frac{a^2}{4\sigma^2(1+\delta^2)}\right)\right) + se^{-\frac{n_2}{24}} + C_1 \left(\frac{s}{2p}\right)^{C_2 s}.$$

The proof of this theorem is given in Appendix 4.8.

We can notice that both types of errors decrease exponentially as the sample size increases to ∞.

Corollary 4.3.2. *Under the conditions of Theorem 4.3.2, if* $a \leq \sigma/\sqrt{3}$ *and* $n_2 \geq A\frac{\log(p-s)+\log(s)}{\log\left(1+\frac{a^2}{4\sigma^2(1+\delta^2)}\right)}$ *for some* $A > 1$*, we have*

$$\sup_{\beta\in\Omega^p_{s,a}} \mathbf{P}_\beta\left(\hat{\eta} \neq \eta_\beta\right) \leq 3\left(s(p-s)\right)^{\frac{1-A}{2}} + C_1\left(\frac{s}{2p}\right)^{C_2 s}.$$

Proof. Since $\log(p-s)+\log(s) \geq \log\left(\frac{p}{s}-1\right)$, and $\log(1+x) \leq x$, we get

$$n_2 \geq \frac{4\sigma^2 \log\left(\frac{p}{s}-1\right)}{a^2}.$$

Hence Theorem 4.3.2 applies. Moreover, since $a \leq \sigma/\sqrt{3}$ we also have

$$e^{-\frac{n_2}{24}} \leq \exp\left(-\frac{n_2}{2}\log\left(1+\frac{a^2}{4\sigma^2\left(1+\delta^2\right)}\right)\right).$$

We can conclude by using the lower bound on n_2 and the inequality $s \leq \sqrt{s(p-s)}$. □

As consequence of the last corollary, sufficient conditions for the selector (4.8) - (4.9) with threshold (4.10) to achieve exact recovery are as follows

$$n_1 > \frac{C_0}{\delta^2} s \log\left(\frac{ep}{s}\right) \quad \text{and} \quad n_2 > (1+\epsilon)\frac{\log(p-s)+\log(s)}{\log\left(1+\frac{a^2}{4\sigma^2(1+\delta^2)}\right)},$$

for some $\delta \in (0,1]$ and $\epsilon > 0$.

Comparing the rate of convergence in Corollary 4.3.2 to the rate for the ML decoder established in Rad (2011), we notice that they have similar form. Indeed, Rad (2011) proves the bound

$$\sup_{\beta\in\Omega^p_{s,a}} \mathbf{P}_\beta\left(\hat{\eta} \neq \eta_\beta\right) \leq s\left(\left(es(p-s)\right)^{-B^*} + \left(\frac{s}{e(p-s)}\right)^{B^* s}\right),$$

for some $B^* > 0$ and $s \leq p/2$.

It is interesting to compare these conditions with the best known in the literature (where only the risk (4.2) was studied). Using (4.5), we see that, in the zone $a/\sigma = \mathcal{O}(1)$, our sufficient condition for exact recovery has the form

$$n = \Omega\left(s\log\left(\frac{p}{s}\right) \vee \frac{\sigma^2\log(p-s)}{a^2}\right). \tag{4.16}$$

As follows from the discussion in the Introduction, this gives the exact phase transition in the zone $a/\sigma = \mathcal{O}(1)$, $a = \mathcal{O}(1/\sqrt{s})$, while in the zone $a/\sigma = \mathcal{O}(1)$, $a = \Omega(1/\sqrt{s})$, combination of the results of Wang et al. (2010) and Rad (2011) shows that the exact phase transition (realized by the ML decoder) is given by

$$n = \Omega\left(\frac{s\log(\frac{p}{s})}{\log(1+s\frac{a^2}{\sigma^2})} \vee \frac{\sigma^2\log(p-s)}{a^2}\right).$$

It remains an open question whether the improvement by the term $\log(1+s\frac{a^2}{\sigma^2})$ appearing here is achievable by computationally tractable methods. Our sufficient condition (4.16) is the same as for the ML decoder Wainwright (2009a), with the advantage that our selector can be computed in polynomial time. Nevertheless, the knowledge of parameters s, a and σ is required for the construction. This motivates us to derive, in the next section, adaptive variants of the proposed selector.

4.4 Nearly optimal adaptive procedures

In this section, we propose three adaptive versions of our selector. The first one assumes that we know only a and do not know s and σ, the second assumes only the knowledge of σ, and the third one is completely adaptive to all the parameters.

We first present the following a tail bound for the Student distribution that will be useful to derive the results.

Lemma 4.4.1. *Let Z be a Student random variable with k degrees of freedom. There exist constants $c, C > 0$ independent of k such that for all $b \geq 1/\sqrt{k}$ we have*

$$c\frac{(1+b^2)^{-\frac{k-1}{2}}}{\sqrt{k}b} \leq \mathbf{P}\left(|Z| \geq \sqrt{k}b\right) \leq C\frac{(1+b^2)^{-\frac{k-1}{2}}}{\sqrt{k}b}.$$

The proof of this lemma is given in Appendix 4.8.

The Square-Root SLOPE estimator $\hat{\beta}$ is adaptive to the sparsity parameter s and to the scale parameter σ. The dependence of the selector $\hat{\eta}$ defined in (4.8) - (4.9) on the parameters s, σ and a only appears in the definition of the threshold $t(\cdot)$. Hence, we will replace it by an adaptive threshold. In this section, we assume that n is an even integer and the sample splitting is done in two subsamples of equal sizes such that $n_1 = n_2 = n/2$. In Theorem 4.3.2, we have shown that the selector $\hat{\eta}$ defined in (4.8) - (4.9) with the threshold function

$$t(u) = \frac{a\|u\|}{2} + \frac{(1+\delta^2)\sigma^2 \log\left(\frac{p}{s} - 1\right)}{a\|u\|}, \quad \forall u \in \mathbb{R}^{n_2}, \tag{4.17}$$

achieves nearly optimal conditions of exact recovery. We now set a new threshold by simply dropping the second term in (4.17):

$$t(u) = \frac{a\|u\|}{2}, \quad \forall u \in \mathbb{R}^{n_2}. \tag{4.18}$$

Then, the procedure becomes adaptive to s and σ. The phase transition for this procedure is given by the following proposition.

Proposition 4.4.1. *Let n be an even integer and $2(1 \vee 1/C_2) \leq s < p$. Set $n_1 = n_2 = n/2$, and let the threshold $t(\cdot)$ be defined in (4.18). Then, the selector $\hat{\eta}$ defined in (4.8) - (4.9) achieves exact recovery under both risks (Hamming and support recovery) if*

$$n \geq 2\Big(C_0 s \log\left(\frac{ep}{s}\right) \vee \frac{2\log p}{\log\left(1+\frac{a^2}{8\sigma^2}\right)} + 1\Big).$$

Proof. Following the lines of the proof of Theorem 4.2.2 and choosing there $\delta = 1$ we get

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left| \hat{\eta} - \eta_\beta \right| \leq p\mathbf{P}\left(\sqrt{2}\sigma \left| \varepsilon \right| \geq \frac{a\|\zeta\|}{2} \right) + C_1 p \Big(\frac{s}{2p} \Big)^{C_2 s}, \tag{4.19}$$

where ε is a standard Gaussian random variable and $\zeta \sim \mathcal{N}(0, \mathbb{I}_{n_2})$ is a standard Gaussian random vector in $\mathbb{R}^{n_2}$ independent of ε. In order to prove exact recovery, we need to show that both terms on the right hand side of (4.19) vanish as p goes to infinity. We first consider the second term. Note that the function $t \mapsto \left(\frac{t}{2p}\right)^t$ is decreasing for $1 \leq t \leq p/2$. Thus, if $2(1 \vee 1/C_2) \leq s \leq p/2$ we have

$$p\Big(\frac{s}{2p} \Big)^{C_2 s} \leq p\Big(\frac{1 \vee 1/C_2}{p} \Big)^2 \to 0 \quad \text{as } p \to \infty,$$

while for $p/2 < s < p$,

$$p\Big(\frac{s}{2p} \Big)^{C_2 s} \leq p 2^{-C_2 p/2}.$$

Thus, to prove the proposition, it remains to show that the first term on the right hand side of (4.19) vanishes. Using the independence between ε and ζ, we have

$$\mathbf{P}\left(\sqrt{2}\sigma \left| \varepsilon \right| \geq \frac{a\|\zeta\|}{2} \right) = \mathbf{P}\left(|Z| \geq \frac{a\sqrt{n_2}}{2\sqrt{2}\sigma} \right),$$

where Z is a Student random variable with n_2 degrees of freedom. To bound the last probability, we use Lemma 4.4.1. Since $\log(1+x) \leq x, \forall x \geq 0$, the assumption on n_2 implies

$$n_2 > \frac{16\sigma^2 \log p}{a^2}.$$

In particular, since $p \geq 3$ we have $\frac{n_2 a^2}{8\sigma^2} \geq 1$. Thus, by Lemma 4.4.1,

$$\begin{aligned} p\mathbf{P}\left(\sqrt{2}\sigma \left| \varepsilon \right| \geq \frac{a\|\zeta\|}{2} \right) &\leq \frac{Cp\sigma}{a\sqrt{n_2}} \left(1 + \frac{a^2}{8\sigma^2} \right)^{-\frac{n_2 - 1}{2}} \\ &= \frac{C\sigma}{a\sqrt{n_2}} \exp\left(\log p - \frac{n_2 - 1}{2} \log\left(1 + \frac{a^2}{8\sigma^2} \right) \right) \\ &\leq \frac{C\sigma \sqrt{\log\left(1 + \frac{a^2}{8\sigma^2}\right)}}{a\sqrt{2\log p}} \end{aligned} \tag{4.20}$$

where we have used the condition $n_2 \geq \frac{2\log p}{\log\left(1+\frac{a^2}{8\sigma^2}\right)} + 1$. The expression in (4.20) tends to 0 as $p \to \infty$. This completes the proof. □

Proposition 4.4.1 shows that the condition

$$n = \Omega\left(s \log\left(\frac{ep}{s} \right) \vee \frac{\log p}{\log\left(1 + \frac{a^2}{8\sigma^2}\right)} \right)$$

is sufficient for exact recovery without knowing the sparsity parameter s.

We now turn to the case where both s and a are unknown. In Proposition 4.4.1, we have used the condition

$$n_2 > \frac{2\log p}{\log\left(1+\frac{a^2}{8\sigma^2}\right)},$$

which is equivalent to

$$a^2 > 8\sigma^2\left(p^{\frac{2}{n_2}}-1\right). \tag{4.21}$$

This inspires us to replace the threshold function $t(u) = a\|u\|/2$ considered in Proposition 4.4.1 by

$$t(u) = \sigma\sqrt{2\left(p^{\frac{2}{n_2}}-1\right)}\|u\|, \quad u \in \mathbb{R}^{n_2}. \tag{4.22}$$

Then, we get the following result analogous to Proposition 4.4.1.

Theorem 4.4.1. *Let $n \geq 4$ be an even integer and $2(1 \vee 1/C_2) \leq s < p$. Set $n_1 = n_2 = n/2$, and let the threshold $t(\cdot)$ be defined in (4.22). Then, the selector $\hat{\eta}$ defined in (4.8) - (4.9) achieves exact recovery under both risks (Hamming and support recovery) if $n \geq 2\left(C_0 s\log\left(\frac{ep}{s}\right) \vee \frac{2\log p}{\log\left(1+\frac{a^2}{8\sigma^2}\right)}\right)$.*

Proof. Acting as in the proof of Theorem 4.2.2 and choosing there $\delta = 1$ we get

$$\sup_{\beta\in\Omega^p_{s,a}} \mathbf{E}_\beta|\hat{\eta}-\eta_\beta| \leq p\mathbf{P}\left(\sqrt{2}\sigma|\varepsilon| > t(\zeta)\right) + s\mathbf{P}\left(\sqrt{2}\sigma|\varepsilon| \geq (a\|\zeta\| - t(\zeta))_+\right) + C_1 p\left(\frac{s}{2p}\right)^{C_2 s}$$

where ε is a standard Gaussian random variable and $\zeta \sim \mathcal{N}(0, \mathbb{I}_{n_2})$ is a standard Gaussian random vector in $\mathbb{R}^{n_2}$ independent of ε. Since $n_2 \geq \frac{2\log p}{\log\left(1+\frac{a^2}{8\sigma^2}\right)}$, we have (4.21), which implies $a\|\zeta\| \geq 2t(\zeta)$. Therefore,

$$\sup_{\beta\in\Omega^p_{s,a}} \mathbf{E}_\beta|\hat{\eta}-\eta_\beta| \leq 2p\mathbf{P}\left(\sqrt{2}\sigma|\varepsilon| \geq t(\zeta)\right) + C_1 p\left(\frac{s}{2p}\right)^{C_2 s}. \tag{4.23}$$

The second summand on the right hand side of (4.23) is treated in the same way as in Proposition 4.4.1. To bound the first summand, we note that due to (4.22),

$$\mathbf{P}\left(\sqrt{2}\sigma|\varepsilon| \geq t(\zeta)\right) = \mathbf{P}\left(|Z| \geq \sqrt{n_2\left(p^{\frac{2}{n_2}}-1\right)}\right)$$

where Z is a Student random variable with n_2 degrees of freedom. Using the inequalities $n_2\left(p^{\frac{2}{n_2}}-1\right) = n_2\left(\exp(2(\log p)/n_2)-1\right) \geq 2\log p$, $n_2 \geq C_0\log p$, and Lemma 4.4.1 we find

$$\mathbf{P}\left(|Z| \geq \sqrt{n_2\left(p^{\frac{2}{n_2}}-1\right)}\right) \leq \frac{p^{-1+1/n_2}}{\sqrt{2\log p}} \leq \frac{p^{-1}\exp(1/C_0)}{\sqrt{2\log p}}.$$

This implies that the first summand on the right hand side of (4.23) tends to 0 as $p \to \infty$. □

Thus, if only σ is known while a and s are not, we can achieve exact recovery under the same condition as for the ML decoder (which is not computationally tractable and depends on s). Next, we show that, replacing σ in (4.22) by a suitable estimator, we can render the procedure completely adaptive to all parameters of the problem.

Define $\hat{\sigma} > 0$ by

$$\hat{\sigma}^2 = \frac{1}{n_2} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij} \hat{\beta}_j \right)^2,$$

where $\hat{\beta}$ is the same Square-Root SLOPE estimator as in (4.9) and consider the threshold function

$$t(u) = \hat{\sigma} \sqrt{2 \left(p^{\frac{2}{n_2}} - 1 \right)} \|u\|, \quad \forall u \in \mathbb{R}^{n_2}. \tag{4.24}$$

We get the following result for the fully adaptive procedure corresponding to this threshold.

Theorem 4.4.2. *Let $n \geq 4$ be an even integer and $2(1 \vee 1/C_2) \leq s < p$. Set $n_1 = n_2 = n/2$, and let the threshold $t(\cdot)$ be defined in (4.24). Then, there exists a constant $\bar{C}_0 > 0$ such that the selector $\hat{\eta}$ defined in (4.8) - (4.9) achieves exact recovery under both risks (Hamming and support recovery) if $n \geq 2 \left(\bar{C}_0 s \log \left(\frac{ep}{s} \right) \vee \frac{2 \log p}{\log \left(1 + \frac{a^2}{16 \sigma^2} \right)} \right)$.*

Proof. Define the random event

$$\mathbb{B} = \left\{ \|\hat{\beta} - \beta\|^2 \leq \sigma^2 \right\} \cap \left\{ \left| \frac{\hat{\sigma}^2}{\|\hat{\beta} - \beta\|^2 + \sigma^2} - 1 \right| \leq \frac{1}{2} \right\}.$$

We have

$$\sup_{\beta \in \Omega_{s,a}} \mathbf{E}_\beta \left| \hat{\eta} - \eta_\beta \right| \leq \sup_{\beta \in \Omega_{s,a}} \mathbf{E}_\beta \left(\left| \hat{\eta} - \eta_\beta \right| \mathbf{1}\{\mathbb{B}\} \right) + p \sup_{\beta \in \Omega_{s,a}} \mathbf{P}_\beta \left(\mathbb{B}^c \right).$$

To control the second term on the right hand side, note that, conditionally on $\hat{\beta}$, the estimator $\hat{\sigma}^2$ has the same distribution as

$$\frac{\|\hat{\beta} - \beta\|^2 + \sigma^2}{n_2} \chi^2(n_2),$$

where $\chi^2(n_2)$ is a chi-squared random variable with n_2 degrees of freedom. We will use the following lemma, cf. Cavalier et al. (2002) or Lounici et al. (2011).

Lemma 4.4.2. *For any $N \geq 1$ and $t > 0$,*

$$\mathbf{P}\left(\left| \chi^2(N)/N - 1 \right| \geq t \right) \leq 2 \exp \left(-\frac{t^2 N}{4(1+t)} \right),$$

where $\chi^2(N)$ is a chi-squared random variable with N degrees of freedom.

From Lemma 4.4.2 with $t = 1/2$ and Proposition 4.2.1 we get

$$p \sup_{\beta \in \Omega_{s,a}} \mathbf{P}_\beta \left(\mathbb{B}^c \right) \leq C_1 p \left(\frac{s}{2p} \right)^{C_2 s} + 2pe^{-(n_2 - 1)/24}.$$

Here, $p\big(\frac{s}{2p}\big)^{C_2 s} \to 0$ as $p \to \infty$ (cf. the proof of Proposition 4.4.1), while $pe^{-n_2/24} \to 0$ as $p \to \infty$ provided that we choose $\bar{C}_0 > 24$.

To evaluate $\Gamma := \mathbf{E}_\beta\big(|\hat{\eta} - \eta_\beta| \mathbf{1}\{\mathbb{B}\}\big)$, we act similarly to the proof of Theorem 4.2.2. We have

$$\Gamma = \sum_{i:\beta_i=0} \mathbf{P}_\beta(\{|\alpha_i| > t(X_i)\} \cap \mathbb{B}) + \sum_{i:\beta_i\neq 0} \mathbf{P}_\beta(\{|\alpha_i| \le t(X_i)\} \cap \mathbb{B}).$$

Set $\sigma_* = \sqrt{\|\hat{\beta} - \beta\|^2 + \sigma^2}$. On the event $\mathbb{B}$, we have $\sigma_*^2 \le 2\hat{\sigma}^2 \le 3\sigma_*^2$ and $\sigma_*^2 \le 2\sigma^2$. The last inequality and the assumption on n_2 imply that $a \ge 2\sqrt{2}\sigma^*(p^{2/n_2} - 1)^{1/2}$. Using these remarks and the fact that, conditionally on $\hat{\beta}$ and X_i, the variable α_i has the same distribution as (4.14) we obtain, for all i such that $\beta_i = 0$,

$$\begin{aligned}\mathbf{P}_\beta(\{|\alpha_i| > t(X_i)\} \cap \mathbb{B}) &\le \mathbf{P}_\beta(\{|\alpha_i| > \sigma_*\|X_i\|(p^{2/n_2} - 1)^{1/2}\} \cap \mathbb{A}) \\ &\le \mathbf{P}_\beta(|\varepsilon| > \|X_i\|(p^{2/n_2} - 1)^{1/2}\}),\end{aligned}$$

where $\mathbb{A} = \big\{\|\hat{\beta} - \beta\|^2 \le \sigma^2\big\}$ and ε is a standard Gaussian random variable independent of $\|X_i\|$. Similarly, for all i such that $\beta_i \neq 0$ (and thus $|\beta_i| \ge a$) we have

$$\begin{aligned}\mathbf{P}_\beta(\{|\alpha_i| \le t(X_i)\} \cap \mathbb{B}) &\le \mathbf{P}_\beta(\{|\alpha_i| \le \sqrt{3}\sigma_*\|X_i\|(p^{2/n_2} - 1)^{1/2}\} \cap \mathbb{A}) \\ &\le \mathbf{P}_\beta(\sigma_*|\varepsilon| \ge a\|X_i\| - \sqrt{3}\sigma_*\|X_i\|(p^{2/n_2} - 1)^{1/2}) \\ &\le \mathbf{P}_\beta(|\varepsilon| \ge (2\sqrt{2} - \sqrt{3})\|X_i\|(p^{2/n_2} - 1)^{1/2}).\end{aligned}$$

Combining the above inequalities we find

$$\Gamma \le p\mathbf{P}\left(|Z| \ge (p^{2/n_2} - 1)^{1/2}\right) + s\mathbf{P}\left(|Z| \ge (2\sqrt{2} - \sqrt{3})(p^{2/n_2} - 1)^{1/2}\right), \qquad (4.25)$$

where Z is a Student random variable with n_2 degrees of freedom. Finally, we apply the same argument as in the proof of Theorem 4.4.1 to obtain that the right hand side of (4.25) vanishes as $p \to \infty$. □

4.5 Generalization to sub-Gaussian distributions

In this section, we generalize our procedure to the case where both the design (sensing) matrix X and the noise ξ are sub-Gaussian. Recall that, for given $\sigma > 0$, a random variable ζ is called σ-sub-Gaussian if

$$\mathbf{E}\exp(t\zeta) \le \exp(\sigma^2 t^2/2), \quad \forall t \in \mathbb{R}.$$

In particular, this implies that ζ is centered.

In this section, we assume that both X and ξ have i.i.d. sub-Gaussian entries, and as above, X is independent of ξ.

The estimation part of our procedure (cf. Proposition 4.2.1) extends to sub-Gaussian designs as follows.

Proposition 4.5.1. *Assume that the entries of matrix X are i.i.d σ_X-sub-Gaussian random variables, the entries of the noise ξ are i.i.d σ-sub-Gaussian random variables*

for some $\sigma > 0$, $\mathbf{E}(X_{ij}^2) = 1$ *for all entries* X_{ij} *of matrix* X, *and* X *is independent of* ξ. *Let* $\hat{\beta}$ *be the Square-Root SLOPE estimator defined in Section 4.1 with large enough* $A > 0$. *There exist constants* $C_0, C_1, C_2 > 0$ *that can depend only on* σ_X, *such that for all* $\delta \in (0, 1]$ *and* $n_1 > \frac{C_0}{\delta^2} s \log\left(\frac{ep}{s}\right)$ *we have*

$$\sup_{|\beta|_0 \le s} \mathbf{P}_\beta \left(\|\hat{\beta} - \beta\| \ge \delta\sigma \right) \le C_1 \Big(\frac{s}{2p}\Big)^{C_2 s}.$$

The proof of this proposition is based on combination of arguments from Bellec et al. (2018) and Comminges et al. (2018). It is given in Appendix 4.8.

We will also need the following lemma proved in Appendix 4.8.

Lemma 4.5.1. *Let* U, V *be two independent random vectors in* $\mathbb{R}^n$, *such that the entries of* U *are i.i.d. random variables and the entries of* V *are i.i.d.* σ*-sub-Gaussian random variables for some* σ. *Assume that* $\mathbf{E}(U_i^2) = 1$ *and* $\mathbf{E}(U_i^4) \le \sigma_1^4$ *for all components* U_i *of* U, *where* $\sigma_1 > 0$. *Then, for any* $t > 0$,

$$\mathbf{P}\left(\frac{\left|U^\top V\right|}{\|U\|^2} \ge t \right) \le 2 \exp\left(-\frac{nt^2}{8\sigma^2} \right) + \exp\left(-\frac{9n}{32\sigma_1^4} \right).$$

We are now ready to state a general result for sub-Gaussian designs.

Theorem 4.5.1. *Let the assumptions of Proposition 4.5.1 be satisfied. Let* $n \ge 4$ *be an even integer and* $2(1 \vee 1/C_2) \le s < p$. *Set* $n_1 = n_2 = n/2$, *and let the threshold* $t(\cdot)$ *be defined in (4.18). Then, there exists a constant* $C > 0$ *such that the selector* $\hat{\eta}$ *defined in (4.8) - (4.9) achieves exact recovery under both risks (Hamming and support recovery) if* $n \ge C\left(s \log\left(\frac{ep}{s}\right) \vee \frac{\sigma^2 \log p}{a^2} \right)$.

Proof. We act similarly to the proof of Theorem 4.2.2 where we set $\delta = 1$ and $t(X_i) = \frac{a}{2}\|X_i\|$. Then, for all i such that $\beta_i = 0$, we have

$$\mathbf{P}_\beta(|\alpha_i| > t(X_i) | \mathcal{D}_1) \le \mathbf{P}\left(\frac{\left|U^\top V\right|}{\|U\|^2} > \frac{a}{2} \Big| \mathcal{D}_1 \right),$$

where $U = X_i$ and $V = Y - \sum_{j \ne i} X_j \hat{\beta}_j = \sigma\xi + \sum_{j \ne i} X_j(\beta_j - \hat{\beta}_j)$. For fixed $\hat{\beta}$, the components of V are i.i.d. σ_*-sub-Gaussian with $\sigma_* = (\sigma^2 + \|\hat{\beta} - \beta\|^2)^{1/2}$. In particular, for fixed $\hat{\beta}$ on the event $\mathbb{A} = \{\|\hat{\beta} - \beta\|^2 \le \sigma^2\}$, they are $\sqrt{2}\sigma$-sub-Gaussian. Thus, from Lemma 4.5.1 we obtain that there exists an absolute constant $c > 0$ such that, for all i with $\beta_i = 0$,

$$\mathbf{P}_\beta(\{|\alpha_i| > t(X_i)\} \cap \mathbb{A}) \le 2 \exp\Big(- cn_2 \Big(\frac{a^2}{\sigma^2} \wedge 1 \Big) \Big).$$

The same bound holds for $\mathbf{P}_\beta(\{|\alpha_i| \le t(X_i)\} \cap \mathbb{A})$ for all i such that $\beta_i \ne 0$. The rest of the proof follows the same lines as the proof of Theorem 4.2.2 using Proposition 4.5.1 to evaluate $\mathbf{P}_\beta(\mathbb{A}^c)$. This yields the bound

$$\sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta |\hat{\eta} - \eta_\beta| \le 4p \exp\Big(- cn_2 \Big(\frac{a^2}{\sigma^2} \wedge 1 \Big) \Big) + C_1 p \Big(\frac{s}{2p} \Big)^{C_2 s}.$$

The second summand on the right hand side of this inequality vanishes as $p \to \infty$ as shown in the proof of Proposition 4.4.1. The second summand vanishes as $p \to \infty$ if $n_2 > c'\big(\frac{\sigma^2}{a^2} \vee 1\big) \log p$ for some $c' > 1/c$. We conclude the proof by noticing that $s \log\left(\frac{ep}{s}\right) \ge \log p$ for all $1 \le s \le p$. □

Theorem 4.5.1 shows that, with no restriction on the joint behavior of s, a and σ, a sufficient condition for exact recovery in the sub-Gaussian case is the same as in the Gaussian case:

$$n = \Omega\left(s \log\left(\frac{ep}{s}\right) \vee \frac{\sigma^2 \log p}{a^2}\right).$$

On the other hand, necessary conditions of exact recovery given in (4.4) are valid for any X with i.i.d centered entries satisfying $\mathbf{E}(X_{ij}^2) = 1$ and for Gaussian noise ξ Wang et al. (2010). It follows that, if under the assumptions of Theorem 4.5.1 the noise ξ is Gaussian, our selector achieves exact phase transition in the zone $a/\sigma = \mathcal{O}(1)$, $a = \mathcal{O}(1/\sqrt{s})$, while for other values of s, a and σ, it achieves the phase transition up to a logarithmic factor.

4.6 Robustness through MOM thresholding

In the previous section, we have shown that the suggested decoder succeeds for independent sub-Gaussian designs. In practice, the observations we have may be corrupted by some outliers, and the assumption of sub-Gaussian noise is not always relevant. This motivates us to introduce a robust version of this selector. In this section, we propose a selector that achieves similar properties as described above under weaker assumptions on the noise and in the presence of outliers.

Suppose that data are partitioned in two disjoint groups O and I, where $(\mathbf{x}_i, Y_i)_{i\in O}$ are outliers, that is arbitrary vectors with $\mathbf{x}_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$, and $(\mathbf{x}_i, Y_i)_{i\in I}$ are informative observations distributed as described below. Here, $|I| + |O| = n$.

We assume that the informative observations satisfy

$$Y_i = \mathbf{x}_i^\top \beta + \xi_i, \quad i \in I, \tag{4.26}$$

where $\beta \in \mathbb{R}^p$ is an unknown vector of parameters and $\xi_1, \dots, \xi_n$ are zero-mean i.i.d. random variables such that for some $q, \sigma > 0$ we have $\mathbf{E}(|\xi_i|^{2+q}) \le \sigma^{2+q}, i \in I$. We also assume that all components X_{ij} of vectors $\mathbf{x}_i$ are σ_X-sub-Gaussian i.i.d. random variables with zero mean and $\mathbf{E}(X_{ij}^2) = 1$. Here, $\sigma_X > 0$ is a constant. The conditions on the design can be further weakened but we consider sub-Gaussian designs for the sake of readability and also because such designs are of major interest in the context of compressed sensing. We also assume that $\xi = (\xi_1, \dots, \xi_n)$ is independent of $X = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top)^\top$.

In this section, we propose a selector based on median of means (MOM). The idea of MOM goes back to Nemirovskii and Yudin (1983), Jerrum et al. (1986), Alon et al. (1999). Our selector uses again sample splitting. We first construct a preliminary estimator $\hat{\beta}^*$ based on the subsample $\mathcal{D}_1$ and then we threshold debiased estimators of the components of β. These debiased estimators are constructed using both $\hat{\beta}^*$ and the second subsample $\mathcal{D}_2$. As a preliminary estimator, we take the MOM-SLOPE estimator of Lecué and Lerasle (2017), for which we have the following version of Proposition 4.2.1.

Proposition 4.6.1. *Let X and ξ satisfy the conditions stated above in this section. Then, there exist constants $c_0, c_1, c_2 > 0$ depending only on q and the sub-Gaussian constant σ_X such that the following holds. If $|O| \le c_0 s \log(ep/s) \le n_1/2$, then the*

MOM-SLOPE estimator $\hat{\beta}^*$ *defined in Lecué and Lerasle (2017) satisfies*

$$\sup_{|\beta|_0 \leq s} \mathbf{P}_\beta \left(\|\hat{\beta}^* - \beta\| \geq \sigma \right) \leq c_1 \left(\frac{s}{p} \right)^{c_2 s}.$$

The proof of this proposition is given in Appendix 4.8.

Proposition 4.6.1 holds uniformly over all outlier sets $|O|$ such that $|O| \leq c_1 s \log(ep/s)$, and uniformly over all distributions of ξ_i satisfying the assumptions of this section. Based on the fact that the MOM-SLOPE estimator satisfies Proposition 4.6.1, we will now present a robust version of our selector. We split our sample in two subsamples of size $n/2$ each. The first subsample is used to construct a pilot estimator, which is the MOM-SLOPE estimator or any other estimator $\hat{\beta}^*$ satisfying Proposition 4.6.1. Then, the selector is constructed based on this estimator $\hat{\beta}^*$ and on the second subsample. To simplify the notation, for the rest of this section we will consider that the size of the second subsample is n rather than $n/2$ and we have an estimator $\hat{\beta}^*$ satisfying Proposition 4.6.1 and independent from the second subsample.

Let $K = \lfloor c_3 \log(p) \rfloor$ be the number of blocks, where $c_3 > 0$ is an absolute constant large enough. Assume that $1 < K < n$. By extracting K disjoint blocks from the observation Y corresponding to the second subsample, we get K independent observations $(\mathbf{Y}^{(i)})_{1 \leq i \leq K}$, where $\mathbf{Y}^{(i)} \in \mathbb{R}^q$ and $q = \lfloor \frac{n}{K} \rfloor$. Each observation $\mathbf{Y}^{(i)}$ satisfies

$$\mathbf{Y}^{(i)} = \mathbf{X}^{(i)} \beta + \xi^{(i)},$$

where $X^{(i)}$ is a submatrix of X with rows indexed by the ith block. For $i = 1, \ldots, K$, consider the new observations

$$\mathbf{Z}^{(i)} = \frac{1}{q} \mathbf{X}^{(i)\top} \mathbf{Y}^{(i)} - \left(\frac{1}{q} \mathbf{X}^{(i)\top} \mathbf{X}^{(i)} - I_p \right) \hat{\beta}^*.$$

We denote by $Z_1^{(i)}, \ldots, Z_p^{(i)}$ the components of $\mathbf{Z}^{(i)}$. Consider the selector defined as a vector

$$\hat{\eta}(X, Y) = (\hat{\eta}_1(X, Y), \ldots, \hat{\eta}_p(X, Y)) \tag{4.27}$$

with components

$$\hat{\eta}_j(X, Y) = \mathbf{1} \left\{ |Med(Z_j)| > t \right\}, \quad j = 1, \ldots, p, \tag{4.28}$$

where $Med(Z_j)$ is the median of $Z_j^{(1)}, \ldots, Z_j^{(K)}$, and $t = c_4 \sigma \sqrt{\frac{\log p}{n}}$ with an absolute constant $c_4 > 0$. The next theorem shows that, when the noise has polynomial tails and contains a portion of outliers, the robust selector (4.27) - (4.28) achieves exact recovery under the same condition on the sample size as when the noise is Gaussian.

Theorem 4.6.1. *Let* X *and* ξ *satisfy the conditions stated at the beginning of this section. Then, there exist absolute constants* $c', c_3, c_4 > 0$ *and a constant* $C' > 0$ *depending only on* q *and on the sub-Gaussian constant* σ_X *such that the following holds. Let* $c' < s < p$. *Then, the selector given in (4.27) - (4.28) achieves exact recovery with respect to both risks (4.2) and (4.3) if* $n \geq C' \left(s \log(p/s) \vee \sigma^2 \frac{\log(p)}{a^2} \right)$ *and* $|O| < K/4$.

Proof. For all $i = 1, \dots, K$, we have

$$\mathbf{Z}^{(i)} = \beta + \varepsilon^{(i)},$$

where

$$\varepsilon^{(i)} = \left(\frac{1}{q}\mathbf{X}^{(i)\top}\mathbf{X}^{(i)} - I_p\right)\left(\beta - \hat{\beta}^*\right) + \frac{1}{q}\mathbf{X}^{(i)\top}\xi^{(i)}.$$

The random vectors $\varepsilon^{(1)}, \dots, \varepsilon^{(K)}$ are independent conditionally on $\hat{\beta}^*$. Let $\varepsilon_j^{(i)}$ denote the jth component of $\varepsilon^{(i)}$. Note that $Med(Z_j) = \beta_j + Med(\varepsilon_j)$, where $Med(\varepsilon_j)$ denotes the median of $\varepsilon_j^{(1)}, \dots, \varepsilon_j^{(K)}$. Choose $C' > 0$ large enough to guarantee that $a > 2t$. Then,

$$\begin{aligned}
\mathbf{E}_\beta |\hat{\eta} - \eta_\beta| &= \sum_{j:\beta_j \neq 0} \mathbf{P}_\beta \left(|Med(Z_j)| \leq t\right) + \sum_{j:\beta_j = 0} \mathbf{P}_\beta \left(|Med(Z_j)| > t\right) \\
&\leq \sum_{j:\beta_j \neq 0} \mathbf{P}_\beta \left(|Med(\varepsilon_j)| \geq a - t\right) + \sum_{j:\beta_j = 0} \mathbf{P}_\beta \left(|Med(\varepsilon_j)| > t\right) \\
&\leq p \sup_{j=1,\dots,p} \mathbf{P}_\beta \left(|Med(\varepsilon_j)| \geq t\right).
\end{aligned}$$

Consider the event $\mathbb{A}_* = \{\|\hat{\beta}^* - \beta\|^2 \leq \sigma^2\}$. The following lemma is proved in Appendix 4.8.

Lemma 4.6.1. *Under the conditions of Theorem 4.6.1 we have*

$$\sup_{j=1,\dots,p} \mathbf{P}_\beta \left(|Med(\varepsilon_j)| \geq t\right) \leq e^{-c_5 K} + \mathbf{P}_\beta(\mathbb{A}_*^c)$$

for some $c_5 > 0$.

From Lemma 4.6.1 and Proposition 4.6.1 we get

$$\sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta |\hat{\eta} - \eta_\beta| \leq p e^{-c_5 K} + p e^{-c_2 s \log(ep/s)}.$$

Since $K = \lfloor c_3 \log(p) \rfloor$, and $s \log(ep/s) \geq c' \log(ep/c')$ the result follows for $c_3, c' > 0$ chosen large enough. □

We see that sufficient conditions of exact recovery for the robust selector are of the same order as in the Gaussian case. If the risk is considered uniformly over all noise distributions under the conditions of this section, clearly the Gaussian noise is in this class. Hence, necessary conditions in the Gaussian case are also necessary for such a uniform risk over noise distributions. We have proved previously that, sufficient conditions for the selector (4.8) - (4.9) to achieve exact recovery are almost optimal in the Gaussian case. As a consequence, the selector (4.27) - (4.28) is almost optimal in this more general setting.

4.7 Conclusion

In this chapter, we proposed computationally tractable algorithms of variable selection that can achieve exact recovery under milder conditions than the ones known so far. Throughout different sections, we have investigated, respectively, the setting with Gaussian observations, sub-Gaussian observations, and heavy-tailed observations corrupted by arbitrary outliers. We have shown that the suggested selectors nearly achieve necessary conditions of exact recovery. For the Gaussian case, we obtained not only the conditions of exact recovery but also accurate upper and lower bounds on the minimax Hamming risk. Furthermore, we constructed a decoder, which is fully adaptive to all parameters of the problem and achieves exact recovery under almost the same sufficient conditions as in the case where sparsity s and the signal strength a and the noise level σ are known.

4.8 Appendix: Proofs

In order to prove Theorem 4.2.1, we use the following result from Butucea et al. (2018). Consider the set of binary vectors

$$A = \{\eta \in \{0,1\}^p : \ |\eta|_0 \le s\}$$

and assume that we are given a family $\{\mathbf{P}_\eta, \eta \in A\}$ where each $\mathbf{P}_\eta$ is a probability distribution on a measurable space $(\mathcal{X}, \mathcal{U})$. We observe X drawn from $\mathbf{P}_\eta$ with some unknown $\eta = (\eta_1, \dots, \eta_p) \in A$ and we consider the Hamming risk of a selector $\hat\eta = \hat\eta(X)$:

$$\sup_{\eta \in A} \mathbf{E}_\eta |\hat\eta - \eta|$$

where $\mathbf{E}_\eta$ is the expectation w.r.t. $\mathbf{P}_\eta$. We call the selector any estimator with values in $\{0,1\}^p$. Let π be a probability measure on $\{0,1\}^p$ (a prior on η). We denote by $\mathbb{E}_\pi$ the expectation with respect to π. Then the following result is proved in Butucea et al. (2018)

Theorem 4.8.1. *Butucea et al. (2018) Let π be a product on p Bernoulli measures with parameter s'/p where $s' \in (0, s]$. Then,*

$$\inf_{\hat\eta} \sup_{\eta \in A} \mathbf{E}_\eta |\hat\eta - \eta| \ge \inf_{\hat T \in [0,1]^p} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{i=1}^p |\hat T_i - \eta_i| - 4s' \exp\Big(- \frac{(s-s')^2}{2s}\Big), \tag{4.29}$$

where $\inf_{\hat\eta}$ is the infimum over all selectors and $\inf_{\hat T \in [0,1]^p}$ is the infimum over all estimators $\hat T = (\hat T_1, \dots, \hat T_p)$ with values in $[0,1]^p$.

Proof of Theorem 4.2.1. Let $\Theta(p, s, a)$ a subset of $\Omega^p_{s,a}$ defined as

$$\Theta(p, s, a) = \{\beta \in \Omega^p_{s,a} : \ \beta_i = a, \ \forall i \in S_\beta\}.$$

Since any $\beta \in \Theta(p, s, a)$ can be written as $\beta = a\eta_\beta$, there is a one-to-one correspondence between A and $\Theta(p, s, a)$. Hence,

$$\inf_{\hat\eta} \sup_{\eta \in A} \mathbf{E}_\eta |\hat\eta - \eta| = \inf_{\hat\eta} \sup_{\beta \in \Theta(p,s,a)} \mathbf{E}_\beta |\hat\eta - \eta_\beta|.$$

Using this remark and Theorem 4.8.1 we obtain that, for all $s' \in (0, s]$,

$$\inf_{\hat\eta} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta |\hat\eta - \eta_\beta| \geq \inf_{\hat T \in [0,1]^p} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{i=1}^p |\hat T_i(X,Y) - \eta_i| - 4s' \exp\Big(- \frac{(s-s')^2}{2s}\Big),$$

where π a product on p Bernoulli measures with parameter s'/p. Thus, to finish the proof it remains to show that

$$\inf_{\hat T \in [0,1]^p} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{i=1}^p |\hat T_i(X,Y) - \eta_i| \geq \frac{s'}{s} \psi_+(n,p,s,a,\sigma).$$

We first notice that

$$\begin{aligned}
\inf_{\hat T \in [0,1]^d} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{i=1}^p |\hat T_i(X,Y) - \eta_i| &\geq \sum_{i=1}^p \mathbb{E}_\pi \mathbf{E}_\eta \Big[\inf_{\hat T_i \in [0,1]} \mathbb{E}_\pi \mathbf{E}_\eta \Big(|\hat T_i(X,Y) - \eta_i| \big| \eta_{(-i)}, X\Big)\Big] \\
&\geq \sum_{i=1}^p \mathbb{E}_\pi \mathbf{E}_\eta \Big[\inf_{\tilde T_i \in [0,1]} \mathbb{E}_\pi \mathbf{E}_\eta \Big(|\tilde T_i(X,Y,\eta_{(-i)}) - \eta_i| \big| \eta_{(-i)}, X\Big)\Big] \\
&= \sum_{i=1}^p \mathbb{E}_\pi \mathbf{E}_\eta [L_i^*]
\end{aligned}$$

where $\eta_{(-i)}$ denotes $(\eta_j)_{j\neq i}$ and $L_i^* = L_i^*(\eta_{(-i)}, X)$ has the form

$$\begin{aligned}
L_i^* = \inf_{\tilde T_i \in [0,1]} \Big(&\frac{s'}{p} \int (1 - \tilde T_i(X,y,\eta_{(-i)})) \varphi_\sigma(y - aX_i - \sum_{j\neq i} a\eta_j X_j) \mathrm{d}y \\
&+ \Big(1 - \frac{s'}{p}\Big) \int \tilde T_i(X,y,\eta_{(-i)}) \varphi_\sigma(y - \sum_{j\neq i} a\eta_j X_j) \mathrm{d}y \Big). \qquad (4.30)
\end{aligned}$$

Here, φ_σ is the density of Gaussian distribution in $\mathbb{R}^n$ with i.i.d. zero-mean and variance σ^2 components. By the Bayesian version of the Neyman-Pearson lemma, the infimum in (4.30) is attained for $\tilde T_i = T_i^*$ given by the formula

$$T_i^*\left(X, Y, \eta_{(-i)}\right) = \mathbf{1}\left\{ \frac{(s'/p)\phi_\sigma(Y - aX_i - \sum_{j\neq i} a\eta_j X_j)}{(1 - s'/p)\phi_\sigma(Y - \sum_{j\neq i} a\eta_j X_j)} > 1 \right\}.$$

Equivalently,

$$T_i^* = \mathbf{1}\left\{ \frac{X_i^\top (Y - \sum_{j\neq i} a\eta_j X_j)}{\|X_i\|} > t(s', X_i) \right\},$$

where

$$t(s', X_i) = \frac{a\|X_i\|}{2} + \frac{\sigma^2 \log(\frac{p}{s'} - 1)}{a\|X_i\|}.$$

Hence,

$$L_i^* = \left(1 - \frac{s'}{p}\right) \mathbf{P}\left(\frac{X_i^T \sigma\xi}{\|X_i\|} > t(s', X_i)\right) + \frac{s'}{p} \mathbf{P}\left(- \frac{X_i^T \sigma\xi}{\|X_i\|} > a\|X_i\| - t(s', X_i)\right).$$

where ξ is a standard Gaussian random vector in $\mathbb{R}^n$ independent of X_i. Notice now that $\varepsilon := \frac{X_i^T \xi}{\|X_i\|}$ is a standard Gaussian random variable and it is independent of $\|X_i\|$ since $X_i \sim \mathcal{N}(0, \mathbb{I}_n)$. Combining the above arguments we find that

$$\inf_{\hat{T}\in[0,1]^p} \mathbb{E}_\pi \mathbf{E}_\eta \sum_{i=1}^{p} |\hat{T}_i(X,Y) - \eta_i| \geq \psi_+(n,p,s',a,\sigma).$$

We conclude the proof by using the fact that the function $u \to \frac{\psi_+(n,p,u,a,\sigma)}{u}$ is decreasing for $u > 0$ (cf. Butucea et al. (2018)), so that $\psi_+(n,p,s',a,\sigma) \geq \frac{s'}{s}\psi_+(n,p,s,a,\sigma)$. □

Proof of Theorem 4.3.1. In view of Theorem 4.2.1 with $s' = s/2$, it is sufficient to bound $\psi_+ = \psi_+(n,p,s,a,\sigma)$ from below. We have

$$\psi_+ \geq (p-s)\,\mathbf{P}\left(\sigma\varepsilon \geq t\left(\zeta\right)\right).$$

We will use the following bound for the tails of standard Gaussian distribution: For some $c' > 0$,

$$\forall y \geq 2/3, \qquad \mathbf{P}\left(\varepsilon \geq y\right) \geq c' \frac{\exp(-y^2/2)}{y}.$$

We also recall that the density f_n of a chi-squared distribution with n degrees of freedom has the form

$$f_n(t) = b_n t^{\frac{n}{2}-1} e^{-\frac{t}{2}}, \quad t > 0, \tag{4.31}$$

and $\lim\limits_{n\to\infty} \frac{b_{n+1}}{b_n}\sqrt{n+1} = 1$, so that for some $c'' > 0$ we have

$$\forall n \geq 1, \qquad b_{n+1} \geq c'' \frac{b_n}{\sqrt{n+1}}.$$

Combining the above remarks we get

$$\begin{aligned}
\psi_+ &\geq (p-s)\int_{2/3}^{\infty} \mathbf{P}\left(\varepsilon \geq \left(\frac{\sqrt{u}a}{2\sigma} + \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}}\right)\right) f_n(u)\mathrm{d}u \qquad (4.32)\\
&\geq c'(p-s)\int_{2/3}^{\infty} \frac{\exp\left(-\frac{1}{2}\left(\frac{\sqrt{u}a}{2\sigma} + \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}}\right)^2\right)}{\frac{\sqrt{u}a}{2\sigma} + \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}}} f_n(u)\mathrm{d}u\\
&\geq \frac{b_{n-1}c}{\sqrt{n}}\sqrt{s(p-s)}\int_{2/3}^{\infty} \frac{\exp\left(-\frac{u}{2}\left(1+\frac{a^2}{4\sigma^2}\right) - \frac{\sigma^2\log\left(\frac{p}{s}-1\right)^2}{2a^2u}\right)}{\frac{\sqrt{u}a}{2\sigma} + \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}}} u^{\frac{n}{2}-1}\mathrm{d}u,
\end{aligned}$$

where $c = c'c''$. Set

$$B = \int_1^{\infty} \frac{\exp\left(-\frac{v}{2} - \frac{\sigma^2\log\left(\frac{p}{s}-1\right)^2\left(1+\frac{a^2}{4\sigma^2}\right)}{2a^2v}\right)}{1 + \frac{\sigma^2\log\left(\frac{p}{s}-1\right)\left(1+\frac{a^2}{4\sigma^2}\right)}{a^2v}} v^{\frac{n-1}{2}-1}\mathrm{d}v.$$

Using the change of variable $v = u\left(1+\frac{a^2}{4\sigma^2}\right)$ and the assumptions of the theorem we get

$$\begin{aligned}
\psi_+ &\geq \frac{cb_{n-1}}{\sqrt{n}}\sqrt{s\,(p-s)}\,e^{-\frac{n}{2}\log\left(1+\frac{a^2}{4\sigma^2}\right)}\int_{\frac{2}{3}\left(1+\frac{a^2}{4\sigma^2}\right)}^{\infty}\frac{\exp\left(-\frac{v}{2}-\frac{\sigma^2\log\left(\frac{p}{s}-1\right)^2\left(1+\frac{a^2}{4\sigma^2}\right)}{2a^2v}\right)}{\frac{\sqrt{v}a}{2\sigma\sqrt{1+\frac{a^2}{4\sigma^2}}}+\frac{\sigma\log\left(\frac{p}{s}-1\right)\sqrt{1+\frac{a^2}{4\sigma^2}}}{a\sqrt{v}}}v^{\frac{n}{2}-1}\mathrm{d}v \\
&\geq cb_{n-1}B\sqrt{\frac{4\sigma^2\left(1+\frac{a^2}{4\sigma^2}\right)}{na^2}}\sqrt{s\,(p-s)}e^{-\frac{n}{2}\log\left(1+\frac{a^2}{4\sigma^2}\right)} \\
&\geq cb_{n-1}B\sqrt{\frac{s\,(p-s)}{n\log\left(1+\frac{a^2}{4\sigma^2}\right)}}e^{-\frac{n}{2}\log\left(1+\frac{a^2}{4\sigma^2}\right)},
\end{aligned}$$

where the second inequality uses the condition $a \leq \sqrt{2}\sigma$ to guarantee that $\frac{2}{3}\left(1+\frac{a^2}{4\sigma^2}\right) \leq 1$, while the last inequality uses the fact that $(1+x)\log(1+x) \geq x,\ \forall x \geq 0$. To finish the proof, we need to bound $b_{n-1}B$ from below. We have

$$\begin{aligned}
B &\geq \int_n^{\infty}\frac{\exp\left(-\frac{v}{2}-\frac{\sigma^2\log\left(\frac{p}{s}-1\right)^2\left(1+\frac{a^2}{4\sigma^2}\right)}{2a^2v}\right)}{1+\frac{\sigma^2\log\left(\frac{p}{s}-1\right)\left(1+\frac{a^2}{4\sigma^2}\right)}{a^2v}}v^{\frac{n-1}{2}-1}\mathrm{d}v \\
&\geq \frac{\exp\left(-\frac{\sigma^2\log\left(\frac{p}{s}-1\right)^2\left(1+\frac{a^2}{4\sigma^2}\right)}{2a^2n}\right)}{1+\frac{\sigma^2\log\left(\frac{p}{s}-1\right)\left(1+\frac{a^2}{4\sigma^2}\right)}{a^2n}}\frac{\int_n^{\infty}f_{n-1}(u)\mathrm{d}u}{b_{n-1}}.
\end{aligned}$$

The last inequality is due to the fact that the function $x \to \frac{e^{-\frac{c}{x}}}{1+\frac{1}{x}}$ is increasing for $x > 0$, for any fixed $c > 0$. Since $n > \frac{2\sigma^2\log\left(\frac{p}{s}-1\right)}{a^2}$ and $a^2 < 2\sigma^2$, we deduce from the last display that

$$b_{n-1}B \geq \frac{4}{7}\exp\left(-\frac{3}{8}\log\left(\frac{p}{s}-1\right)\right)\int_n^{\infty}f_{n-1}(u)\mathrm{d}u.$$

Proposition 3.1 from Inglot (2010) implies that, for some absolute constant $c > 0$,

$$\int_n^{\infty}f_{n-1}(u)\mathrm{d}u > c$$

(indeed, n is very close to the median of a chi-squared random variable with $n-1$ degrees of freedom). Combining the above inequalities we obtain

$$\psi_+ \geq C\sqrt{\frac{s^{7/4}(p-s)^{1/4}}{n\log\left(1+\frac{a^2}{4\sigma^2}\right)}}e^{-\frac{n}{2}\log\left(1+\frac{a^2}{4\sigma^2}\right)}.$$

□

Proof of Theorem 4.3.2. In view of Theorem 4.2.2, it is sufficient to bound from above the expression

$$\psi\,(n,p,s,a,\sigma) = (p-s)\,\mathbf{P}\left(\sigma\varepsilon \geq t\,(\zeta)\right) + s\mathbf{P}\left(\sigma\varepsilon \geq (a\|\zeta\| - t\,(\zeta))_+\right).$$

Introducing the event $\mathbb{D} = \{a\|\zeta\| \geq t(\zeta)\}$ we get

$$\mathbf{P}\left(\sigma\varepsilon \geq (a\|\zeta\| - t(\zeta))_+\right) \leq \mathbf{P}\left(\{\sigma\varepsilon \geq a\|\zeta\| - t(\zeta)\} \cap \mathbb{D}\right) + \frac{1}{2}\mathbf{P}\left(\mathbb{D}^c\right).$$

Using the assumption on n_2 we obtain

$$\mathbf{P}\left(\mathbb{D}^c\right) = \mathbf{P}\left(\|\zeta\|^2 \leq \frac{2\sigma^2 \log(\frac{p}{s} - 1)}{a^2}\right) \leq \mathbf{P}\left(\|\zeta\|^2 \leq \frac{n_2}{2}\right).$$

Here, $\|\zeta\|^2$ is a chi-squared random variable with n_2 degrees of freedom. Lemma 4.4.2 implies

$$\frac{1}{2}\mathbf{P}\left(\mathbb{D}^c\right) \leq e^{-\frac{n_2}{24}}.$$

Thus, to finish the proof it remains to show that

$$(p-s)\,\mathbf{P}\left(\sigma\varepsilon \geq t(\zeta)\right) + s\mathbf{P}\left(\{\sigma\varepsilon \geq a\|\zeta\| - t(\zeta)\} \cap \mathbb{D}\right) \leq 2\sqrt{s(p-s)}e^{-\frac{n_2}{2}\log\left(1+\frac{a^2}{4\sigma^2}\right)}.$$

The bound $\mathbf{P}\left(\varepsilon \geq y\right) \leq e^{-\frac{y^2}{2}}$, $\forall y > 0$, on the tail of standard Gaussian distribution yields

$$\begin{aligned}
(p-s)\,\mathbf{P}\left(\sigma\varepsilon \geq t(\zeta)\right) &\leq (p-s)\int_0^\infty \frac{e^{-\frac{1}{2}\left(\frac{\sqrt{u}a}{2\sigma} + \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}}\right)^2}}{1 + \frac{\sqrt{u}a}{2\sigma} + \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}}} f_{n_2}(u)\mathrm{d}u \\
&\leq b_{n_2}\sqrt{s\,(p-s)}\int_0^\infty e^{-\frac{u}{2}\left(1+\frac{a^2}{4\sigma^2}\right)} u^{\frac{n_2}{2}-1}\mathrm{d}u,
\end{aligned}$$

where $f_{n_2}(\cdot)$ is the density of chi-squared distribution with n_2 degrees of freedom and b_{n_2} is the corresponding normalizing constant, cf. (4.31). Using again the bound $\mathbf{P}\left(\varepsilon \geq y\right) \leq e^{-\frac{y^2}{2}}$, $\forall y > 0$, and the inequality

$$\frac{\sqrt{u}a}{2\sigma} - \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}} \geq 0, \qquad \forall u \geq \frac{2\sigma^2\log\left(\frac{p}{s}-1\right)}{a^2},$$

we get

$$\begin{aligned}
s\mathbf{P}\left(\{\sigma\varepsilon \geq a\|\zeta\| - t(\zeta)\} \cap \mathbb{D}\right) &\leq s\int_{\frac{2\sigma^2\log\left(\frac{p}{s}-1\right)}{a^2}}^\infty e^{-\frac{1}{2}\left(\frac{\sqrt{u}a}{2\sigma} - \frac{\sigma\log\left(\frac{p}{s}-1\right)}{a\sqrt{u}}\right)^2} f_{n_2}(u)\mathrm{d}u \\
&\leq b_{n_2}\sqrt{s\,(p-s)}\int_0^\infty e^{-\frac{u}{2}\left(1+\frac{a^2}{4\sigma^2}\right)} u^{\frac{n_2}{2}-1}\mathrm{d}u.
\end{aligned}$$

The change of variable, $v = u\left(1 + \frac{a^2}{4\sigma^2}\right)$ yields

$$\begin{aligned}
b_{n_2}\int_0^\infty e^{-\frac{u}{2}\left(1+\frac{a^2}{4\sigma^2}\right)} u^{\frac{n_2}{2}-1}\mathrm{d}u &= b_{n_2}e^{-\frac{n_2}{2}\log\left(1+\frac{a^2}{4\sigma^2}\right)}\int_0^\infty e^{-\frac{v}{2}} v^{\frac{n_2}{2}-1}\mathrm{d}v \\
&= e^{-\frac{n_2}{2}\log\left(1+\frac{a^2}{4\sigma^2}\right)}.
\end{aligned}$$

That concludes the proof. □

Proof of Lemma 4.4.1. Recall that the density of a Student random variable Z with k degrees of freedom is given by:

$$f_Z(t) = c_k^* \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad t \in \mathbb{R},$$

where $c_k^* > 0$ satisfies

$$\lim_{k\to\infty} c_k^* = \sqrt{2\pi}. \tag{4.33}$$

Define, for $t > 0$,

$$g(t) = -c_k^* t^{-1} \left(1 + \frac{t^2}{k}\right)^{-\frac{k-1}{2}}.$$

It is easy to check that the derivative of g has the form

$$g'(t) = \left(1 + \frac{1}{t^2}\right) f_Z(t).$$

Hence, for all $b \geq 1/\sqrt{k}$,

$$-2g(\sqrt{k}b) = 2\int_{\sqrt{k}b}^{\infty} g'(t)\mathrm{d}t \leq \mathbf{P}\left(|Z| \geq \sqrt{k}b\right) \leq 4\int_{\sqrt{k}b}^{\infty} g'(t)\mathrm{d}t = -4g(\sqrt{k}b).$$

The lemma follows since, in view of (4.33), there exist two positive constants c and C such that $c \leq c_k^* \leq C$ for all $k \geq 1$. □

Proof of Lemma 4.5.1. It is not hard to check that the random variable $\frac{|u^\top V|}{\|u\|}$ is σ-sub-Gaussian for any fixed $u \in \mathbb{R}^n$. Also, any σ-sub-Gaussian random ζ variable satisfies $\mathbf{P}(|\zeta| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$ for all $t > 0$. Therefore, we have the following bound for the conditional probability:

$$\mathbf{P}\left(\frac{|U^\top V|}{\|U\|} \geq t\|U\| \,\middle|\, U\right) \leq 2e^{-\frac{t^2\|U\|^2}{2\sigma^2}}, \quad \forall\, t > 0.$$

This implies

$$\begin{aligned}\mathbf{P}\left(\frac{|U^\top V|}{\|U\|^2} \geq t\right) &\leq 2\mathbf{E}\left[e^{-\frac{t^2\|U\|^2}{2\sigma^2}} \mathbf{1}\left\{\|U\| \geq \sqrt{n}/2\right\}\right] + \mathbf{P}\left(\|U\| \leq \sqrt{n}/2\right) \\ &\leq 2e^{-\frac{nt^2}{8\sigma^2}} + \mathbf{P}\left(\|U\| \leq \sqrt{n}/2\right). \end{aligned} \tag{4.34}$$

To bound the last probability, we apply the following inequality (Wegkamp, 2003, Proposition 2.6).

Lemma 4.8.1. *Let $Z_1, Z_2, \ldots, Z_n$ be independent, nonnegative random variables with $\mathbf{E}(Z_i) = \mu_i$ and $\mathbf{E}(Z_i^2) \leq v^2$. Then, for all $x > 0$,*

$$\mathbf{P}\Big(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu_i) \leq -x\Big) \leq e^{-\frac{nx^2}{2v^2}}.$$

Using this lemma with $Z_i = U_i^2$, $\mu_i \equiv 1$, $x = 3/4$, and $v^2 = \sigma_1^4$ we find

$$\mathbf{P}\left(\|U\| \le \sqrt{n}/2\right) \le e^{-\frac{9n}{32\sigma_1^4}},$$

which together with (4.34) proves the lemma. □

Proof of Proposition 4.5.1. Under the assumptions of the proposition, the columns of matrix X have the covariance matrix $\mathbb{I}_p$. Without loss of generality, we may assume that this covariance matrix is $\frac{1}{2}\mathbb{I}_p$ and replace σ by $\frac{\sigma}{\sqrt{2}}$. We next define the event

$$\mathbb{A} = \{\text{the design matrix } X \text{ satisfies the } WRE(s, 20) \text{ condition}\},$$

where the WRE condition is defined in Bellec et al. (2018). It is easy to check that the assumptions of Theorem 8.3 in Bellec et al. (2018) are fulfilled, with $\Sigma = \frac{1}{2}\mathbb{I}_p$, $\kappa = \frac{1}{2}$ and $n_1 \ge C_0 s \log(2p/s)$ for some $C_0 > 0$ large enough. Using Theorem 8.3 in Bellec et al. (2018) we get

$$\mathbf{P}\left(\mathbb{A}^c\right) \le 3e^{-C' s \log 2p/s},$$

for some $C' > 0$. Now, in order to prove the proposition, we use the bound

$$\mathbf{P}_\beta\left(\|\hat\beta - \beta\|^2 \ge \sigma^2\delta^2\right) \le \mathbf{P}_\beta\left(\left\{\|\hat\beta - \beta\|^2 \ge \sigma^2\delta^2\right\} \cap \mathbb{A}\right) + \mathbf{P}\left(\mathbb{A}^c\right).$$

Under the assumption $n_1 \ge C_0 s \log(ep/s)/\delta^2$, we have

$$\mathbf{P}_\beta\left(\left\{\|\hat\beta - \beta\|^2 \ge \sigma^2\delta^2\right\} \cap \mathbb{A}\right) \le \mathbf{P}_\beta\left(\left\{\|\hat\beta - \beta\|^2 \ge C_0\sigma^2 \frac{s \log ep/s}{n_1}\right\} \cap \mathbb{A}\right).$$

By choosing C_0 large enough, and using Proposition 4 from Comminges et al. (2018) we get that, for some $C'' > 0$,

$$\mathbf{P}_\beta\left(\left\{\|\hat\beta - \beta\|^2 \ge \sigma^2\delta^2\right\} \cap \mathbb{A}\right) \le C''\left(e^{-s\log(2p/s)/C''} + e^{-n_1/C''}\right).$$

Recalling that $n_1 \ge C_0 s \log(2p/s)$ and comibinig the above inequalities we obtain the result of the proposition with $C_1 = 2C'' + 3$ and $C_2 = C' \wedge 1/C'' \wedge C_0/C''$. □

Proof of Proposition 4.6.1. We apply Theorem 6 in Lecué and Lerasle (2017). Thus, it is enough to check that items 1-5 of Assumption 6 in Lecué and Lerasle (2017) are satisfied. Item 1 is immediate since $|I| = n_1 - |O| \ge n_1/2$, and $|O| \le c_0 s \log(ep/s)$. To check item 2, we first note that the random variable $\mathbf{x}_1^\top t$ is $\|t\|\sigma_X$-sub-Gaussian for any $t \in \mathbb{R}^p$. It follows from the standard properties of sub-Gaussian random variables (Vershynin, 2012, Lemma 5.5) that, for some $C > 0$,

$$\left(\mathbf{E}|\mathbf{x}_1^\top t|^d\right)^{1/d} \le C\|t\|\sqrt{d}, \quad \forall t \in \mathbb{R}^p, \forall d \ge 1.$$

On the other hand, since the elements of $\mathbf{x}_1$ are centered random variables with variance 1,

$$\left(\mathbf{E}|\mathbf{x}_1^\top t|^2\right)^{1/2} = \|t\|, \quad \forall t \in \mathbb{R}^p. \tag{4.35}$$

Combining the last two displays proves item 2. Item 3 holds since we assume that $\mathbf{E}(|\xi_i|^{q_0}) \le \sigma^{q_0}, i \in I$, with $q_0 = 2 + q$. To prove item 4, we use (4.35) and the fact that, for some $C > 0$,

$$\mathbf{E}|\mathbf{x}_1^\top t| \ge C\|t\|, \quad \forall t \in \mathbb{R}^p,$$

due to Marcinkiewicz-Zygmund inequality (Petrov, 1995, page 82). Finally we have that, for some $c > 0$,

$$\mathrm{Var}(\xi_1 \mathbf{x}_1^\top t) \le \mathbf{E}[\xi_1^2]\, \mathbf{E}\big[(\mathbf{x}_1^\top t)^2\big] \le c\mathbf{E}\big[(\mathbf{x}_1^\top t)^2\big], \quad \forall t \in \mathbb{R}^p.$$

Thus, all conditions of Theorem 6 in Lecué and Lerasle (2017) are satisfied. Application of this theorem yields the result. □

Proof of Lemma 4.6.1. We first prove that for all $i \in I$ and $1 \le j \le p$,

$$\mathbf{E}_\beta\big((\varepsilon_j^{(i)})^2 \mathbf{1}\{\mathbb{A}_*\}\big) \le C\frac{K\sigma^2}{n}, \tag{4.36}$$

where $C > 0$ depends only on the sub-Gaussian constant σ_X. Indeed, the components of $\varepsilon^{(i)}$ have the form

$$\varepsilon_j^{(i)} = \left(\frac{1}{q}\|\mathbf{X}_j^{(i)}\|^2 - 1\right)(\beta_j - \hat\beta_j^*) + \frac{1}{q}\sum_{k\neq j} \mathbf{X}_j^{(i)\top}\mathbf{X}_k^{(i)}(\beta_k - \hat\beta_k^*) + \frac{1}{q}\mathbf{X}_j^{(i)\top}\xi,$$

where $\mathbf{X}_j^{(i)}$ is the jth column of $\mathbf{X}^{(i)}$. Conditioning first on $\mathbf{X}_j^{(i)}$, we get

$$\begin{aligned}\mathbf{E}_\beta\big((\varepsilon_j^{(i)})^2\mathbf{1}\{\mathbb{A}_*\}\big) &\le 2\mathbf{E}_\beta\left[\left(\|\hat\beta^* - \beta\|^2\left(\frac{1}{q}\|\mathbf{X}_j^{(i)}\|^2 - 1\right)^2 + \frac{1}{q^2}\left(\|\hat\beta^* - \beta\|^2 + \sigma^2\right)\|\mathbf{X}_j^{(i)}\|^2\right)\mathbf{1}\{\mathbb{A}_*\}\right] \\ &\le 2\sigma^2\mathbf{E}\left[\left(\frac{1}{q}\|\mathbf{X}_j^{(i)}\|^2 - 1\right)^2 + \frac{2}{q^2}\|\mathbf{X}_j^{(i)}\|^2\right].\end{aligned}$$

Since $\mathbf{E}(X_{kl}^2) = 1$ for all k and l, we have $\mathbf{E}\|\mathbf{X}_j^{(i)}\|^2 = q$. Furthermore, $\mathbf{E}(X_{kl}^4) \le \bar C$ where $\bar C$ depends only on the sub-Gaussian constant σ_X. Using these remarks we obtain from the last display that

$$\mathbf{E}_\beta\big((\varepsilon_j^{(i)})^2\mathbf{1}\{\mathbb{A}_*\}\big) \le \frac{2(\bar C + 2)\sigma^2}{q}.$$

As $q = \lfloor n/K \rfloor$ this yields (4.36).

Next, the definition of the median immediately implies that

$$\{|Med(\varepsilon_j)| \ge t\} \subseteq \left\{\sum_{i=1}^K \mathbf{1}_{\{|\varepsilon_j^{(i)}|\ge t\}} \ge \frac{K}{2}\right\}, \quad \forall t > 0.$$

It follows that

$$\begin{aligned}\mathbf{P}_\beta\left(|Med(\varepsilon_j)| \ge t\right) &\le \mathbf{P}_\beta\left(\{|Med(\varepsilon_j)| \ge t\} \cap \mathbb{A}_*\right) + \mathbf{P}(\mathbb{A}_*^c) \\ &\le \mathbf{P}_\beta\left(\left\{\sum_{i=1}^K \mathbf{1}_{\{|\varepsilon_j^{(i)}|\ge t\}} \ge \frac{K}{2}\right\} \cap \mathbb{A}_*\right) + \mathbf{P}_\beta(\mathbb{A}_*^c) \\ &\le \mathbf{P}_\beta\left(\sum_{i=1}^K \mathbf{1}_{\{|\varepsilon_j^{(i)}|\ge t\}\cap\mathbb{A}_*} \ge \frac{K}{2}\right) + \mathbf{P}_\beta(\mathbb{A}_*^c).\end{aligned}$$

Since the number of outliers $|O|$ does not exceed $\lfloor K/4 \rfloor$ there are at least $K' := K - \lfloor K/4 \rfloor$ blocks that contain only observations from I. Without loss of generality, assume that these blocks are indexed by $1, \dots, K'$. Hence

$$\mathbf{P}_\beta(|Med(\varepsilon_j)| \geq t) \leq \mathbf{P}_\beta\left(\sum_{i=1}^{K'} \mathbf{1}_{\{|\varepsilon_j^{(i)}| \geq t\} \cap \mathbb{A}_*} \geq \frac{K}{4} \right) + \mathbf{P}_\beta(\mathbb{A}_*^c). \tag{4.37}$$

Note that using (4.36) we have, for all $i = 1, \dots, K'$,

$$\mathbf{P}_\beta\left(\{|\varepsilon_j^{(i)}| \geq t\} \cap \mathbb{A}_* \right) \leq \mathbf{E}_\beta\big((\varepsilon_j^{(i)})^2 \mathbf{1}\{\mathbb{A}_*\}\big)/t^2 \leq \frac{CK\sigma^2}{t^2 n} \leq \frac{1}{5}.$$

The last inequality is granted by a choice of large enough constant c_4 in the definition of t. Thus, introducing the notation $\zeta_i = \mathbf{1}_{\{|\varepsilon_j^{(i)}| \geq t\} \cap \mathbb{A}_*}$ we obtain

$$\begin{aligned} \mathbf{P}_\beta\left(\sum_{i=1}^{K'} \mathbf{1}_{\{|\varepsilon_j^{(i)}| \geq t\} \cap \mathbb{A}_*} \geq \frac{K}{4} \right) &\leq \mathbf{P}_\beta\left(\sum_{i=1}^{K'} (\zeta_i - \mathbf{E}_\beta(\zeta_i)) \geq \frac{K}{4} - \frac{K'}{5} \right) \\ &\leq \mathbf{P}_\beta\left(\sum_{i=1}^{K'} (\zeta_i - \mathbf{E}_\beta(\zeta_i)) \geq \frac{K}{20} \right) \leq e^{-c_5 K} \end{aligned} \tag{4.38}$$

where the last inequality is an application of Hoeffding's inequality. Combining (4.37) and (4.38) proves the lemma.

□

Chapter 5

Interplay of minimax estimation and minimax support recovery under sparsity

In this chapter, we study a new notion of scaled minimaxity for sparse estimation in high-dimensional linear regression model. We present more optimistic lower bounds than the one given by the classical minimax theory and hence improve on existing results. We recover sharp results for the global minimaxity as a consequence of our study. Fixing the scale of the signal-to-noise ratio, we prove that the estimation error can be much smaller than the global minimax error. We construct a new optimal estimator for the scaled minimax sparse estimation. An optimal adaptive procedure is also described.

Based on Ndaoud (2019): Ndaoud, M. (2019). Interplay of minimax estimation and minimax support recovery under sparsity. *ALT 2019*.

5.1 Introduction

Statement of the problem

Assume that we observe the vector of measurements $Y \in \mathbb{R}^p$ satisfying

$$Y = \beta + \sigma\xi \tag{5.1}$$

where $\beta \in \mathbb{R}^p$ is the unknown signal, $\sigma > 0$ and the noise $\xi \sim \mathcal{N}(0, \mathbb{I}_p)$ is a standard Gaussian vector. Here, $\mathbb{I}_p$ denotes the $p \times p$ identity matrix. This model is a specific case of the more general model where $Y \in \mathbb{R}^n$ satisfies

$$Y = X\beta + \sigma\xi \tag{5.2}$$

where $X \in \mathbb{R}^{n\times p}$ is a given design or sensing matrix, and the noise is independent of X. Model (5.1) corresponds to the orthogonal design. In this chapter, we mostly focus on model (5.1). We denote by $\mathbf{P}_\beta$ the distribution of Y in model (5.1) or of (Y, X) in model (5.2), and by $\mathbf{E}_\beta$ the corresponding expectation.

We consider the problem of estimating the vector β. We will also explore its relation to the problem of recovering the support of β, that is the set S_β of non-zero components of β. For an integer $s \leq p$, we assume that β is s-sparse, that is it has at most s non-zero

components. We also assume that these components cannot be arbitrarily small. This motivates us to define the following set $\Omega^p_{s,a}$ of s-sparse vectors:

$$\Omega^p_{s,a} = \{\beta \in \mathbb{R}^p : \quad |\beta|_0 \leq s \quad \text{and} \quad |\beta_i| \geq a, \quad \forall i \in S_\beta\},$$

where $a \geq 0$, β_i are the components of β for $i = 1, \ldots, p$, and $|\beta|_0$ denotes the number of non-zero components of β. The value a characterizes the scale of the signal. In the rest of the paper, we will always denote by β the vector to estimate, while $\hat{\beta}$ will denote the corresponding estimator. In our setting, we do not constrain $\hat{\beta}$ to be sparse. Let us denote by ϕ the scaled minimax risk

$$\phi(s, a) = \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|^2 \right),$$

where the infimum is taken over all possible estimators $\hat{\beta}$. It is easy to check that ϕ is increasing with respect to s and decreasing with respect to a. Note that, for Y following model (5.1), the global minimax error over $\mathbb{R}^p$ is given by

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{R}^p} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|^2 \right) = \sigma^2 p.$$

The previous equality is achieved for $s = p$ and $a = 0$ through the naive estimator $\hat{\beta} = Y$. Under the sparsity assumption, the previous result can be improved. In the seminal paper Donoho et al. (1992), it is shown that the global sparse minimax estimation error has asymptotics:

$$\inf_{\hat{\beta}} \sup_{|\beta|_0 \leq s} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|^2 \right) = 2\sigma^2 s \log(p/s)(1 + o(1)) \quad \text{as } \frac{s}{p} \to 0. \tag{$*$}$$

Inspecting the proof of the minimax lower bound, one can see that ($*$) is achieved for

$$a = \sigma \sqrt{2 \log(p/s)}(1 + o(1)).$$

We may also notice that the global sparse minimax estimation error is more optimistic than the global over $\mathbb{R}^p$. In this chapter, we present an even more optimistic solution inspired by a notion of scaled minimax sparse estimation given by ϕ. By doing so, we recover the global sparse estimation by taking the supremum over all a. In the rest of the paper, we will always denote by $SMSE$ the quantity ϕ.

It is well known that minimax lower bounds are pessimistic. The worst case is usually specific to a critical region. Hence, a minimax optimal estimator can be good globally but may not be optimal outside of the critical region. By studying the quantity ϕ for fixed sparsity, we will emphasize this phenomenon.

An optimistic lower bound for estimation of s-sparse vectors is given by $\sigma^2 s$ and can be achieved when the support of vector β is known. We say that an estimator $\hat{\beta}$ achieves *exact estimation* in model (5.1) if

$$\lim_{s/p \to 0} \frac{\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|^2 \right)}{\sigma^2 s} = 1.$$

We also say that estimator β achieves *exact support recovery* in model (5.1) if

$$\lim_{p\to\infty} \sup_{\beta\in\Omega^p_{s,a}} \mathbf{P}_\beta\left(S_{\hat\beta} = S_\beta\right) = 1,$$

where the asymptotics are considered as $p \to \infty$ when all other parameters of the problem (namely, s, a, σ) depend on p. In this chapter, we shed some light on the relation between exact support recovery and exact estimation. Specifically, we give an answer to the following questions that motivate the present work.

- How pessimistic is the result (∗)? Can we do any better by fixing the scale value a?
- Is exact support recovery necessary to achieve exact estimation?
- Can we achieve minimax optimality with respect to SMSE adaptively to the scale value a?

In the dense regime where $s \asymp p$, the minimax estimation error is of order $\sigma^2 p$ independently of a. Hence, in the rest of the paper, we focus on the regime where $\frac{s}{p} = o(1)$. All the proofs are deferred Appendix 5.6.

Notation. In the rest of this paper we use the following notation. For given sequences a_n and b_n, we say that $a_n = \mathcal{O}(b_n)$ (resp $a_n = \Omega(b_n)$) when $a_n \le cb_n$ (resp $a_n \ge cb_n$) for some absolute constant $c > 0$. We write $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, $\|\mathbf{x}\|$ is the Euclidean norm of $\mathbf{x}$, and $\mathbf{x}^\top\mathbf{y}$ the corresponding inner product. For $q \ge 1$, and $\mathbf{x} \in \mathbb{R}^p$, we denote by $\|\mathbf{x}\|_q$ the l_q norm of $\mathbf{x}$. For a matrix X, we denote by X_j its jth column. For $x, y \in \mathbb{R}$, we denote by $x \vee y$ the maximum of x and y and we set $x_+ = x \vee 0$. For $q \ge 1$ and ξ a centered Gaussian random variable with variance σ^2, we denote by σ_q the quantity $\mathbf{E}(|\xi|^q)^{1/q}$. The notation $\mathbf{1}(\cdot)$ stands for the indicator function. We denote by C and C_q positive constants where the second one depends on q for some $q \ge 1$.

Related literature

The literature on minimax sparse estimation in high-dimensional linear regression (for both random and orthogonal design) is very rich and its complete overview falls beyond the format of this paper. We mention here only some recent results close to our work. All sharp results are considered in the regime $\frac{s}{p} \to 0$.

- In Bellec et al. (2018), the authors show that SLOPE estimator, which is defined in Bogdan et al. (2015), is minimax optimal for estimation under sparsity constraint in model (5.2), as long as X satisfies some general conditions. This result is non-asymptotic.
- Bellec (2018) proves that the minimax estimation rate of convex penalized destimators cannot be improved for sparse vectors, even when the scale parameter a is large. This fact is mainly due to the bias caused by convex penalization as it is the case for LASSO and SLOPE estimators.

- In Su and Candes (2016), it is shown that SLOPE is asymptotically minimax optimal on $\{|\beta|_0 \leq s\}$ giving the asymptotic optimal estimation error $2\sigma^2 s \log \frac{p}{s}$ in model (5.1). In model (5.2), where X has i.i.d standard Gaussian entries and under the asymptotic condition $\frac{s \log p}{n} \to 0$, SLOPE gives the asymptotic optimal error $2\frac{\sigma^2}{n} s \log \frac{p}{s}$, cf. Su and Candes (2016). Both results are asymptotically minimax optimal and adaptive to the sparsity level s.

- Wu and Zhou (2013) show that the penalized least squares estimator with a penalty that grows like $2\sigma^2 s \log \frac{p}{s}$, is asymptotically minimax optimal on $\{|\beta|_0 \leq s\}$ under additional assumptions on s and p.

Main contribution

Inspired by the related literature, the present work is also motivated by the following questions.

- In model (5.1), the proof of lower bounds uses a worst case vector with non-zero components that scale as $\sigma\sqrt{2 \log \frac{p}{s}}$ in order to get the best lower bound. In other words, the worst case happens for a specific vector β. Can we do better far from this vector?

- One of the popular approaches is to recover the support of a sparse vector and then estimate this vector on the obtained support. In this case the error of estimation is of order $s\sigma^2$ and is the best one can hope to achieve. Is it necessary to recover the true support in order to get this error? This is an important question that we address in this chapter.

- If the answer to the previous question is negative, can we propose an algorithm that would be optimal in the sense of SMSE, practical and adaptive?

The main contribution in this chapter is a sharp study of the minimax risk ϕ. What is more, we study a more general quantity given by

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right), \tag{5.3}$$

for any $q \geq 1$. We give lower bounds and corresponding upper bounds for (5.3). We show that in the interesting regime $\frac{s}{p} = o(1)$, our lower and upper bounds match not only in the rate but also in the constant up to a factor 4 under a mild condition on sparsity. As a result of our study, we recover two interesting phase transitions when $\frac{s}{p} = o(1)$.

The first one is that there are basically two regimes in estimation. For $a \leq \sigma\sqrt{2\log(p/s)}(1-\epsilon)$ and $\epsilon > 0$, the asymptotic SMSE is $2s\sigma^2 \log(p/s)$. This regime is called the hard recovery regime, where we prove that the error is due to misspecification in recovering the support. Alternatively, for $a \geq (1+\epsilon)\sigma\sqrt{2\log(p/s)}$, the error is of order $s\sigma^2$. This regime is presented as the hard estimation regime. In this regime, we can recover a good fraction of the support but still have to pay for the estimation on the support. Hence, and surprisingly, the SMSE is almost piece-wise constant as a function of a. This shows

that the sparse minimax risk can be made much smaller once we get far from some critical region.

Another contribution of this paper is a new phase transition related to sparsity. In Butucea et al. (2018), it is shown that a necessary condition to achieve exact recovery is given by

$$a \geq \sigma\sqrt{2\log(p-s)} + \sigma\sqrt{2\log(s)}.$$

To achieve exact estimation, a necessary condition is

$$a \geq \sigma\sqrt{2\log(p/s-1) + 2\log\log(p/s-1)} + \sigma\sqrt{2\log\log(p/s-1)}.$$

Hence exact recovery is not necessary for exact estimation. In fact, when $s \gg \log(p)$ then exact estimation is easier and when $s \ll \log(p)$ exact recovery becomes easier. This shows that there is no direct implication of exact recovery on exact estimation, hence the latter task should be considered as a separate problem.

Finally, one more contribution of this paper is adaptivity. We give an optimal adaptive variant of our procedure, that achieves the sparse minimax optimal rate and whenever exact estimation is possible achieves it as well. By doing so, our procedure improves on the existing literature. In fact, Lasso is known to have an unavoidable bias of order $\sigma^2 s\log(p/s)$ even on the class $\Omega^p_{s,a}$, cf. Bellec (2018). We show that our procedure is better in the sense that it gets rid of the bias whenever it is possible.

5.2 Towards more optimistic lower bounds for estimation

In several papers, lower bounds for minimax risk are derived using the Fano lemma. These lower bounds are usually far from being sharp in the non-asymptotic setting. We establish, in this section, non-asymptotic lower bounds on the minimax risk based on some revisited two-hypothesis testing techniques.

We derive two lower bounds for the SMSE. The scaled error of estimation of sparse vectors can be decomposed into two types of error. A first one based on the error of estimation when the true support S_β is known and a second one is given by the error of recovery of the true support when the vector components are known but not the support. For this purpose, we prove first a general lower bound for constrained minimax sparse estimation.

In the next theorem, we reduce the constrained minimax risk over all estimators to a Bayes risk with arbitrary prior measure π on $\mathbb{R}^p$ and give a bound on the difference between the two risks. This result is true in a general setup, non necessarily for Gaussian models. For a particular choice of measure π, we provide an explicit bound of the remainder term.

Consider the set of vectors $\Theta_{s,a} \subseteq \mathbb{R}^p$, and assume that we are given a family $\{P_\beta, \beta \in \Theta_{s,a}\}$ where each P_β is a probability distribution on a measurable space $(\mathcal{X},\mathcal{U})$. We observe Y drawn from P_β with some unknown $\beta \in \Theta_{s,a}$ and we consider the risk of an estimator $\hat\beta = \hat\beta(Y)$:

$$\sup_{\beta\in\Theta_{s,a}} \mathbf{E}_\beta \|\hat\beta - \beta\|_q^q$$

where $\mathbf{E}_\beta$ is the expectation with respect to P_β. Let π be a probability measure on $\mathbb{R}^p$ (a prior on β). We denote by $\mathbb{E}_\pi$ the expectation with respect to π.

Theorem 5.2.1. *For any $s < p$, $q \geq 1$ and any probability measure π on $\mathbb{R}^p$, there exists $C_q > 0$ such that*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Theta_{s,a}} \mathbf{E}_\beta \|\hat{\beta} - \beta\|_q^q \geq \inf_{\hat{T} \in \mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |\hat{T}_j(Y) - \beta_j|^q - C_q \, \mathbb{E}_\pi \left[\left(\mathbf{E}(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta_{s,a}) \right], \tag{5.4}$$

where $\beta^A := \beta.\mathbf{1}(\beta \in \Theta_{s,a}) = (\beta_1 \mathbf{1}(\beta \in \Theta_{s,a}), \ldots, \beta_p \mathbf{1}(\beta \in \Theta_{s,a}))$, $\inf_{\hat{\beta}}$ is the infimum over all estimators and $\inf_{\hat{T} \in \mathbb{R}^p}$ is the infimum over all estimators $\hat{T}(Y) = (\hat{T}_1(Y), \ldots, \hat{T}_p(Y))$ with values in $\mathbb{R}^p$.

Theorem 5.2.1 is valid in a very general setting. We present now specific lower bounds in the general model of linear regression. Assume that $Y \in \mathbb{R}^n$ follows model (5.2), where X is a deterministic design. The following lemma is useful to get more precise lower bounds in model (5.2). It is based on the simple observation that under independent prior distributions of the entries of β the oracle estimator of a given component does not depend on the rest of the components.

Lemma 5.2.1. *Assume that Y satisfies model (5.2) with a deterministic design X. Then*

$$\inf_{\hat{T} \in \mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |\hat{T}_j(X, Y) - \beta_j|^q \geq \sum_{j=1}^p \inf_{\hat{T}_j \in \mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_{\beta_j} |\hat{T}_j(X_j, \tilde{Y}_j) - \beta_j|^q,$$

where $\tilde{Y}_j = Y - \sum_{i \neq j} X_i \beta_i = \beta_j X_j + \sigma \xi$.

Using the previous lemma, we are now ready to give two sharp lower bounds for the SMSE. A first one supposed to capture the error of estimation when the support is known, while the second one handles the case where the support is not known.

Theorem 5.2.2. *Assume that Y follows model (5.2) with a deterministic design X. For any $a > 0$, $q \geq 1$ and $s < p$ we have*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq \sigma_q^q \max_{|S|=s} \sum_{i \in S} \frac{1}{\|X_i\|_2^q}.$$

In order to derive the next lower bound, we define the quantity Ψ introduced in Ndaoud and Tsybakov (2018) in the context of support recovery:

$$\Psi(p, s, a, \sigma, X) := \sum_{j=1}^p \left(\frac{s}{p} \mathbf{P}(\sigma\varepsilon \geq (a - t_j(a)) \|X_j\|) + (1 - \frac{s}{p}) \mathbf{P}(\sigma\varepsilon \geq t_j(a) \|X_j\|) \right),$$

where ε is standard Gaussian random variable and

$$t_j(a) := \frac{a}{2} + \frac{\sigma^2 \log\left(\frac{p}{s} - 1\right)}{a \|X_j\|^2}, \quad \forall j = 1, \ldots, p.$$

Theorem 5.2.3. *Assume that Y follows model (5.2) with deterministic design X. For any $a > 0$, $q \geq 1$ and $s < p$ we have*

$$\forall s' \in (0, s), \quad \inf_{\hat{\beta}} \sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq a^q \frac{s'}{s} \left(\frac{1}{2^q} \Psi(p, s, a, \sigma, X) - 2se^{-\frac{(s-s')^2}{2s}} \right).$$

The proof is based on arguments similar to Butucea et al. (2018). Assume now that we are under model (5.1) and set

$$\psi\left(p, s, a, \sigma\right) := (p-s)\,\mathbf{P}\left(\sigma\varepsilon > t\left(a\right)\right) + s\mathbf{P}\left(\sigma\varepsilon > a - t(a)\right),$$

where ε is a standard Gaussian random variable and

$$t\left(a\right) := \frac{a}{2} + \frac{\sigma^2 \log\left(\frac{p}{s} - 1\right)}{a}. \tag{5.5}$$

The minimax Hamming loss for model (5.1) was studied in Butucea et al. (2018), where it was shown that it is very linked to ψ. One may notice that, under model (5.1), $\Psi(p, s, a, \sigma, \mathbb{I}_p) = \psi(p, s, a, \sigma)$. We define now the following estimation rate

$$\Phi(a) := \begin{cases} a^q \psi(s, p, a, \sigma) \vee \sigma^q_q s & \text{if } a \geq t^*, \\ s\sigma^q \left(2\log(\frac{p}{s} - 1)\right)^{\frac{q}{2}} & \text{else,} \end{cases}$$

where

$$t^* = \sigma\sqrt{2\log\frac{p}{s} - 1}. \tag{5.6}$$

The next proposition is a consequence of previous theorems and shows the link between the minimax Hamming loss and the minimax estimation risk.

Proposition 5.2.1. *Assume that Y follows model (5.1). For any $a > 0$, $q \geq 1$, $s < p/2$ and $s \geq 8q\log\log(p)$, we have*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|^q_q \right) \geq C_q \Phi(a),$$

where $C_q > 0$.

Remark 5.2.1. *The mild condition $s = \Omega(\log\log p)$ is an artifact of the proof of the lower bound. We believe that this condition can be removed or further relaxed.*

A more careful proof of the previous result can lead us to $C_q = (1 + o(1))$ as $\frac{s}{p} \to 0$. We omit the proof of this, since we give a more accurate result in the next section. Analyzing the lower bound of Proposition 5.2.1, it turns out that the minimax rate $\sigma^2 s \log\left(p/s\right)$, for $q = 2$, cannot be improved when $a \leq t^*$. We will see later that this is not the case for large values of a. The next section is devoted to closing this gap by deriving matching upper bounds.

5.3 Optimal scaled minimax estimators

In this section, we consider upper bounds for the scaled minimax risk under model (5.1). For $a > 0$ define the following estimator:

$$\hat{\beta}^a_j := Y_j \mathbf{1}_{\{|Y_j| \geq t(a \vee t^*)\}}, \quad \forall j \in 1, \ldots, p, \tag{5.7}$$

where $t(.)$ and t^* are defined respectively in (5.5) and (5.6). The following result gives a matching upper bound for the scaled minimax risk. Set

$$\Phi_+(a) := \begin{cases} a^q \psi_+(p,s,a,\sigma) \vee \sigma_q^q s & \text{if } a \geq t^*, \\ s\sigma^q \left(2\log(\frac{p}{s}-1)\right)^{\frac{q}{2}} & \text{else,} \end{cases}$$

where ψ_+ is given by

$$\psi_+(p,s,a,\sigma) := (p-s)\,\mathbf{P}\left(\sigma\varepsilon > t\,(a)\right) + s\mathbf{P}\left(\sigma\varepsilon > (a - t(a)_+)\right),$$

and ε is a standard Gaussian random variable. Notice that $\Phi_+(a) \leq \Phi(a)$. This remark, combined with the next theorem, shows minimax optimality of the estimator (5.7).

Theorem 5.3.1. *Assume that Y follows model (5.1). For all $a > 0$, let $\hat{\beta}^a$ be the estimator (5.7). For all $q \geq 1$ and $s < p/2$ we have*

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta}^a - \beta\|_q^q \right) \leq C_q \Phi_+(a),$$

where C_q is a universal constant depending only in q.

Combining this result with Proposition 5.2.1, we deduce the next corollary.

Corollary 5.3.1. *Assume that Y follows model (5.1). For all $a > 0$, let $\hat{\beta}^a$ be the estimator (5.7). For all $q \geq 1$, $s < p/2$ and $s \geq 8q\log\log(p)$, there exists $C_q > 0$ such that*

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta}^a - \beta\|_q^q \right) \leq C_q \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right).$$

We give now a more accurate upper bound in the regime $\frac{s}{p} \to 0$. Assume that $s \leq p/4$. For $q \geq 1$, and $\epsilon \in [0,1]$ define

$$a_q(\epsilon) = \sigma\sqrt{2\log(p/s-1) + q\epsilon\log\log(p/s-1)} + \sigma\sqrt{q\epsilon\log\log(p/s-1)}.$$

Set

$$\Phi_o(a) := \begin{cases} s\sigma^q \left(2\log(\frac{p}{s}-1)\right)^{\frac{q}{2}} & \text{if } a \leq a_q(0), \\ \frac{s\sigma^q \left(2\log(\frac{p}{s}-1)\right)^{\frac{q}{2}(1-\epsilon)}}{1+\sigma\sqrt{\frac{\pi}{2}\epsilon q\log\log(\frac{p}{s}-1)}} \vee \sigma_q^q s & \text{if } a = a_q(\epsilon), \epsilon \in (0,1), \\ s\sigma_q^q & \text{if } a \geq a_q(1). \end{cases}$$

The next theorem gives sharp upper bounds in the regime $\frac{s}{p} \to 0$.

Theorem 5.3.2. *Assume that Y follows model (5.1). For all $a > 0$, let $\hat{\beta}^a$ be the estimator (5.7). In the regime where $\frac{s}{p} \to 0$, for all $q \geq 1$, we have*

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta}^a - \beta\|_q^q \right) \leq \Phi_o(a)(1 + o(1)).$$

As a consequence of previous results, we derive the next corollary that gives an almost sharp bound for SMSE when $\frac{s}{p} \to 0$.

Corollary 5.3.2. *Assume that Y follows model (5.1). For all $a > 0$, $q \geq 1$ and $s \geq 8q \log\log(p)$, in the regime $\frac{s}{p} \to 0$, we have*

$$\frac{1}{4} + o(1) \leq \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \frac{\mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right)}{\Phi_o(a)} \leq 1 + o(1),$$

and

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \frac{\mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right)}{s\sigma_q^q} = 1 + o(1) \quad \text{if } a \geq a_q(1).$$

Inspecting the proof of Corollary 5.3.2, we may notice that the discrepancy between the bounding constants is mainly caused by values of the scale $a = a_q(\epsilon)$ such that $\epsilon \in (0, 1)$. Corollary 5.3.2 shows that we can construct an almost sharp optimal minimax estimator provided a and s. The next section is devoted to the question of adaptivity.

5.4 Adaptative scaled minimax estimators

In Section 5.3 we have shown that the minimax rate is given by the quantity $\Phi_o(a)$ in a sharp way if $\frac{s}{p} \to 0$. Note that $\Phi_o(a)$ is almost piece-wise constant as a function of a. In fact the study of $\Phi_o(a)$ gives rise to three different regimes that we describe below.

1. *Hard recovery regime:*

   Let $a \leq \sigma\sqrt{2\log\left(\frac{p}{s} - 1\right)}$. We call this the hard recovery regime. In this regime, $\Phi_o(a)$ is constant and has a value of order $\sigma^q s \left(2\log\left(\frac{p}{s} - 1\right)\right)^{q/2}$. It turns out that the worst case of estimation happens for $a = \sigma\sqrt{2\log\left(\frac{p}{s} - 1\right)}$. This error is mainly due to the fact the we cannot achieve almost full recovery as defined in Butucea et al. (2018).

2. *Hard estimation regime:*

   This regime corresponds to values of a such that

   $$a \geq \sigma\sqrt{2\log\left(p/s - 1\right)}\sqrt{1 + 4\frac{q\log\log(p/s - 1)}{\log(p/s - 1)}}.$$

   In this regime $\Phi_o(a)$ is of order $\sigma_q^q s$. In this region, the error of estimation on a known support dominates the error of recovering the support.

3. *Transition regime:*

   This regime concerns the remaining values of a falling between the two previous regimes. In this regime $\Phi_o(a)$ is not constant any more. It represents a monotonous and continuous transition from one regime to another.

After analyzing the SMSE, we give a couple of remarks.

Remark 5.4.1.

- *If $s = o(p)$ there are basically two regimes around the threshold $\sigma\sqrt{2\log\left(\frac{p}{s}-1\right)}$. Notice also that the hard estimation error is very small compared to the hard recovery error. We may notice that the SMSE is very small compared to the minimax sparse estimation error in the hard estimation regime. This proves how pessimistic the general minimax lower bounds are and that we can do much better for the scaled minimax risk.*

- *The case $s \sim p$ is of small interest. There is no phase transition in this case, since the SMSE is of order $\sigma^q p$ for every a.*

- *In the Hard estimation regime, the minimax error rate is the same as if the support were exactly known. It is interesting to notice that we need a weaker condition to get this rate when $s \gg \log(p)$, while a stronger necessary condition is needed for exact recovery, cf. Butucea et al. (2018). Hence exact support recovery is not necessary to achieve exact estimation.*

Notice also that the transition regime happens in a very small neighborhood around the universal threshold $\sigma\sqrt{2\log(p/s)}$. Thus, it is very difficult to be adaptive to a in the transition regime. For $s \leq p/4$, define the following estimator:

$$\hat{\beta}_j^s := Y_j \mathbf{1}_{\{|Y_j| \geq t_s^*\}}, \quad \forall j \in 1, \dots, p, \tag{5.8}$$

where

$$t_s^* := \sigma\sqrt{2\log(p/s-1) + q\log\log(p/s-1)}.$$

We define a more convenient adaptive estimation error. Set

$$\Phi_{ad}(a) := \begin{cases} \sigma_q^q s & \text{if } a \geq a_q(1), \\ s\sigma^q \left(2\log(\frac{p}{s}-1)\right)^{\frac{q}{2}} & \text{else.} \end{cases}$$

The following result gives a matching upper bound for the adaptive scaled minimax risk.

Theorem 5.4.1. *Assume that Y follows model (5.1). Let $\hat{\beta}^s$ be the estimator (5.8). For all $q \geq 1$, $a > 0$ and $s < p/4$ we have*

$$\sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta \left(\|\hat{\beta}^s - \beta\|_q^q \right) \leq C_q \Psi_{ad}(a),$$

where C_q is a universal constant depending only in q.

Since the two main regimes are hard estimation and hard recovery, we restricted the notion of adaptivity to these regimes. By doing so, we constructed an almost optimal estimator adaptively to the parameter a. This estimator is minimax optimal over the set of s-sparse vectors and achieves exact estimation when necessary conditions are satisfied. Our estimator has a phase transition around the universal threshold. Based on a procedure similar to Butucea et al. (2018), we can also construct an optimal estimator adaptive to sparsity. We do not give further details here for the sake of brevity.

5.5 Conclusion

In this chapter, we define and study a new notion that we call scaled minimax sparse estimation. We assess how pessimistic are minimax lower bounds for the problem of sparse estimation. We also show that exact recovery is not necessary for exact estimation in general. As a result, we construct a new estimator optimal for the SMSE and present its adaptive version, improving on existing procedures for the problem of estimation.

5.6 Appendix: Proofs

Sharp Gaussian tail bounds

The following bounds for the tails of Gaussian distribution will be useful:

$$\frac{e^{-y^2/2}}{\sqrt{2\pi}y+4} \le \frac{1}{\sqrt{2\pi}}\int_y^\infty e^{-u^2/2}du \le \frac{e^{-y^2/2}}{\sqrt{2\pi}y\vee 2}.$$

for all $y \ge 0$. These bounds are an immediate consequence of formula 7.1.13 in Abramowitz and Stegun (1964) with $x = y/\sqrt{2}$.

Proof of Theorem 5.2.1

Throughout the proof, we write for brevity $A = \Theta_{s,a}$. Set $\beta^A = \beta.\mathbf{1}(\beta \in A)$ and denote by π_A the probability measure π conditioned by the event $\{\beta \in A\}$, that is, for any $C \subseteq \mathbb{R}^d$,

$$\pi_A(C) = \frac{\pi(C \cap \{\beta \in A\})}{\pi(\beta \in A)}\,.$$

The measure π_A is supported on A and we have

$$\begin{aligned}\inf_{\hat\beta}\sup_{\beta\in A}\mathbf{E}_\beta|\hat\beta-\beta|_q^q &\ge \inf_{\hat\beta}\mathbb{E}_{\pi_A}\mathbf{E}_\beta|\hat\beta-\beta|_q^q = \inf_{\hat\beta}\mathbb{E}_{\pi_A}\mathbf{E}_\beta|\hat\beta-\beta^A|_q^q\\ &\ge \sum_{j=1}^p \inf_{\hat T_j}\mathbb{E}_{\pi_A}\mathbf{E}_\beta|\hat T_j-\beta_j^A|^q\end{aligned}$$

where $\inf_{\hat T_j}$ is the infimum over all estimators $\hat T_j = \hat T_j(Y)$ with values in $\mathbb{R}$. According to Theorem 1.1 and Corollary 1.2 on page 228 in Lehmann and Casella (2006), there exists a Bayes estimator $B_j^A = B_j^A(Y)$ such that

$$\inf_{\hat T_j}\mathbb{E}_{\pi_A}\mathbf{E}_\beta|\hat T_j-\beta_j^A|^q = \mathbb{E}_{\pi_A}\mathbf{E}_\beta|B_j^A-\beta_j^A|^q.$$

In particular, for any estimator $\hat T_j(Y)$ we have

$$\mathbb{E}^A\big(|B_j^A(Y)-\beta_j^A|^q\big|Y\big) \le \mathbb{E}^A\big(|\hat T_j(Y)-\beta_j^A|^q\big|Y\big) \tag{5.9}$$

almost surely. Here, the superscript A indicates that the conditional expectation $\mathbb{E}^A(\cdot|Y)$ is taken when β is distributed according to π_A. Therefore,

$$\inf_{\hat\beta}\sup_{\beta\in A}\mathbf{E}_\beta|\hat\beta-\beta|_q^q \ge \mathbb{E}_{\pi_A}\mathbf{E}_\beta\sum_{j=1}^p|B_j^A-\beta_j^A|^q. \tag{5.10}$$

Using this, we obtain

$$\begin{aligned}\inf_{\hat{T}\in\mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}-\beta|_q^q &\le \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |B_j^A - \beta_j|^q \\ &= \mathbb{E}_\pi \mathbf{E}_\beta \Big(\sum_{j=1}^p |B_j^A-\beta_j|^q \mathbf{1}(\beta \in A)\Big) + \mathbb{E}_\pi \mathbf{E}_\beta \Big(\sum_{j=1}^p |B_j^A-\beta_j|^q \mathbf{1}(\beta \notin A)\Big) \\ &\le \mathbb{E}_{\pi_A} \mathbf{E}_\beta \sum_{j=1}^p |B_j^A-\beta_j^A|^q + \mathbb{E}_\pi \mathbf{E}_\beta \Big(\sum_{j=1}^p |B_j^A-\beta_j|^q \mathbf{1}(\beta \notin A)\Big) \\ &\le \mathbb{E}_{\pi_A} \mathbf{E}_\beta \sum_{j=1}^p |B_j^A-\beta_j^A|^q + \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p 2^{q-1}(|B_j^A|^q+|\beta_j|^q)\mathbf{1}(\beta \notin A).\end{aligned} \tag{5.11}$$

Our next step is to bound the term

$$\mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |B_j^A|^q \mathbf{1}(\beta \notin A).$$

For this purpose, we first note that inequality (5.9) with $\hat{T}_j(Y)=0$ implies that

$$|B_j^A(Y)|^q = \mathbb{E}^A(|B_j^A(Y)|^q|Y) \le 2^q \mathbb{E}^A(|\beta_j^A|^q|Y).$$

Thus

$$\mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |B_j^A|^q \mathbf{1}(\beta \notin A) \le 2^q \mathbb{E}_\pi \mathbb{E}^A(\|\beta^A\|_q^q|Y)\mathbf{1}(\beta \notin A).$$

Combining this inequality with (5.10) and (5.11) yields (5.4).

Proof of Lemma 5.2.1

We begin by noticing that

$$\inf_{\hat{T}\in\mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |\hat{T}_j(X,Y)-\beta_j|^q = \sum_{j=1}^p \inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y)-\beta_j|^q.$$

It is easy to check that

$$\forall a \in \mathbb{R}^p, \forall j=1,\dots,p \quad \inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y)-\beta_j|^q = \inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y-a)-\beta_j|^q. \tag{5.12}$$

Using conditioning, one may also notice that

$$\inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y)-\beta_j|^q \ge \mathbb{E}_{\pi_{-j}} \left(\inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X,Y)-\beta_j|^q \Big| \beta_{-j} \right), \tag{5.13}$$

where β_{-j} represents the vector β deprived of β_j and π_{-j} the corresponding prior. Hence, we get from (5.12) and (5.13) that

$$\inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y)-\beta_j|^q \ge \mathbb{E}_{\pi_{-j}} \left(\inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X,\tilde{Y}_j)-\beta_j|^q \Big| \beta_{-j} \right),$$

where $\tilde{Y}_j = Y - \sum_{i \neq j} X_i \beta_i = \beta_j X_j + \sigma \xi$. We remove the last conditional expectation and replace the dependence on X by X_j, since the observable $\tilde{Y}_j$ depends only on β_j and X_j.

Proof of Theorem 5.2.2

We apply Theorem 5.2.1 with $\Theta_{s,a} = \Omega_{s,a}$. Let S a support of size s, and consider the prior β such that $\beta_{S^c} = 0$ and $\beta_S = Z$, where $Z \in \mathbb{R}^s$ is a Gaussian random vector distributed following $\mathcal{N}(\mu, \nu^2 \mathbb{I}_s)$ where $\mu, \nu > 0$ are defined later. We have

$$\inf_{\hat{\beta}} \sup_{\beta \in \beta_{s,a}} \mathbf{E}_\beta |\hat{\beta} - \beta|^q \geq \inf_{\hat{T} \in \mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |\hat{T}_j(X, Y) - \beta_j|^q - C_q \, \mathbb{E}_\pi \left[\left(\mathbf{E}(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Omega_{s,a}) \right].$$

We first upper-bound the second term

$$\mathbb{E}_\pi \left[\left(\mathbf{E}^A(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta_{s,a}) \right] \leq 2 \sqrt{\mathbb{E}_\pi \|\beta\|_q^{2q}} \sqrt{\mathbf{P}(\beta \notin \Theta s, a)},$$

since $\|\beta^A\|_q^q \leq \|\beta\|_q^q$. It is easy to check that for some $C > 0$ we have

$$\mathbf{P}(\beta \notin \Theta_{s,a}) \leq s \mathbf{P}(|\beta_1| \leq a) \leq C s e^{-\frac{(\mu_1 - a)_+^2}{2\nu^2}}.$$

By choosing $\mu_1 = a + \nu^2$, we get for some $C_q > 0$

$$\mathbb{E}_\pi \left[\left(\mathbb{E}^A(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta_{s,a}) \right] \leq C_q \sqrt{s} p \sqrt{a^{2q} + \nu^{4q} + \nu^{2q}} e^{-\frac{\nu^2}{2}}.$$

Using lemma 5.2.1 combined with Anderson lemma for Gaussian priors we get

$$\inf_{\hat{T}_j \in \mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X, \tilde{Y}_j) - \beta_j|^q = \mathbf{E} \left(\left(\frac{\nu \sigma}{\nu \|X_j\| + \sigma} \right)^q |\xi_1|^q \right).$$

We conclude that $\forall \nu > 0$, we have

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq \sum_{j \in S} \left(\frac{\nu \sigma}{\nu \|X_j\| + \sigma} \right)^q \mathbf{E} \left(|\xi_1|^q \right) - C_q \sqrt{s} p \sqrt{a^{2q} + \nu^{4q} + \nu^{2q}} e^{-\frac{\nu^2}{2}}.$$

The result follows by taking the limit $\nu \to \infty$.

Proof of Theorem 5.2.3

We are going to mimic the previous proof using a different prior. We apply Theorem 5.2.1 with $\Theta_{s,a} = \Omega_{s,a}$. Consider the prior β such that $\beta = a\eta$, where $\eta \in \{0, 1\}^p$ be a Bernoulli random vector with i.i.d entries and $\mathbf{E}(\eta_i) = \frac{s'}{p}$, $s' \in (0, s)$. We have

$$\inf_{\hat{\beta}} \sup_{\beta \in \Theta s, a} \mathbf{E}_\beta |\hat{\beta} - \beta|^q \geq \inf_{\hat{T} \in \mathbb{R}^d} \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |\hat{T}_j(X, Y) - \beta_j|^q - C_q \, \mathbb{E}_\pi \left[\left(\mathbf{E}(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta s, a) \right].$$

First notice that in this case

$$\beta \in \Theta s, a \quad \text{if and only if} \quad |\eta|_0 \leq s.$$

Hence $\|\beta^A\|_q^q \leq a^q |\eta|_0 \leq s a^q$. We first upper-bound the second term

$$\mathbb{E}_\pi \left[\left(\mathbf{E}^A(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta_{s,a}) \right] \leq a^q \mathbb{E}_\pi \left[2|\eta|_0 \mathbf{1}(|\eta|_0 \geq s+1) \right],$$

since $|\eta|_0 > s$. Using same arguments as in Butucea et al. (2018), we conclude that

$$\mathbb{E}_\pi \left[\left(\mathbf{E}^A(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta_{s,a}) \right] \leq 2a^q s' e^{-\frac{(s-s')^2}{2s}}.$$

Going back to the first term, we get the following lower bound using Lemma 5.2.1

$$\inf_{\hat{T}_j \in \mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X, \tilde{Y}_j) - \beta_j|^q = a^q \inf_{\hat{T}_j \in \mathbb{R}} \left(\frac{s'}{p} \mathbf{E}_a |\hat{T}_j(X, \tilde{Y}_j) - 1|^q + (1 - \frac{s'}{p}) \mathbf{E}_0 |\hat{T}_j(X, \tilde{Y}_j)|^q \right)$$

Minimizing the posterior risk, the Bayes rule gives

$$\forall q > 1, \quad T_j^*(X, \tilde{Y}_j) = \frac{1}{1 + e^{\frac{a}{q-1}(t_j(a)\|X_j\|^2 - \langle \tilde{Y}_j, X_j \rangle)}},$$

and for $q = 1$ we get

$$T_j^*(X, \tilde{Y}_j) = \mathbf{1}(\langle \tilde{Y}_j, X_j \rangle \geq t_j(a)\|X_j\|^2).$$

Hence we deduce that

$$\inf_{\hat{T}_j \in \mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X, \tilde{Y}_j) - \beta_j|^q \geq \frac{a^q}{2^q} \Psi,$$

where

$$\Psi = \left(\frac{s'}{p} \mathbf{P}_a(\langle \tilde{Y}_j, X_j \rangle \leq t_j(a)\|X_j\|^2) + (1 - \frac{s'}{p}) \mathbf{P}_0(\langle \tilde{Y}_j, X_j \rangle \geq t_j(a)\|X_j\|^2) \right).$$

Notice that for $q = 1$ the term 2^q is not needed. Replacing $\tilde{Y}_j$ by its expression, we recover the lower bound

$$\Psi(p, s', a, \sigma, X) = \sum_{j=1}^{p} \left(\frac{s'}{p} \mathbf{P}(\varepsilon \geq (a - t_j(a))\|X_j\|) + (1 - \frac{s'}{p}) \mathbf{P}(\varepsilon \geq t_j(a)\|X_j\|) \right).$$

Following the proof of Ndaoud and Tsybakov (2018), we may use the fact that $s \to \frac{\Psi(s)}{s}$ is decreasing to conclude the proof.

Proof of Proposition 5.2.1

Combining Theorem 5.2.2 and Theorem 5.2.3 with $s' = s/2$, we get

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq s\sigma_q^q \vee \left(\frac{a^q}{2^{q+1}} \psi(p, s, a, \sigma) - s a^q e^{-s/8} \right).$$

We remind the reader the notation $t^* := \sigma\sqrt{2\log(p/s - 1)}$. In order to prove the result, we handle several cases.

- case $a \geq 10t^*$:
  It is easy to check that $a - t(a) \geq a/4$ and that $t(a) \geq a/4 + t^*$. Hence

  $$a^q \psi(p, s, a, \sigma) \leq C s a^q e^{-a^2/32\sigma^2} \leq C_q s.$$

  This shows that the term $\sigma_q^q s$ is dominating. As a result

  $$s\sigma_q^q \vee \left(\frac{a^q}{2^{q+1}} \psi(p, s, a, \sigma) - s a^q e^{-s/8} \right) \asymp s,$$

  and

  $$s\sigma_q^q \vee a^q \psi(p, s, a, \sigma) \asymp s.$$

  This suffises to prove the lower bound.

- case $t^* \leq a \leq 10t^*$:
  Since $s \geq 8q \log\log p$, then

  $$a^q e^{-s/8} \leq C_q a^{-q/2} \leq C_{q'}.$$

  This leads to

  $$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq s\sigma_q^q \vee \left(\frac{a^q}{2^{q+1}} \psi(p, s, a, \sigma) - s C_{q'} \right).$$

  We conclude by noticing that $a \vee b \asymp a \vee (b - a)$ for $a, b \geq 0$.

- case $a \leq t^*$:
  We observe that $t(t^*) = t^*$. In this case

  $$\begin{aligned} \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq & \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,t^*}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \\ \geq & \frac{1}{2^{q+1}} s t^{*q} \mathbf{P}\left(\sigma\varepsilon \geq 0\right) - C_{q'} s \geq C_{q''} s t^{*q}. \end{aligned}$$

  Hence

  $$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq C_{q''} \sigma^q s \log\left(\frac{p}{s} - 1 \right)^{q/2}.$$

Proof of Theorem 5.3.1

Let β be a vector in $\Omega^p_{s,a}$, we have

$$\|\beta^a - \beta\|_q^q = \sum_{i \in S} \mathbf{E} \left| \hat{\beta}_i^a - \beta_i \right|^q + \sum_{i \in S^c} \mathbf{E} \left| \hat{\beta}_i^a - \beta_i \right|^q.$$

On S^c, we have

$$\hat{\beta}_i^a - \beta_i = \xi_i \mathbf{1}_{\{|\xi_i| > t(a \vee t^*)\}}.$$

Hence we get that

$$\mathbf{E} \left| \hat{\beta}_i^a - \beta_i \right|^q = \mathbf{E} \left(|\xi_i|^q \mathbf{1}_{\{|\xi_i| > t(a \vee t^*)\}} \right).$$

Using integration by parts and induction we get

$$\forall q \geq 0, \quad \mathbf{E}\left(|\xi_i|^q \mathbf{1}_{\{|\xi_i|>t(a\vee t^*)\}}\right) \leq C_q(t(a\vee t^*)^q + \sigma^q)\mathbf{P}\left(|\xi_i| \geq t(a\vee t^*)\right),$$

where C_q is a universal constant depending only in q. Applying this we get

$$\begin{aligned}\mathbf{E}\left|\hat{\beta}_i^a - \beta_i\right|^q &= \mathbf{E}\left(|\xi_i|^q \mathbf{1}_{\{|\xi_i|>t(a\vee t^*)\}}\right)\\ &\leq C_q(t(a\vee t^*)^q + \sigma^q)\mathbf{P}\left(|\xi_i| \geq t(a\vee t^*)\right).\end{aligned}$$

Hence

$$\mathbf{E}\sum_{i\in S^c}\left|\hat{\beta}_i - \beta_i\right|^q \leq 2C_q|S^c|t(a\vee t^*)^q\mathbf{P}\left(\sigma\varepsilon \geq t\left(a\vee t^*\right)\right).$$

The last inequality holds since $t(a\vee t^*) \geq c\sigma$ for $s \leq p/4$.

On S, we have

$$\hat{\beta}_i^a - \beta_i = Y_i\mathbf{1}_{\{|Y_i|>t(a)\}} - \beta_i = -\xi_i - Y_i\mathbf{1}_{\{|Y_i|\leq t(a)\}}.$$

Hence and since $|x+y|^q \leq 2^{q-1}(|x|^q + |y|^q)$ we get

$$\begin{aligned}\mathbf{E}\left|\hat{\beta}_i^a - \beta_i\right|^q &\leq 2^{q-1}\sigma_q^q + 2^{q-1}\mathbf{E}\left(|Y_i|^q\mathbf{1}_{\{|Y_i|\leq t(a\vee t^*)\}}\right)\\ &\leq 2^{q-1}\sigma_q^q + 2^{q-1}t(a\vee t^*)^q\mathbf{P}\left(|Y_i| \leq t(a\vee t^*)\right)\\ &\leq 2^{q-1}\sigma_q^q + 2^{q-1}t(a\vee t^*)^q\mathbf{P}\left(|\xi_i| \geq (a - t(a\vee t^*))_+\right).\end{aligned}$$

We get that on S we have

$$\mathbf{E}\sum_{i\in S}|\hat{\beta}_i^a - \beta_i|^q \leq C_q\left(s\sigma_q^q + t(a)^q|S|\mathbf{P}\left(\sigma\varepsilon > (a - t(a\vee t^*)_+)\right)\right). \tag{5.14}$$

Since $(a - t(a\vee t^*)_+) \leq (a\vee t^* - t(a\vee t^*)_+)$, we get

$$\mathbf{E}\sum_{i\in S}|\hat{\beta}_i^a - \beta_i|^q \leq C_q s\sigma_q^q + C_q t(a\vee t^*)^q|S|\mathbf{P}\left(\sigma\varepsilon > (a\vee t^* - t(a\vee t^*)_+)\right).$$

We conclude that

$$\mathbf{E}\left(\|\hat{\beta}^a - \beta\|_q^q\right) \leq C_q\sigma_q^q s + C_q t(a\vee t^*)^q\psi_+(p,s,t^*\vee a,\sigma).$$

Hence for $a \geq t^*$, the result is immediate, since $t(a\vee t^*) \leq t(a) \leq a$. For $a < t^*$ we have

$$\mathbf{E}\left(\|\hat{\beta}^a - \beta\|_q^q\right) \leq C_q\sigma_q^q s + C_q\sigma^q\log(\frac{p}{s} - 1)^{q/2}\psi_+(p,s,t^*,\sigma).$$

It is easy to verify

$$\psi_+(p,s,t^*,\sigma) \leq s + (p-s)\frac{s}{p-s} \leq 2s,$$

and hence

$$\mathbf{E}\left(\|\hat{\beta}^a - \beta\|_q^q\right) \leq C_q\left(\sigma^q s\log(p/s-1)^{q/2} + \sigma_q^q s\right) \leq C_{q'}\sigma^q s\log(p/s-1)^{q/2},$$

since $s \leq \frac{p}{4}$ and $\log(p/s-1) \geq 1$.

Proof of Theorem 5.3.2

Let us first notice that for $\epsilon \in [0,1]$ we have

$$t(a_q(\epsilon)) = \sigma\sqrt{2\log(p/s-1) + q\epsilon\log\log(p/s-1)},$$

and

$$a_q(\epsilon) - t(a_q(\epsilon)) = \sqrt{q\epsilon\sigma^2\log\log(p/s-1)}.$$

Following the previous proof we have

$$\mathbb{E}\sum_{i\in S^c}\left|\hat{\beta}_i - \beta_i\right|^q \leq 2C_q p t(a)^q \mathbf{P}\left(\sigma\varepsilon > t\left(a\right)\right).$$

Since $\epsilon \in [0,1]$ we have that $t(a_q(\epsilon)) \leq \sigma\sqrt{2\log(p/s-1)}(1+o(1))$. Moreover

$$\mathbf{P}\left(\sigma\varepsilon > t\left(a\right)\right) \leq C\sigma\frac{e^{-t(a)^2/2\sigma^2}}{t(a)} \leq \frac{C\sigma}{t(a)}\frac{s}{p-s}\frac{1}{\log(p/s-1)^{q\epsilon/2}}.$$

Hence

$$\mathbf{E}\sum_{i\in S^c}\left|\hat{\beta}_i - \beta_i\right|^q \leq C_q\frac{s\log(p/s-1)^{\frac{q}{2}(1-\epsilon)}}{\sqrt{\log(p/s-1)}}.$$

We can now notice that on S^c we have

$$\mathbf{E}\sum_{i\in S^c}\left|\hat{\beta}_i - \beta_i\right|^q = \underset{\frac{s}{p}\to 0}{o}\left(\Phi_o\right).$$

In order to prove the Theorem we focus on the error in the support. Remember that on S we have

$$\hat{\beta}_i^a - \beta_i = Y_i\mathbf{1}_{\{|Y_i|>t(a)\}} - \beta_i = -\xi_i - Y_i\mathbf{1}_{\{|Y_i|\leq t(a)\}}.$$

- case $a \leq a(0)$:
  In this case $a(0) = t(a(0)) = \sigma\sqrt{2\log(p/s-1)}$. We use the following inequality

  $$\forall a,b\in\mathbb{R}, q\geq 1, \quad |a+b|^q \leq |a|^q + q|a+b|^{q-1}|b|.$$

  Hence

  $$\begin{aligned}|\xi_i - Y_i\mathbf{1}_{\{|Y_i|\leq t(a)\}}|^q &\leq |Y_i\mathbf{1}_{\{|Y_i|\leq t(a)\}}|^q + q|\xi_i||\xi_i - Y_i\mathbf{1}_{\{|Y_i|\leq t(a)\}}|^{q-1}\\ &\leq |Y_i 1_{\{|Y_i|\leq t(a)\}}|^q + q|\xi_i|2^q\left(|\xi_i|^{q-1} + |Y_i\mathbf{1}_{\{|Y_i|\leq t(a)\}}|^{q-1}\right)\\ &\leq t(a)^q + q2^q\left(|\xi i|^q + |\xi_i|t(a)^{q-1}\right).\end{aligned}$$

  As a consequence

  $$\mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \leq t(a)^q + q2^q(\sigma_q^q + \sigma_1 t(a)^{q-1}) \leq t(a)^q(1+o(1)).$$

  The last inequality holds since $t(a)\to\infty$ as $s/p\to 0$. We conclude that

  $$\sum_{i\in S}\mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \leq \Phi_o(1+o(1)).$$

- case $a = a(\epsilon)$ for $\epsilon \in (0,1)$:
  In this case and following same steps in previous case

$$|\xi_i - Y_i 1_{\{|Y_i| \le t(a)\}}|^q \le t(a)^q \mathbf{1}_{\{|Y_i| \le t(a)\}} + q2^q \left(|\xi_i|^q + |\xi_i| t(a)^{q-1} \mathbf{1}_{\{|Y_i| \le t(a)\}}\right).$$

  Remember that

$$\forall q \ge 0, \quad \mathbf{E}\left(|\xi_i|^q 1_{\{|\xi_i| > t(a)\}}\right) \le C_q (t(a)^q + \sigma_2^q) \mathbf{P}\left(|\xi_i| \ge t(a)\right).$$

  Hence

$$\begin{aligned}\mathbf{E}(|\xi_i| \mathbf{1}_{\{|Y_i| \le t(a)\}}) &\le \mathbf{E}(|\xi_i| 1_{\{|\xi_i| \ge a - t(a)\}}) \\ &\le ((a-t) + \sigma)\mathbf{P}(|\xi_i| \ge a - t) \le \log(t)\mathbf{P}(|Y_i| \le t).\end{aligned}$$

  We get that

$$\mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \le t^{*q}(1 + o(1))\mathbf{P}(|Y_i| \le t(a)) + C_q \sigma_q^q.$$

  One may notice that

$$\mathbf{P}(|Y_i| \le t(a)) \le 2\mathbf{P}(\sigma\epsilon \ge (a - t(a))_+).$$

  Using the Gaussian tail inequality and the fact that $\epsilon > 0$, we get

$$\mathbf{P}(|Y_i| \le t(a)) \le \frac{t^{*-q\epsilon}}{1 + \sqrt{\frac{\pi}{2}} q\epsilon \log\log(p/s - 1)}(1 + o(1)).$$

  Since $t^{*q(1-\epsilon)} / \log(t) \to \infty$ we conclude that

$$\sum_{i \in S} \mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \le \frac{s t^{*q(1-\epsilon)}}{1 + \sqrt{\frac{\pi}{2}} q\epsilon \log\log(p/s - 1)}(1 + o(1)).$$

- case $a \ge a(1)$:
  In this case it suffices to prove the result for $a = a(1)$ since the minimax risk in increasing with respect to a.

$$\begin{aligned}|\xi_i - Y_i \mathbf{1}_{\{|Y_i| \le t(a)\}}|^q &\le |\xi_i|^q + q|Y_i \mathbf{1}_{\{|Y_i| \le t(a)\}}||\xi_i - Y_i \mathbf{1}_{\{|Y_i| \le t(a)\}}|^{q-1} \\ &\le |\xi_i|^q + C_q \left(t^* |\xi_i|^{q-1} \mathbf{1}_{\{|Y_i| \le t(a)\}} + t^{*q} \mathbf{1}_{\{|Y_i| \le t(a)\}}\right).\end{aligned}$$

  In the previous case we proved that

$$\mathbf{E}(|\xi_i|^{q-1} \mathbf{1}_{\{|Y_i| \le t(a)\}}) \le \log(t^*)^{q-1} \mathbb{P}(|Y_i| \le t(a)),$$

  and that

$$\mathbf{P}(|Y_i| \le t(a)) \le C \frac{t^{*-q}}{\log\log(p/s) + 1}.$$

  Hence

$$\mathbf{E}\left(t^* |\xi_i|^{q-1} \mathbf{1}_{\{|Y_i| \le t(a)\}} + t^{*q} \mathbf{1}_{\{|Y_i| \le t(a)\}}\right) \le \frac{C}{\log\log(p/s)} = o(\sigma_q^q).$$

  It follows that

$$\sum_{i \in S} \mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \le s\sigma_q^q (1 + o(1)).$$

  This concludes the proof of this theorem.

Proof of Corollary 5.3.2

Based on the fact that

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq s\sigma_q^q,$$

we observe that the second result is a direct consequence of Theorem 5.3.2. In order to conclude, we need to show that for $a < a_q(1)$ we have

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \frac{\mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right)}{\Phi_o(a)} \geq \frac{1}{4} + o(1).$$

In what follows, we assume that $a < a_q(1)$. Going back to the initial lower bound with $s' = s/2$, and using the fact that $s \geq 8q \log\log p$, we have

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq \frac{1}{2} s a^q \mathbf{E}\left(T_1^q\right) - C_{q''} s a(1)^q a(0)^{-2q},$$

where

$$\forall q > 1, \quad T_1 = \frac{1}{1 + e^{-\frac{a}{q-1}(t(a) - a + \xi_1)}},$$

and for $q = 1$

$$T_1 = \mathbf{1}(\xi_1 \geq a - t(a)).$$

Since

$$s a(1)^q a(0)^{-2q} = o(s),$$

we get immediately that

$$s a(1)^q a(0)^{-2q} = o(\Phi_o(a)).$$

It is sufficient to prove that

$$s a^q \mathbf{E}\left(T_1^q\right) \geq \frac{1}{2} \Phi_o(a)(1 + o(1)).$$

For $q = 1$, we have $\mathbf{E}(T_1) = \mathbf{P}(\xi_1 \geq a - t(a)) \geq \mathbf{P}(|\xi_1| \geq (a - t(a))_+)/2$. Using the fact that the Gaussian tail bounds presented in the beginning of this Appendix are sharp combined with the proof of the previous upper bound we can verify that for $q = 1$

$$s a \mathbf{E}\left(T_1\right) \geq \frac{1}{2} \Phi_o(a)(1 + o(1)).$$

For $q > 1$ it is enough to prove that

$$\mathbf{E}(T_1^q) \geq \mathbf{P}(\xi_1 \geq a - t(a))(1 + o(1)).$$

For $a = a_q(0)$ we have that $t(a) = a$, hence

$$\mathbf{E}(T_1^q) = \mathbf{E}\left(\left(\frac{1}{1 + e^{\frac{a}{q-1}\xi_1}} \right)^q \right) \to \mathbf{P}(\xi_1 \geq 0).$$

The limit is a consequence of the dominated convergence theorem and proves the result. The last case is when $a = a_q(\epsilon)$ with $\epsilon \in (0, 1)$. Let us just recall that $a \asymp \sqrt{\log(p/s)}$ and $a - t(a) \asymp \sqrt{\log\log(p/s)}$. Let $\alpha_s > 0$ be a sequence satisfying

$$\alpha_s . a \to \infty \text{ and } \alpha_s . (a - t) \to 0.$$

We have

$$\mathbf{E}(T_1^q) \geq \mathbf{E}\left(\left(\frac{1}{1+e^{\frac{-a\alpha_s}{q-1}}}\right)^q \mathbf{1}(\xi_1 \geq a - t(a) + \alpha_s)\right) \geq \left(\frac{1}{1+e^{\frac{-a\alpha_s}{q-1}}}\right)^q \mathbf{P}(\xi_1 \geq a - t(a) + \alpha_s).$$

Using the monotony of cumulative distribution functions, we get that

$$\mathbf{E}(T_1^q) \geq (1+o(1))\left(\mathbf{P}(\xi_1 \geq a - t(a)) - Ce^{-(a-t(a)+\alpha_s)^2/2\sigma^2}\alpha_s\right).$$

Using the limiting behaviour of α_s we get

$$\mathbf{E}(T_1^q) \geq \mathbf{P}(\xi_1 \geq a - t(a))(1+o(1)).$$

This concludes the proof.

Proof of Theorem 5.4.1

First notice that $t_s^* = t(a(1))$ and that $\hat{\beta}^s = \hat{\beta}^{a(1)}$. Hence and using Theorem 5.3.2 we get for all $a \geq a(1)$

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta\left(\|\hat{\beta}^s - \beta\|_q^q\right) \leq \sup_{\beta \in \Omega^p_{s,a(1)}} \mathbf{E}_\beta\left(\|\hat{\beta}^s - \beta\|_q^q\right) \leq C_q s\sigma_q^q.$$

On S, we have

$$\hat{\beta}_i^s - \beta_i = Y_i \mathbf{1}_{\{|Y_i| > t_s^*\}} - \beta_i = -\xi_i - Y_i \mathbf{1}_{\{|Y_i| \leq t_s^*\}}.$$

Hence

$$\mathbf{E}_\beta\left(\sum_{i \in S} |\hat{\beta}_i^s - \beta_i|^q\right) \leq C_q(s\sigma_q^q + st_s^{*q}).$$

On S^c, and since $t_s > t^*$, it is easy to check using previous proofs that

$$\mathbf{E}_\beta\left(\sum_{i \in S^c} |\hat{\beta}_i^s - \beta_i|^q\right) \leq C'_q s t_s^{*q}.$$

This concludes the proof.

Part II

From Variable Selection to Community Detection

Chapter 6

Sharp optimal recovery in the two component Gaussian Mixture Model

In this chapter, we study the problem of clustering in the Two component Gaussian mixture model when the centers are separated by some $\Delta > 0$. We present a non-asymptotic lower bound for the corresponding minimax Hamming risk improving on existing results. We also propose an optimal, efficient and adaptive procedure that is minimax rate optimal. The rate optimality is moreover sharp in the asymptotics when the sample size goes to infinity. Our procedure is based on a variant of Lloyd's iterations initialized by a spectral method. As a consequence of non-asymptotic results, we find a sharp phase transition for the problem of exact recovery in the Gaussian mixture model. We prove that the phase transition occurs around the critical threshold $\bar{\Delta}$ given by

$$\bar{\Delta}^2 = \sigma^2 \left(1 + \sqrt{1 + \frac{2p}{n \log n}}\right) \log n.$$

Based on Ndaoud (2018b): Ndaoud, M. (2018b). Sharp optimal recovery in the two component Gaussian mixture model. *arXiv preprint arXiv:1812.08078*.

6.1 Introduction

The problems of supervised or unsupervised clustering have gained huge interest in the machine learning literature. In particular, many clustering algorithms are known to achieve good empirical results. A very useful model to study and compare these algorithms is the Gaussian mixture model. In this model, we assume that the data are attributed to different centers and that we only have access to observations corrupted by Gaussian noise. For this specific model, one can consider the problem of estimation of the centers, see, e.g., Klusowski and Brinda (2016),Mixon et al. (2016) or the problem of detecting the communities, see, e.g., Lu and Zhou (2016),Giraud and Verzelen (2018),Royer (2017). This paper focuses on community detection.

The Gaussian Mixture Model

We observe n independent random vectors $Y_1, \dots, Y_n \in \mathbf{R}^p$. We assume that there exist two unknown vectors $\boldsymbol{\theta} \in \mathbf{R}^p$ and $\eta \in \{-1, 1\}^n$, such that, for all $i = 1, \dots, n$,

$$Y_i = \boldsymbol{\theta}\eta_i + \sigma\xi_i, \tag{6.1}$$

where $\sigma > 0$, $\xi_1, \dots, \xi_n$ are standard Gaussian random vectors and η_i is the ith component of η. We denote by Y (respectively, W) the matrix with columns $Y_1, \dots, Y_n$ (respectively, $\sigma\xi_1, \dots, \sigma\xi_n$). Model (6.1) can be written in matrix form

$$Y = \boldsymbol{\theta}\eta^\top + W.$$

We denote by $\mathbf{P}_{(\boldsymbol{\theta},\eta)}$ the distribution of Y in model (6.1) and by $\mathbf{E}_{(\boldsymbol{\theta},\eta)}$ the corresponding expectation. We assume that $(\boldsymbol{\theta}, \eta)$ belongs to the set

$$\Omega_\Delta = \{\boldsymbol{\theta} \in \mathbf{R}^p : \|\boldsymbol{\theta}\| \geq \Delta\} \times \{-1, 1\}^n,$$

where $\Delta > 0$ is a given constant. The value Δ characterizes the separation between the clusters and equivalently the strength of the signal.

In this chapter, we study the problem of recovering the communities, that is, of estimating the vector η. As estimators of η, we consider any measurable functions $\hat{\eta} = \hat{\eta}(Y_1, \dots, Y_n)$ of $(Y_1, \dots, Y_n)$ taking values in $\{-1, 1\}^n$. We characterize the loss of a given $\hat{\eta}$ by the Hamming distance between $\hat{\eta}$ and η, that is, by the number of positions at which $\hat{\eta}$ and η differ:

$$|\hat{\eta} - \eta| := \sum_{j=1}^n |\hat{\eta}_j - \eta_j| = 2\sum_{j=1}^n \mathbf{1}(\hat{\eta}_j \neq \eta_j).$$

Here, $\hat{\eta}_j$ and η_j are the jth components of $\hat{\eta}$ and η, respectively. Since for community detection it is enough to determine η up to a sign change, one can also consider the loss defined by

$$r(\hat{\eta}, \eta) := \min_{\nu \in \{-1,1\}} |\hat{\eta} - \nu\eta|.$$

In what follows, we use this loss. The expected loss of $\hat{\eta}$ is defined as $\mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\hat{\eta}, \eta)$.

In the rest of the paper, we will always denote by η the vector to estimate, while $\hat{\eta}$ will denote the corresponding estimator. We consider the following minimax risk

$$\Psi_\Delta := \inf_{\tilde{\eta}} \sup_{(\boldsymbol{\theta},\eta)\in\Omega_\Delta} \frac{1}{n} \mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\tilde{\eta}, \eta), \tag{6.2}$$

where $\inf_{\tilde{\eta}}$ denotes the infimum over all estimators $\tilde{\eta}$ in $\{-1, 1\}^n$. A simple lower bound for the risk Ψ_Δ is given by (cf. Proposition 6.3 below):

$$\Psi_\Delta \geq \frac{c}{1 + \Delta/\sigma} e^{-\frac{\Delta^2}{2\sigma^2}} \tag{6.3}$$

for some $c > 0$. Inspecting the proof one may also notice that this bound is attained at the oracle η^* given by

$$\eta_i^* = \operatorname{sign}\left(Y_i^\top \boldsymbol{\theta}\right).$$

This oracle assumes a prior knowledge of $\boldsymbol{\theta}$. It turns out that for $p \geq n$, there exists a regime where the lower bound (6.3) is not optimal, as pointed by Giraud and Verzelen (2018). The intuitive explanation is that for p larger than n, the vector $\boldsymbol{\theta}$ is hard to estimate. To the best of our knowledge, there are no lower bounds for Ψ_Δ that capture the issue of estimating $\boldsymbol{\theta}$. This is one of the main questions addressed in the present paper.

Notation. In the rest of this paper we use the following notation. For given sequences a_n and b_n, we write that $a_n = \mathcal{O}(b_n)$ (respectively, $a_n = \Omega(b_n)$) when $a_n \leq cb_n$ (respectively, $a_n \geq cb_n$) for some absolute constant $c > 0$. We write $a_n \asymp b_n$ when $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. For $x, y \in \mathbf{R}^p$, we denote by $x^\top y$ the Euclidean scalar product, by $\|x\|$ the corresponding norm of x and by $\mathrm{sign}(x)$ the vector of signs of the components of x. For $x, y \in \mathbf{R}$, we denote by $x \vee y$ (respectively, $x \wedge y$) the maximum (respectively, minimum) value between x and y. To any matrix $M \in \mathbf{R}^{n\times p}$, we denote by $\|M\|_{op}$ its operator norm with respect to the L^2-norm , by $M^\top$ its transpose and by $\mathrm{Tr}(M)$ its trace in case $p = n$. Further, $\mathbf{I}_n$ denotes the identity matrix of dimension n and $\mathbf{1}(.)$ denotes the indicator function. We denote by $\Phi^{\mathsf{c}}(.)$ the complementary cumulative distribution function of the standard Gaussian random variable z i.e., $\forall t \in \mathbf{R}, \Phi^{\mathsf{c}}(t) = \mathbf{P}(z > t)$. We denote by c and C positive constants that may vary from line to line.

We assume that p, σ and Δ depend on n and the asymptotic results correspond to the limit as $n \to \infty$. All proofs are deferred to Appendix 6.8.

Related literature

The present work can be related to two parallel lines of work.

1. Community detection in the sub-Gaussian mixture model:

   Lu and Zhou (2016) were probably the first to present statistical guarantees for community detection in the sub-Gaussian mixture model using the well-known Lloyd's algorithm, cf. Lloyd (1982). The results of Lu and Zhou (2016) require a better initialization than a random guess in addition to the condition

   $$\Delta^2 = \Omega\left(\sigma^2\left(1 \vee \frac{p}{n}\right)\right), \tag{6.4}$$

   in order to achieve *almost full recovery recovery* and

   $$\Delta^2 = \Omega\left(\sigma^2 \log n\left(1 \vee \frac{p}{n}\right)\right), \tag{6.5}$$

   in order to achieve *exact recovery*. The notions of almost full and exact recovery are defined in Section 6.5 and Section 6.7. More recently, Royer (2017) and Giraud and Verzelen (2018) have shown that conditions (6.4) and (6.5) are not optimal in high dimension i.e. for $n = o(p)$. In particular, Giraud and Verzelen (2018) study an SDP relaxation of the K-means criterion that achieves *almost full recovery* under a milder condition

   $$\Delta^2 = \Omega\left(\sigma^2\left(1 \vee \sqrt{\frac{p}{n}}\right)\right), \tag{6.6}$$

and *exact recovery* under the condition

$$\Delta^2 = \Omega\left(\sigma^2\left(\log n \vee \sqrt{\frac{p\log n}{n}}\right)\right). \tag{6.7}$$

To the best of our knowledge, conditions (6.6) and (6.7) are the mildest in the literature, but no matching necessary conditions are known so far. Giraud and Verzelen (2018) provide insightful heuristics about optimality of these conditions. In the supervised setting, where all labels are known similar conditions seem necessary to achieve either almost full or exact recovery. It is still not clear whether optimal conditions in supervised mixture learning are also optimal in the unsupervised setting.

Another difference between the previous papers is in computational aspects. While, in Giraud and Verzelen (2018), an SDP relaxation is proposed, a faster algorithm based on Lloyd's iterations is developed in Lu and Zhou (2016). It remains not clear whether we can achieve almost full (respectively, exact) recovery under condition (6.6) (respectively, (6.7)) through faster methods than SDP relaxations, for instance, through Lloyd's iterations.

Lu and Zhou (2016) suggest to initialize Lloyd's algorithm using a spectral method. It would be interesting to investigate whether Lloyd's algorithm initialized by a spectral method, in the same spirit as in Vempala and Wang (2004), can achieve optimal performance in the more general setting where p is allowed to be larger than n.

In this chapter, we shed some light on these issues. Specifically, we address the following questions.

- Are conditions (6.6) and (6.7) necessary for both almost full and exact recovery?
- Are optimal requirements similar in both supervised and unsupervised settings?
- Can we achieve results similar to Giraud and Verzelen (2018) using a faster algorithm?
- In case the answer to previous questions is positive, can we achieve the same results adaptively to all parameters?

2. Community detection in the Stochastic Block Model (SBM):

   The Stochastic Block Model, cf. Holland et al. (1983), is probably the most popular framework for node clustering. This model with two communities can be seen as a particular case of model (6.1) when both the signal and the noise are symmetric matrices. A non symmetric variant of SBM is the Bipartite SBM, cf. Feldman et al. (2015). Unlike the case of sub-Gaussian mixtures where most results in the literature are non-asymptotic, results on almost full or exact recovery for the SBM and its variants are mostly asymptotic and focus on sharp phase transitions. Abbe (2017) poses an open question on whether it is possible to characterize sharp phase transitions in other related problems, for instance, in the Gaussian mixture model.

The first polynomial method achieving exact recovery in the SBM with two communities is due to Abbe et al. (2014). The algorithm splits the initial sample into two independent samples. A black-box algorithm is used on the first sample for almost full recovery, then a local improvement is applied on the second sample. As stated in Abbe et al. (2014), it is not clear whether algorithms achieving almost full recovery can be used to achieve exact recovery. It remains interesting to understand whether similar results can be achieved through direct algorithms ideally without the splitting step.

For the Bipartite SBM, sufficient computational conditions for exact recovery are presented in Feldman et al. (2015), Florescu and Perkins (2016). While the sharp phase transition for the problem of detection is fully answered in Florescu and Perkins (2016), it is still not clear whether the condition they require, for exact recovery, is optimal. More interestingly, the sufficient condition for exact recovery is different for p of the same order as n and for p larger than n^2 for instance. This shows a kind of phase transition with respect to p, where for some critical dimension p^* the hardness of the problem changes.

We resume potential connections between our work and these recent developments in the following questions.

- Is it possible to characterize a sharp phase transition for exact recovery in model (6.1)?
- Are algorithms achieving almost full recovery useful in order to achieve exact recovery in the Gaussian mixture model?
- Is there a critical dimension p^* that separates different regimes of hardness in the problem of exact recovery?

Main contribution

In this work, we provide a sharp analysis of almost full and exact recovery in the two component Gaussian mixture model. Moreover, we give non-asymptotic lower bounds for the risk Ψ_Δ and matching upper bounds through a variant of Lloyd's iterations initialized by a spectral method. To do so, we define a key quantity $\mathbf{r}_n$ that turns out to be the right signal-to-noise ratio (SNR) of the problem:

$$\mathbf{r}_n = \frac{\Delta^2/\sigma^2}{\sqrt{\Delta^2/\sigma^2 + p/n}}. \tag{6.8}$$

This SNR is strictly smaller than the "naive" one Δ/σ, cf. (6.3). In particular, it states that the hardness of the problem depends on the dimension p. Among other results, we prove that for some $c_1, c_2, C_1, C_2 > 0$, we have

$$C_1 e^{-c_1 \mathbf{r}_n^2} \leq \Psi_\Delta \leq C_2 e^{-c_2 \mathbf{r}_n^2}.$$

Moreover, we give a sharp characterization of the constants in this relation.

Inspecting the proofs of the lower bounds in Section 6.2, one may learn that, in a setting where no prior information on $\boldsymbol{\theta}$ is given, the supervised learning estimator is optimal. Interestingly, supervised and unsupervised risks are almost equal, and the

problem of community detection in the Gaussian mixture model is almost transparent to any supervised information on labels as long as the centers are unknown.

As for the upper bound, we introduce and analyze a fully adaptive rate optimal and computationally simple procedure. In order to achieve optimal decay of the risk, it turns out that it is enough to consider $\mathbf{H}(Y^\top Y)$ where for any squared matrix M, $\mathbf{H}(M) = M - \mathrm{diag}(M)$ and $\mathrm{diag}(M)$ is the diagonal of M. We set the initializer η^0 such that $\eta^0 = \mathrm{sign}(\hat{v})$ and $\hat{v}$ is the eigenvector corresponding to the top eigenvalue of $\mathbf{H}(Y^\top Y)$. The risk of η^0 is studied in Section 6.3. In particular, we observe that η^0 can achieve almost full recovery but cannot show it is rate optimal. The lack of rate optimality is probably due to the fact that spectral methods do not benefit from the structure of binary vectors. As an improvement, we consider in Section 6.4 the iterative sequence of estimators $(\eta^k)_{k\geq 1}$ defined as

$$\forall k \geq 0, \quad \eta^{k+1} = \mathrm{sign}\left(\mathbf{H}(Y^\top Y)\eta^k\right).$$

In comparison to Lu and Zhou (2016), we get better results, in particular for large p. In their approach, a spectral initialization on $\boldsymbol{\theta}$ is considered and estimation of $\boldsymbol{\theta}$ is handled at each iteration. The main difference compared to our procedure lies in the fact that we get around the step of estimating $\boldsymbol{\theta}$. We only need the matrix $\mathbf{H}(Y^\top Y)$ that is almost blind to the direction of $\boldsymbol{\theta}$. Giraud and Verzelen (2018) present a rate optimal procedure without capturing the sharp optimality. Our procedure differs in two ways from Giraud and Verzelen (2018). First, it is not an SDP relaxation method and hence is faster. Second, by using the operator $\mathbf{H}$, we do not need to de-bias the Gram matrix, as this operator handles the task.

In Section 6.5, we show the existence of a sharp phase transition for exact recovery in the Gaussian mixture model, around the threshold $\Delta = \bar{\Delta}_n$ such that

$$\bar{\Delta}_n^2 = \sigma^2\left(1 + \sqrt{1 + \frac{2p}{n\log n}}\right)\log n.$$

In particular, this phase transition gives rise to two different regimes around a critical dimension $p^* = n\log n$, showing that the hardness of exact recovery depends on whether p is larger or smaller than p^*.

6.2 Non-asymptotic fundamental limits in the Gaussian mixture model

In this section, we derive a sharp optimal lower bound for the risk Ψ_Δ. As stated in the Introduction, a simple lower bound is given by (6.3). The next proposition provides a sharper statement.

Proposition 6.2.1. *For any $\Delta > 0$, we have*

$$\Psi_\Delta \geq c\Phi^{\mathsf{c}}(\Delta/\sigma),$$

for some $c > 0$.

Following the same lines as in Ndaoud (2019), we obtain two different lower bounds for the minimax risk. Proposition 6.2.1 gives a bound responsible for the hardness of recovering communities due to the lack of information on the labels. It still benefits from the knowledge of $\boldsymbol{\theta}$. In Giraud and Verzelen (2018), it becomes clear that for large p, the hardness of the problem results from the hardness of estimating $\boldsymbol{\theta}$. Hence, in order to capture this phenomenon, one may try to hide the information about the direction of $\boldsymbol{\theta}$ in order to make its estimation difficult.

More precisely, in order to bound the risk Ψ_Δ from below, we place a prior on both η and $\boldsymbol{\theta}$. Ideally, we would choose a Gaussian prior for $\boldsymbol{\theta}$ in order to make its estimation harder, but one should keep in mind that $\boldsymbol{\theta}$ is constrained to the set Ω_Δ. To derive lower bounds on constrained sets, we act as in Butucea et al. (2018). Let $\pi = \pi_{\boldsymbol{\theta}} \times \pi_\eta$ be a product probability measure on $\mathbf{R}^p \times \{-1,1\}^n$ (a prior on $(\boldsymbol{\theta}, \eta)$). We denote by $\mathbb{E}_\pi$ the expectation with respect to π.

Theorem 6.2.1. *Let $\Delta > 0$ and $\pi = \pi_{\boldsymbol{\theta}} \times \pi_\eta$ a product probability measure on $\mathbf{R}^p \times \{-1,1\}^n$. Then,*

$$\Psi_\Delta \geq c \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \inf_{\hat{T}_i \in [-1,1]} \mathbb{E}_\pi \mathbf{E}_{(\boldsymbol{\theta},\eta)} |\hat{T}_i - \eta_i| - \pi_{\boldsymbol{\theta}} \left(\|\boldsymbol{\theta}\| < \Delta \right) \right),$$

where $\inf_{\hat{T}_i \in [-1,1]}$ is the infimum over all estimators $\hat{T}_i(Y)$ with values in $[-1,1]$ and $c > 0$.

Theorem 6.2.1 is useful to derive non-asymptotic lower bounds for constrained minimax risks. For the corresponding lower bound to be optimal, we need the remainder term $\pi_{\boldsymbol{\theta}} \left(\|\boldsymbol{\theta}\| < \Delta \right)$ to be negligible. In other words, the prior on $\boldsymbol{\theta}$ must ensure that $\|\boldsymbol{\theta}\|$ is greater than Δ with high probability. This would make the problem of recovery easier. Hence, it is clear that there exists some trade-off concerning the choice of $\pi_{\boldsymbol{\theta}}$.

Let $\pi^\alpha = \pi_{\boldsymbol{\theta}}^\alpha \times \pi_\eta$ be a product prior on $\mathbf{R}^p \times \{-1,1\}^n$, such that $\pi_{\boldsymbol{\theta}}^\alpha$ is the distribution of the Gaussian random vector with i.i.d. centered entries of variance α^2, π_η is the distribution of the vector with i.i.d. Rademacher entries, and $\boldsymbol{\theta}$ is independent of η. For this specific choice of prior we get the following result.

Proposition 6.2.2. *For any $\alpha > 0$, we have for all $i = 1, \ldots, n$,*

$$\inf_{\hat{T}_i \in [-1,1]} \frac{1}{n} \mathbb{E}_{\pi^\alpha} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |\hat{T}_i - \eta_i| \geq \frac{1}{n} \mathbb{E}_{\pi^\alpha} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |\eta_i^{**} - \eta_i|,$$

where η^{**} is a supervised learning oracle given by

$$\forall i = 1, \ldots, n, \quad \eta_i^{**} = \operatorname{sign} \left(Y_i^\top \left(\sum_{j \neq i} \eta_j Y_j \right) \right).$$

It is interesting to notice that each entry of the supervised learning oracle η^{**} only depends on $\boldsymbol{\theta}$ through its best estimator under the Gaussian prior when the labels for other entries are known. The lower bound of Proposition 6.2.2 confirms the intuition

that the supervised learning oracle is optimal in a minimax sense. For $\sigma > 0$, define G_σ by the relation:

$$\forall t \in \mathbf{R}, \quad G_\sigma(t, \boldsymbol{\theta}) = \mathbf{P}\left((\boldsymbol{\theta} + \sigma\xi_1)^\top \left(\boldsymbol{\theta} + \frac{\sigma}{n-1} \sum_{j=2}^{n} \xi_j \right) \leq \|\boldsymbol{\theta}\|^2 t \right), \tag{6.9}$$

where $\xi_1, \dots, \xi_n$ are i.i.d. standard Gaussian random vectors. Combining Theorem 6.2.1 and Proposition 6.2.2 and using the fact that all entries of the prior π^α are i.i.d. we obtain the next proposition.

Proposition 6.2.3. *Let $\Delta > 0$ and let G_σ be the function defined in (6.9). For any $\alpha > 0$, we have*

$$\Psi_\Delta \geq c\mathbb{E}_{\pi_{\boldsymbol{\theta}}^\alpha} G_\sigma(0, \boldsymbol{\theta}) - c\mathbf{P}\left(\sum_{j=1}^{p} \varepsilon_j^2 \leq \frac{\Delta^2}{\alpha^2} \right),$$

where ε_j are i.i.d. standard Gaussian random variables and $c > 0$.

We are now ready to state the main result of this section. As explained in Giraud and Verzelen (2018), the main limitation of the analysis in Lu and Zhou (2016) is partially due to the choice of the signal-to-noise ratio (SNR) as Δ/σ. We use here the SNR $\mathbf{r}_n$ given in (6.8). It is of the same order as the SNR presented in Giraud and Verzelen (2018).

Theorem 6.2.2. *Let $\Delta > 0$. For n large enough, there exists a sequence ϵ_n such that $\epsilon_n = o(1)$ and*

$$\Psi_\Delta \geq c\Phi^{\mathsf{c}}((\mathbf{r}_n(1 + \epsilon_n)),$$

for some $c > 0$.

It is worth saying that the result of Theorem 6.2.2 holds without any assumption on p and can be interpreted in a non-asymptotic sense by replacing ϵ_n by some small $c > 0$. Moreover, since $\mathbf{r}_n < \Delta/\sigma$, it improves upon the lower bound in Proposition 6.2.1. This improvement is most dramatic in the regime $\Delta^2/\sigma^2 = o(p/n)$ that can be called the hard estimation regime.

6.3 Spectral initialization

In this section, we analyze the non-asymptotic minimax risk of the spectral initializer η^0. As it is the case in SDP relaxations of the problem, the matrix of interest is the Gram matrix $Y^\top Y$. It is well known that is suffers from a bias that grows with p. In Royer (2017), a de-biasing procedure is proposed using an estimator of the covariance of the noise. This step is important to obtain a procedure adaptive to the noise level. Our approach is different but is still adaptive and consists in removing the diagonal entries of the Gram matrix. We give here some intuition about this procedure. Define the linear operator $\mathbf{H} : \mathbf{R}^{n\times n} \to \mathbf{R}^{n\times n}$ as follows:

$$\forall M \in \mathbf{R}^{n\times n}, \quad \mathbf{H}(M) = M - \mathrm{diag}(M),$$

where $\text{diag}(M)$ is a diagonal matrix with the same diagonal as M. Going back to Proposition 6.2.2, we may observe that the oracle η^{**} can be written as

$$\eta^{**} = \text{sign}\left(\mathbf{H}\left(Y^\top Y\right)\eta\right), \tag{6.10}$$

where the sign is applied entry-wise. This suggests that the matrix $\mathbf{H}\left(Y^\top Y\right)$ appears in a natural way. We can decompose it as follows:

$$\mathbf{H}(Y^\top Y) = \|\boldsymbol{\theta}\|^2\eta\eta^\top + \mathbf{H}(W^\top W) + \mathbf{H}(W^\top\boldsymbol{\theta}\eta^\top + \eta\boldsymbol{\theta}^\top W) - \|\boldsymbol{\theta}\|^2\mathbf{I}_n. \tag{6.11}$$

Apart from the scalar factor $\|\boldsymbol{\theta}\|^2$, this expression is similar to SBM or symmetric spiked model, with the noise having a more complex structure. It turns out that the main driver of the noise is $\mathbf{H}(W^\top W)$. A simple lemma (cf. Appendix 6.9) shows that our approach is probably an alternative to de-biasing the Gram matrix. Specifically, Lemma 6.9.2 gives that

$$\|\mathbf{H}(W^\top W)\|_{op} \leq 2\left\|W^\top W - \mathbf{E}\left(W^\top W\right)\right\|_{op}$$

for any random matrix W with independent columns. Hence, the noise term can be controlled as if its covariance were known. Nevertheless, the operator $\mathbf{H}(.)$ may affect dramatically the signal since it also removes its diagonal entries. Fortunately, the signal term is almost insensitive to this operation since it is a rank-one matrix where the spike energy is spread all over the spike. For instance, we have

$$\|\mathbf{H}(\eta\eta^\top)\|_{op} = \left(1 - \frac{1}{n}\right)\|\eta\eta^\top\|_{op}.$$

Hence as n grows the signal does not get affected by removing the diagonal terms while we get rid of the bias in the noise. It is worth noticing that our approach succeeds thanks to the specific form of η and cannot be generalized to any spiked model. For the general case, a more consistent approach is proposed in Zhang et al. (2018), where the diagonal entries can be used to achieve optimal estimation accuracy. Motivated by (6.11), the spectral estimator η^0 is defined by

$$\eta^0 = \text{sign}(\hat{v}), \tag{6.12}$$

where $\hat{v}$ is the eigenvector corresponding to the top eigenvalue of $\mathbf{H}(Y^\top Y)$. The next result characterizes the non-asymptotic minimax risk of η^0.

Theorem 6.3.1. *Let $\Delta > 0$ and let η^0 be the estimator given by (6.12). Under the condition $\mathbf{r}_n \geq C$, for some absolute constant $C > 0$, we have*

$$\sup_{(\boldsymbol{\theta},\eta)\in\Omega_\Delta}\frac{1}{n}\mathbf{E}_{(\boldsymbol{\theta},\eta)}r(\eta^0,\eta) \leq \frac{C'}{\mathbf{r}_n^2} + \frac{32}{n^2},$$

and

$$\sup_{(\boldsymbol{\theta},\eta)\in\Omega_\Delta}\mathbf{P}_{(\boldsymbol{\theta},\eta)}\left(\frac{1}{n}\left|\eta^\top\eta^0\right| \leq 1 - \frac{\log n}{n} - \frac{C'}{\mathbf{r}_n^2}\right) \leq \epsilon_n\Phi^{\mathsf{c}}(\mathbf{r}_n),$$

for some sequence ϵ_n such that $\epsilon_n = o(1)$ and $C' > 0$.

As we may expect the appropriate Hamming distance risk is decreasing with respect to $\mathbf{r}_n$. The residual term $\frac{32}{n^2}$ is due to removing the diagonal and can be seen as the price to pay for adaptation. It is obvious that as $\mathbf{r}_n$ gets larger than n, removing the diagonal terms is sub-optimal.

As $n, \mathbf{r}_n \to \infty$, η^0 achieves almost full recovery (cf. Definition 6.7.1). We show later that this condition is optimal but cannot show that η^0 is rate optimal. In particular, it is not clear whether η^0 can achieve exact recovery. To bring some evidence that η^0 cannot achieve exact recovery, we rely on asymptotic random matrix theory. In Benaych-Georges and Nadakuditi (2012), it is shown that, in the asymptotics when $p/n \to c \in (0, 1]$ and when the noise is Gaussian, detection is possible only for $\Delta^2 \geq \sqrt{c}\sigma^2$. Moreover, the asymptotic correlation between η and its spectral approximation is given by $\sqrt{1 - \frac{c\sigma^2+\Delta^2}{\Delta^2(1+\Delta^2/\sigma^2)}}$. When $\mathbf{r}_n = \Omega(1)$, we observe that $\frac{c\sigma^2+\Delta^2}{\Delta^2(1+\Delta^2/\sigma^2)} \asymp \frac{1}{\mathbf{r}_n^2}$. Hence, the decay in Theorem 6.3.1 is expected for general spiked models, but not necessarily rate optimal in our specific setting. The condition $\mathbf{r}_n = \Omega(1)$ is very natural, since it is necessary even for detection as shown in Banks et al. (2018).

6.4 A rate optimal practical algorithm

In this section, we present an algorithm that is minimax optimal, adaptive to Δ and σ and faster than SDP relaxation. In the same spirit as in Lu and Zhou (2016), we are tempted by using Lloyd's iterations. If properly initialized, Lloyd's algorithm may achieve the optimal rate under mild conditions after only a logarithmic number of steps. We present here a variant of Lloyd's iterations. Motivated by (6.10), and given an estimator $\hat{\eta}^0$, we define a sequence of estimators $(\hat{\eta}^k)_{k\geq 0}$ such that

$$\forall k \geq 0, \quad \hat{\eta}^{k+1} = \operatorname{sign}\left(\mathbf{H}\left(Y^\top Y\right)\hat{\eta}^k\right). \tag{6.13}$$

Notice that Lloyd's iterations correspond to the procedure (6.13), where $\mathbf{H}\left(Y^\top Y\right)$ is replaced by $Y^\top Y$. If the initialization is good in a sense that we describe below, then at each iteration $\hat{\eta}^k$ gets closer to η and achieves the minimax optimal rate after a logarithmic number of steps. The logarithmic number of steps is crucial computationally as it is the case in many other iterative procedures.

Theorem 6.4.1. *Let $\Delta > 0$ and let $\hat{\eta}^0$ be an estimator satisfying*

$$\frac{1}{n}\eta^\top \hat{\eta}^0 \geq 1 - \frac{C'}{\mathbf{r}_n^2} - \nu_n$$

for some $C' > 0$ and $\nu_n = o(1)$. Let $(\hat{\eta}^k)_{k\geq 0}$ be the corresponding iterative sequence (6.13). If $\mathbf{r}_n \geq C$ for some $C > 0$, then after $k = \lfloor 3 \log n \rfloor$ steps, we have

$$\sup_{(\boldsymbol{\theta},\eta)\in\Omega_\Delta} \mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\hat{\eta}^k, \eta) \leq C'\mathbf{r}_n^2 \sup_{\|\boldsymbol{\theta}\|\geq\Delta} G_\sigma\left(\epsilon_n + \frac{C'}{\mathbf{r}_n}, \boldsymbol{\theta}\right) + \epsilon_n \Phi^{\mathsf{c}}(\mathbf{r}_n),$$

for some sequence ϵ_n such that $\epsilon_n = o(1)$ and $C' > 0$.

Recall that $G(t, \boldsymbol{\theta})$ is close to $G(0, \boldsymbol{\theta})$ for small t. Theorem 6.4.1 can be interpreted as follows. Given a good initialization, the iterative procedure (6.13) achieves an error close

to the supervised learning risk within a logarithmic number of steps. Observing that under the condition $\mathbf{r}_n \geq C$ for some $C > 0$, the spectral estimator η^0 is a good initializer, we state a general result showing that our variant of Lloyd's iterations initialized with a spectral estimator is minimax optimal.

Theorem 6.4.2. *Let $\Delta > 0$. Let η^0 be the spectral estimator defined in (6.12) and let $(\eta^k)_{k\geq 0}$ be the iterative sequence (6.13). Assume that $\mathbf{r}_n > C$ for some $C > 0$. Then, after $k = \lfloor 3 \log n \rfloor$ steps we have*

$$\sup_{(\boldsymbol{\theta},\eta)\in\Omega_\Delta} \mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\eta^k, \eta) \leq C' \Phi^{\mathsf{c}} \left(\mathbf{r}_n \left(1 - \epsilon_n - \frac{C' \log \mathbf{r}_n}{\mathbf{r}_n} \right) \right),$$

for some sequence ϵ_n such that $\epsilon_n = o(1)$ and $C' > 0$.

Notice that the upper bound in Theorem 6.4.2 is almost optimal, and gets closer to the optimal minimax rate as $n, \mathbf{r}_n \to \infty$. Hence, under mild conditions, we get a matching upper bound to the lower bound in Theorem 6.2.2. Moreover, we figure out that a good initialization combined with smart iterations is almost equivalent to the supervised learning oracle. In fact, the rate in Theorem 6.4.2 is almost the same as the rate of the supervised oracle η^{**}. We conclude that unsupervised learning is asymptotically as easy as supervised learning in the Gaussian mixture model. The next proposition gives a full picture of the minimax risk Ψ_Δ.

Proposition 6.4.1. *Let $\Delta > 0$. For some $c_1, c_2, C_1, C_2 > 0$ and n large enough, we have*

$$C_1 e^{-c_1 \mathbf{r}_n^2} \leq \Psi_\Delta \leq C_2 e^{-c_2 \mathbf{r}_n^2}.$$

Notice that the procedure we present here has a different rate of decay compared to the spectral procedure (6.12), that may be non-asymptotically sub-optimal. Recent papers by Xia and Zhou (2017) and Abbe et al. (2017) show that a simple spectral algorithm can achieve exact recovery using refined sup-norm perturbation techniques. Although their results are striking, they match the optimal conditions for exact recovery in the Gaussian mixture model only in the zone $\mathbf{r}_n \asymp \Delta/\sigma$.

6.5 Asymptotic analysis. Phase transitions

This section deals with asymptotic analysis of the problem of community detection in the two component Gaussian mixture model. The results are derived as corollaries of the minimax bounds of previous sections. We will assume that $n \to \infty$ and that parameters p, σ and Δ depend on n. For the sake of readability we do not equip some parameters with the index n.

The two asymptotic properties we study here are *exact recovery* and *almost full recovery*. The complete characterization of the sharp phase transition for almost full recovery is deferred to Section 6.7. We use the terminology following Butucea et al. (2018) that we recall here.

Definition 6.5.1. *Let $(\Omega_{\Delta_n})_{n\geq 2}$ be a sequence of classes corresponding to $(\Delta_n)_{n\geq 2}$:*

- *We say that* exact recovery is possible *for* $(\Omega_{\Delta_n})_{n\geq 2}$ *if there exists an estimator* $\hat{\eta}$ *such that*

$$\lim_{n\to\infty} \sup_{(\boldsymbol{\theta},\eta)\in\Omega_{\Delta_n}} \mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\hat{\eta},\eta) = 0. \tag{6.14}$$

  In this case, we say that $\hat{\eta}$ *achieves exact recovery.*

- *We say that* exact recovery is impossible *for* $(\Omega_{\Delta_n})_{n\geq 2}$ *if*

$$\liminf_{n\to\infty} \inf_{\tilde{\eta}} \sup_{(\boldsymbol{\theta},\eta)\in\Omega_{\Delta_n}} \mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\tilde{\eta},\eta) > 0, \tag{6.15}$$

  where $\inf_{\tilde{\eta}}$ *denotes the infimum over all estimators in* $\{-1,1\}^n$.

Informally, we would like to get a "phase transition" value $\bar{\Delta}_n$ such that exact recovery is possible for Δ_n greater than $\bar{\Delta}_n$ and is impossible for Δ_n smaller than $\bar{\Delta}_n$. Our aim now is to find such "phase transition" values. For the problem of exact recovery, the "phase transition" is described in the next theorem. Let $\bar{\Delta}_n > 0$ be defined by

$$\bar{\Delta}_n^2 = \sigma^2 \left(1 + \sqrt{1 + \frac{2p}{n\log n}}\right) \log n. \tag{6.16}$$

The next theorem is a direct consequence of Theorem 6.7.1, cf. Section 6.7.

Theorem 6.5.1. *(i) If* $\Delta_n \geq \bar{\Delta}_n(1+\epsilon)$ *for some* $\epsilon > 0$. *Then, the estimator* η^k *defined in* (6.12)-(6.13), *with* $k = \lfloor 3\log n \rfloor$, *achieves exact recovery.*

(ii) If the complementary condition holds, i.e, $\Delta_n \leq \bar{\Delta}_n(1-\epsilon)$ *for some* $\epsilon > 0$, *then exact recovery is impossible.*

Some remarks are in order here. First of all, Theorem 6.5.1 shows that the "phase transition" for exact recovery occurs at $\bar{\Delta}_n$ given in (6.16). It is worth noticing that this sharp threshold for exact recovery holds for all values of p. In particular, there exists a critical dimension $p^* = n\log n$. If $p = o(p^*)$, then $\bar{\Delta}_n = (1+o(1))\sigma\sqrt{2\log n}$. In this case, the phase transition threshold for exact recovery, is the same as if $\boldsymbol{\theta}$ were known. While if $p^* = o(p)$, then $\bar{\Delta}_n = (1+o(1))\sigma\left(\frac{2p\log n}{n}\right)^{1/4}$. This new condition reflects the hardness of estimation, and p^* can be interpreted as a phase transition with respect to the cluster dimension p.

6.6 Discussion and open problems

A key objective of this paper was to establish sharp phase transition for exact recovery in the two component Gaussian mixture model. All upper bounds remain valid in the case of sub-Gaussian noise. It would be interesting to generalize the methodology used to derive both lower and upper bounds to the case of multiple communities and general covariance structure of the noise. We also expect the procedure (6.12)-(6.13) to achieve exact recovery in asymptotically sharp way in other problems, for instance in the Bipartite Stochastic Block Model.

We conclude this paper with an open question. Let $p^* = n \log n$. In the regime $p^* = o(p)$, we proved that for any $\epsilon > 0$, the condition

$$\Delta^2 \geq (1-\epsilon)\sigma^2 \left(\frac{2p}{p^*}\right)^{1/2} \log n$$

is necessary to achieve exact recovery. This is a consequence of considering a Gaussian prior on $\boldsymbol{\theta}$ which makes recovering its direction the hardest. We give here a heuristics that this should hold independently on the choice of prior as long as $\boldsymbol{\theta}$ is uniformly well-spread (i.e., not sparse). Suppose that we put a Rademacher prior on $\boldsymbol{\theta}$ such that $\boldsymbol{\theta} = \frac{\Delta}{\sqrt{p}}\zeta$, where ζ is a random vector with i.i.d. Rademacher entries. Following the same argument as in Proposition 6.2.1, it is clear that a necessary condition to get non-trivial correlation with ζ is given by

$$\Delta^2 \geq c\sigma^2 \frac{p}{n},$$

for some $c > 0$. Observing that, in the hard estimation regime, we have

$$\left(\frac{p}{p^*}\right)^{1/2} \log n = o\left(\frac{p}{n}\right),$$

it comes that, while exact recovery of η is possible, non-trivial correlation with ζ is impossible. Consequently, there is no hope achieving exact recovery through non-trivial correlation with $\boldsymbol{\theta}$ in the hard estimation regime.

Conjecture 6.6.1. *Let $\Delta > 0$. Assume that Y follows model (6.1). Let η be a random vector with i.i.d. Rademacher random entries, and $\boldsymbol{\theta} = \frac{\Delta}{\sqrt{p}}\zeta$ where ζ is a random vector with i.i.d. Rademacher entries and independent of η. Assume that $n \log n = o(p)$. Prove or disprove that, for any $\epsilon > 0$,*

$$\Delta^2 \geq (1-\epsilon)\sigma^2 \sqrt{\frac{2p \log n}{n}}$$

is necessary to achieve exact recovery.

In particular, a positive answer to the previous question will be very useful to derive optimal conditions for exact recovery in bipartite graph models among other problems.

6.7 Asymptotic analysis: almost full recovery

In this section, we conduct the asymptotic analysis of the problem of almost full recovery in the two component Gaussian mixture model. We first recall the terminology used in Butucea et al. (2018) that we adopt for the problem of almost full recovery.

Definition 6.7.1. *Let $(\Omega_{\Delta_n})_{n\geq 2}$ be a sequence of classes corresponding to $(\Delta_n)_{n\geq 2}$:*

- *We say that* almost full recovery is possible *for $(\Omega_{\Delta_n})_{n\geq 2}$ if there exists an estimator $\hat{\eta}$ such that*

$$\lim_{n\to\infty} \sup_{(\boldsymbol{\theta},\eta)\in\Omega_{\Delta_n}} \frac{1}{n}\mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\hat{\eta},\eta) = 0. \tag{6.17}$$

  In this case, we say that $\hat{\eta}$ achieves almost full recovery.

- *We say that* almost full recovery is impossible *for* $(\Omega_{\Delta_n})_{n\geq 2}$ *if*

$$\liminf_{n\to\infty} \inf_{\tilde{\eta}} \sup_{(\boldsymbol{\theta},\eta)\in\Omega_{\Delta_n}} \frac{1}{n}\mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\tilde{\eta},\eta) > 0, \tag{6.18}$$

where $\inf_{\tilde{\eta}}$ *denotes the infimum over all estimators in* $\{-1,1\}^n$.

The following general characterization theorem is a straightforward corollary of the results of previous sections.

Theorem 6.7.1. (i) *Almost full recovery is possible for* $(\Omega_{\Delta_n})_{n\geq 2}$ *if and only if*

$$\Phi^{\mathsf{c}}(\mathbf{r}_n) \to 0 \qquad \text{as } n\to\infty. \tag{6.19}$$

In this case, the estimator η^k *defined in* (6.12)-(6.13)*, with* $k = \lfloor 3\log n \rfloor$*, achieves almost full recovery.*

(ii) *Exact recovery is impossible for* $(\Omega_{\Delta_n})_{n\geq 2}$ *if for some* $\epsilon > 0$

$$\liminf_{n\to\infty} n\Phi^{\mathsf{c}}(\mathbf{r}_n(1+\epsilon)) > 0 \qquad \text{as } n\to\infty, \tag{6.20}$$

and possible if for some $\epsilon > 0$

$$n\Phi^{\mathsf{c}}(\mathbf{r}_n(1-\epsilon)) \to 0 \qquad \text{as } n\to\infty, \tag{6.21}$$

In this case, the estimator η^k *defined in* (6.12)-(6.13)*, with* $k = \lfloor 3\log n \rfloor$*, achieves exact recovery.*

Although this theorem gives a complete solution to the problem of almost full and exact recovery, conditions (6.19), (6.20) and (6.21) are not quite explicit. The next theorem is a consequence of Theorem 6.7.1. It describes a "phase transition" for Δ_n in the problem of almost full recovery.

Theorem 6.7.2. *(i) If* $\sigma^2\left(1+\sqrt{p/n}\right) = o(\Delta_n^2)$*. Then, the estimator* η^k *defined in* (6.12)-(6.13)*, with* $k = \lfloor 3\log n \rfloor$*, achieves almost full recovery.*

(ii) Moreover, if $\Delta_n^2 = \mathcal{O}\left(\sigma^2(1+\sqrt{p/n})\right)$*. Then, almost full recovery is impossible.*

Theorem 6.7.2 shows that almost full recovery occurs if and only if

$$\sigma^2\left(1+\sqrt{p/n}\right) = o(\Delta_n^2). \tag{6.22}$$

6.8 Appendix: Main proofs

In all the proofs of lower bounds, we follow the same argument as in Theorem 2 in Gao et al. (2018) in order to substitute the minimax risk of $r(\tilde{\eta}, \eta)$ by a Hamming minimax risk. Let z^* be a vector of labels in $\{-1, 1\}^n$ and let T be a subset of $\{1, \dots, n\}$ of size $\lfloor n/2 \rfloor + 1$. A lower bound of the minimax risk is given on the subset of labels $\mathbf{Z}$, such that for all $i \in T$,we have $\eta_i = z_i^*$. Observe that in that case

$$r(\eta^1, \eta^2) = |\eta^1 - \eta^2|,$$

for any $\eta^1, \eta^2 \in \mathbf{Z}$. The argument in Gao et al. (2018), leads to

$$\Psi_\Delta \geq \frac{c}{|T^{\mathsf{c}}|} \sum_{i \in T^{\mathsf{c}}} \inf_{\tilde{\eta}_i} \mathbb{E}_\pi \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\tilde{\eta}_i - \eta_i|,$$

for some $c > 0$ and for any prior π such that π_η is invariant by a sign change. That is typically the case under Rademacher prior on labels. As a consequence, a lower bound of Ψ_Δ is given by a lower bound of the R.H.S minimax Hamming risk.

Proof of Proposition 6.2.1

Let $\bar{\boldsymbol{\theta}}$ be a vector in $\mathbf{R}^p$ such that $\|\bar{\boldsymbol{\theta}}\| = \Delta$. Placing an independent Rademacher prior π on η, and fixing $\boldsymbol{\theta}$, it follows that

$$\inf_{\tilde{\eta}_j} \mathbb{E}_\pi \mathbf{E}_{(\bar{\boldsymbol{\theta}}, \eta)} |\tilde{\eta}_j - \eta_j| \geq \inf_{\bar{\eta}_j} \mathbb{E}_\pi \mathbf{E}_{(\bar{\boldsymbol{\theta}}, \eta)} |\bar{\eta}_j(Y_j) - \eta_j|, \tag{6.23}$$

where $\bar{\eta}_j \in [-1, 1]$. The last inequality holds because of independence between the priors. We define, for $\epsilon \in \{-1, 1\}$, $\tilde{f}_\epsilon(.)$ the density of the observation Y_j conditionally on the value of $\eta_j = \epsilon$. Now, using Neyman-Pearson lemma and the explicit form of $\tilde{f}_\epsilon$, we get that the selector η^* given by

$$\eta_j^* = \operatorname{sign}\left(\bar{\boldsymbol{\theta}}^\top Y_j\right), \quad \forall j = 1, \dots, n$$

is the optimal selector that achieves the minimum of the RHS of (6.23). Plugging this value in (6.23), we get further that

$$\inf_{\bar{\eta}_j} \mathbf{E}_\pi |\bar{\eta}_j(Y_j) - \eta_j| = 2\Phi^{\mathsf{c}}(\Delta/\sigma).$$

Proof of Theorem 6.2.1

Throughout the proof, we write for brevity $A = \Omega_\Delta$. Set $\eta^A = \eta \mathbf{1}((\boldsymbol{\theta}, \eta) \in A)$ and denote by $\bar{\pi}_A$ the probability measure π conditioned by the event $\{(\boldsymbol{\theta}, \eta) \in A\}$, that is, for any $C \subseteq \mathbf{R}^p \times \{-1, 1\}^n$,

$$\bar{\pi}_A(C) = \frac{\pi(\{(\boldsymbol{\theta}, \eta) \in C\} \cap \{(\boldsymbol{\theta}, \eta) \in A\})}{\pi((\boldsymbol{\theta}, \eta) \in A)}.$$

The measure $\bar{\pi}_A$ is supported on A and we have

$$\begin{aligned} \inf_{\tilde{\eta}_j} \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\tilde{\eta}_j - \eta_j| &\geq \inf_{\tilde{\eta}_j} \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\tilde{\eta}_j - \eta_j^A| \\ &\geq \inf_{\hat{T}_j} \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{T}_j - \eta_j^A| \end{aligned}$$

where $\inf_{\hat{T}_j}$ is the infimum over all estimators $\hat{T}_j = \hat{T}_j(Y)$ with values in $\mathbf{R}$. According to Theorem 1.1 and Corollary 1.2 on page 228 in Lehmann and Casella (2006), there exists a Bayes estimator $B_j^A = B_j^A(Y)$ such that

$$\inf_{\hat{T}_j} \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |\hat{T}_j - \eta_j^A| = \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |B_j^A - \eta_j^A|,$$

and this estimator is a conditional median of η_j^A given Y. Therefore,

$$\Psi_\Delta \geq c \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{j=1}^{\lfloor n/2 \rfloor} \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |B_j^A - \eta_j^A| \right). \tag{6.24}$$

Note that $B_j^A \in [-1,1]$ since η_j^A takes its values in $[-1,1]$. Using this, we obtain

$$\begin{aligned}
\inf_{\hat{T}_j \in [-1,1]} \mathbb{E}_\pi \mathbf{E}_{(\boldsymbol{\theta},\eta)} |\hat{T}_j - \eta_j| &\leq \mathbb{E}_\pi \mathbf{E}_{(\boldsymbol{\theta},\eta)} |B_j^A - \eta_j| \\
&= \mathbb{E}_\pi \mathbf{E}_{(\boldsymbol{\theta},\eta)} \Big(|B_j^A - \eta_j| \mathbf{1}((\boldsymbol{\theta},\eta) \in A) \Big) + \mathbb{E}_\pi \mathbf{E}_{(\boldsymbol{\theta},\eta)} \Big(|B_j^A - \eta_j| \mathbf{1}((\boldsymbol{\theta},\eta) \in A^c) \Big) \\
&\leq \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |B_j^A - \eta_j^A| + \mathbb{E}_\pi \mathbf{E}_{(\boldsymbol{\theta},\eta)} \Big(|B_j^A - \eta_j| \mathbf{1}((\boldsymbol{\theta},\eta) \in A^c) \Big) \\
&\leq \mathbb{E}_{\bar{\pi}_A} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |B_j^A - \eta_j^A| + 2\mathbf{P}((\boldsymbol{\theta},\eta) \notin A). \qquad (6.25)
\end{aligned}$$

The result follows combining (6.24) and (6.25).

Proof of Proposition 6.2.2

We start by using the fact that

$$\mathbb{E}_{\pi^\alpha} \mathbf{E}_{(\boldsymbol{\theta},\eta)} |\hat{\eta}_i - \eta_i| = \mathbb{E}_{p_{-i}} \mathbb{E}_{p_i} \mathbf{E}_{(\boldsymbol{\theta},\eta)} \left(|\hat{\eta}_i - \eta_i| | (\eta_j)_{j \neq i} \right),$$

where p_i is the marginal of π^α on $(\boldsymbol{\theta},\eta_i)$, while p_{-i} is the marginal of π^α on $(\eta_j)_{j \neq i}$. Using the independence between different priors, one may observe that $\pi^\alpha = p_i \times p_{-i}$. We define, for $\epsilon \in \{-1,1\}$, $\tilde{f}_\epsilon^i$ the density of the observation Y given $(\eta_j)_{j \neq i}$ and given $\eta_i = \epsilon$. Using Neyman-Pearson lemma, we get that

$$\eta_i^{**} = \begin{cases} 1 & \text{if } \tilde{f}_1^i(Y) \geq \tilde{f}_{-1}^i(Y), \\ -1 & \text{else,} \end{cases}$$

minimizes $\mathbb{E}_{p_i} \mathbf{E}_{(\boldsymbol{\theta},\eta)} \left(|\hat{\eta}_i - \eta_i| | (\eta_j)_{j \neq i} \right)$ over all functions of $(\eta_j)_{j \neq i}$ and of Y with values in $[-1,1]$. Using the independence of the rows of Y we have

$$\tilde{f}_\epsilon^i(Y) = \prod_{j=1}^p \frac{e^{-\frac{1}{2} L_j^\top \Sigma_\epsilon^{-1} L_j}}{(2\pi)^{p/2} |\Sigma_\epsilon|},$$

where L_j is the j-th row of Y and $\Sigma_\epsilon = \mathbf{I}_n + \alpha^2 \eta_\epsilon \eta_\epsilon^\top$. We denote by η_ϵ the binary vector such that $\eta_i = \epsilon$ and the other components are known. It is easy to check that $|\Sigma_\epsilon| = 1 + \alpha^2 n$, hence it does not depend on ϵ. A simple calculation leads to

$$\Sigma_\epsilon^{-1} = \mathbf{I}_n - \frac{\alpha^2}{1 + \alpha^2 n} \eta_\epsilon \eta_\epsilon^\top.$$

Hence

$$\begin{aligned}\frac{\tilde{f}_1^i(Y)}{\tilde{f}_{-1}^i(Y)} &= \prod_{j=1}^{p} e^{-\frac{1}{2}L_j^\top(\Sigma_1^{-1}-\Sigma_{-1}^{-1})L_j} \\ &= \prod_{j=1}^{p} e^{\frac{\alpha^2}{1+\alpha^2 n}L_{ji}\sum_{k\neq i}L_{jk}\eta_k} \\ &= e^{\frac{\alpha^2}{1+\alpha^2 n}\sum_{k\neq i}\eta_k\sum_{j=1}^{p}L_{jk}L_{ji}} = e^{\frac{\alpha^2}{1+\alpha^2 n}\langle Y_i,\sum_{k\neq i}\eta_k Y_k\rangle}.\end{aligned}$$

It is now immediate that

$$\eta_i^{**} = \operatorname{sign}\left(Y_i^\top\left(\sum_{k\neq i}\eta_k Y_k\right)\right).$$

Proof of Proposition 6.2.3

Combining Theorem 6.2.1 and Proposition 6.2.2, we get that

$$\Psi_\Delta \geq c\left(\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{E}_{\pi^\alpha}\mathbf{E}_{(\boldsymbol{\theta},\eta)}|\eta_i^{**}-\eta_i| - \pi_{\boldsymbol{\theta}}^{\alpha}\left(\|\boldsymbol{\theta}\|\leq\Delta\right)\right).$$

Recall that here $\boldsymbol{\theta}$ has i.i.d. centered Gaussian entries with variance α^2. This yields the second term on the R.H.S of the inequality of Proposition 6.2.3. While, for the first term, one may notice that the vectors $\eta_i Y_i$ for $i = 1,\ldots,n$ are i.i.d. and that

$$|\eta_i^{**}-\eta_i| = 2\mathbf{1}\left(\eta_i Y_i^\top\left(\sum_{j\neq i}\eta_j Y_j\right)\leq 0\right).$$

Then, we use the definition of G_σ (6.9) in order to conclude.

Proof of Theorem 6.2.2

We prove the result by considering separately the following three cases.

1. Case $\Delta \leq \frac{\log^2(n)}{\sqrt{n}}$. In this case we use Proposition 6.2.1.

   Since $0 \leq \frac{\Delta^2}{\sqrt{\Delta^2+p/n}} \leq \Delta$, we have $\left|\Delta - \frac{\Delta^2}{\sqrt{\Delta^2+p/n}}\right| \leq \frac{\log^2(n)}{\sqrt{n}}$. Hence

   $$\left|\Phi^{\mathsf{c}}(\Delta) - \Phi^{\mathsf{c}}\left(\frac{\Delta^2}{\sqrt{\Delta^2+p/n}}\right)\right| \leq c\frac{\log^2(n)}{\sqrt{n}}\Phi^{\mathsf{c}}\left(\frac{\Delta^2}{\sqrt{\Delta^2+p/n}}\right),$$

   for some $c > 0$. Hence we get the result with $\epsilon_n = c\frac{\log^2(n)}{\sqrt{n}}$.

2. Case $\Delta \geq \sqrt{\frac{p\log n}{n}}$. In this case, we have $\sqrt{1+\frac{p}{n\Delta^2}}\frac{\Delta^2}{\sqrt{\Delta^2+p/n}} = \Delta$. It is easy to check that

   $$\left|\sqrt{1+\frac{p}{n\Delta^2}} - 1\right| \leq \frac{1}{\log n}.$$

Hence

$$\Delta \leq \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}}(1+\epsilon_n),$$

for $\epsilon_n = \frac{1}{\log n}$. We conclude using Proposition 6.2.1.

3. Case $\frac{\log^2(n)}{\sqrt{n}} < \Delta < \sqrt{\frac{p\log n}{n}}$. Notice that $p \geq \log^3(n)$ in this regime. We will use Proposition 6.2.3. Set α^2 such that

$$\alpha^2 = \frac{\Delta^2}{p(1-\nu_n)} \quad \text{and} \quad \nu_n = \sqrt{\frac{n\Delta^2}{p\log^2(n)}}.$$

It is easy to check that $0 < \nu_n^2 \leq 1/\log n$, Hence

$$\mathbf{P}\left(\sum_{j=1}^p \varepsilon_j^2 \leq \frac{\Delta^2}{\alpha^2}\right) = \mathbf{P}\left(\frac{1}{p}\sum_{j=1}^p(\varepsilon_j^2 - 1) \leq -\nu_n\right) \leq e^{-c\frac{n}{\log^2(n)}\Delta^2},$$

for some $c > 0$. Hence, for any $\epsilon_n \to 0$ we have

$$\mathbf{P}\left(\sum_{j=1}^p \varepsilon_j^2 \leq \frac{\Delta^2}{\alpha^2}\right) \leq e^{-c'\log n}\Phi^{\mathsf{c}}\left(\Delta(1+\epsilon_n)\right) \leq e^{-c'\log n}\Phi^{\mathsf{c}}\left(\frac{\Delta^2}{\sqrt{\Delta^2+p/n}}(1+\epsilon_n)\right),$$

for some $c' > 0$. Since $e^{-c'\log n} \underset{n\to\infty}{\to} 0$, then in order to conclude, we just need to prove that

$$\mathbb{E}_{\pi_{\boldsymbol{\theta}}^{\alpha}} G_\sigma(0,\boldsymbol{\theta}) \geq (1-\epsilon_n)\Phi^{\mathsf{c}}\left(\frac{\Delta^2}{\sqrt{\Delta^2+p/n}}(1+\epsilon_n)\right),$$

for some sequence $\epsilon_n \to 0$.

We recall that

$$\mathbb{E}_{\pi_{\boldsymbol{\theta}}^{\alpha}} G_\sigma(0,\boldsymbol{\theta}) = \mathbf{P}\left((\boldsymbol{\theta}+\xi_1)^\top\left(\boldsymbol{\theta}+\frac{\xi_2}{\sqrt{n-1}}\right) \leq 0\right),$$

where ξ_1, ξ_2 are two independent random vectors with i.i.d. standard Gaussian entries and $\boldsymbol{\theta}$ is an independent Gaussian prior. Moreover, using independence, we have

$$\mathbf{P}\left((\boldsymbol{\theta}+\xi_1)^\top\left(\boldsymbol{\theta}+\frac{\xi_2}{\sqrt{n-1}}\right) \leq 0\right) = \mathbf{P}\left(\varepsilon\sqrt{\|\boldsymbol{\theta}\|^2 + \frac{\|\xi_2\|^2}{n-1} + \frac{2}{\sqrt{n-1}}\boldsymbol{\theta}^\top\xi_2} \geq \|\boldsymbol{\theta}\|^2 + \frac{1}{\sqrt{n-1}}\boldsymbol{\theta}^\top\xi_2\right)$$

where ε is a standard Gaussian random variable. Fix $\boldsymbol{\theta}$ and define the random event

$$\mathcal{A} = \left\{\frac{\|\xi_2\|^2}{n-1} \geq \frac{p}{n-1}(1-\zeta_n)\right\} \cap \left\{|\boldsymbol{\theta}^\top\xi_2| \leq \sqrt{n-1}\beta_n\|\boldsymbol{\theta}\|^2\right\},$$

where $\beta_n > 0$ and $\zeta_n \in (0,1)$. It is easy to check that

$$\mathbf{P}\left(\mathcal{A}^{\mathsf{c}}\right) \leq e^{-c\log^3(n)\zeta_n^2} + e^{-c\beta_n^2 n\|\boldsymbol{\theta}\|^2}, \tag{6.26}$$

for some $c > 0$. Hence conditioning on $\boldsymbol{\theta}$, we have

$$\mathbf{P}\left((\boldsymbol{\theta}+\xi_1)^\top\left(\boldsymbol{\theta}+\frac{\xi_2}{\sqrt{n-1}}\right)\leq 0\right) \geq \mathbf{E}\left[\Phi^{\mathsf{c}}\left(\frac{\|\boldsymbol{\theta}\|^2(1+\beta_n)}{\sqrt{\|\boldsymbol{\theta}\|^2(1-2\beta_n)+\frac{p}{n-1}(1-\zeta_n)}}\right)\mathbf{P}(\mathcal{A})\right].$$

where the last expectation is over $\boldsymbol{\theta}$. Define now the random event $\mathcal{B} = \{|\|\boldsymbol{\theta}\|^2 - \Delta^2| \leq \Delta^2\gamma_n\}$ where $\gamma_n \in (0,1)$. Then, using (6.26), we get

$$\mathbf{P}\left((\boldsymbol{\theta}+\xi_1)^\top\left(\boldsymbol{\theta}+\frac{\xi_2}{\sqrt{n-1}}\right)\leq 0\right) \geq \Phi^{\mathsf{c}}\left(U_n\right)\left(1-e^{-c\log^3(n)\zeta_n^2}-e^{-c\beta_n^2(1-\gamma_n)\log^4(n)}\right)\mathbf{P}(\mathcal{B}), \tag{6.27}$$

where $U_n := \frac{\Delta^2(1+\beta_n)(1+\gamma_n)}{\sqrt{\Delta^2(1-2\beta_n)(1-\gamma_n)+\frac{p}{n-1}(1-\zeta_n)}}$. Now we may check that

$$\mathbf{P}(\mathcal{B}^{\mathsf{c}}) = \mathbf{P}\left(\left|\sum_{j=1}^p \varepsilon_j^2 - \frac{\Delta^2}{\alpha^2}\right| \geq \frac{\Delta^2}{\alpha^2}\gamma_n\right).$$

Hence

$$\mathbf{P}(\mathcal{B}^{\mathsf{c}}) \leq \mathbf{P}\left(\left|\sum_{j=1}^p \varepsilon_j^2 - p\right| \geq \frac{\Delta^2}{\alpha^2}\gamma_n - \left|p - \frac{\Delta^2}{\alpha^2}\right|\right).$$

Using the definition of α^2 we get

$$\mathbf{P}(\mathcal{B}^{\mathsf{c}}) \leq \mathbf{P}\left(\left|\sum_{j=1}^p \varepsilon_j^2 - p\right| \geq p((1-\nu_n)\gamma_n - \nu_n)\right) \leq 2e^{-c\log^3(n)\gamma_n^2}, \tag{6.28}$$

for some $c > 0$ whenever $4\nu_n \leq \gamma_n \leq 1$. Using the inequality $\nu_n^2 \leq 1/\log n$, and choosing $\beta_n^2 = 1/\log n$, $\gamma_n^2 = 16/\log n$ and $\zeta_n^2 = 1/\log n$, we get the desired result by combining (6.27) and (6.28).

Proof of Theorem 6.3.1

We begin by writing that

$$\frac{1}{n}Y^\top Y = \frac{\|\boldsymbol{\theta}\|^2}{n}\eta\eta^\top + Z_1,$$

where

$$Z_1 = \frac{1}{n}\eta\boldsymbol{\theta}^\top W + \frac{1}{n}W^\top\boldsymbol{\theta}\eta^\top + \frac{1}{n}W^\top W.$$

Next observe that

$$H\left(\frac{1}{n}Y^\top Y\right) = \frac{\|\boldsymbol{\theta}\|^2}{n}\eta\eta^\top + Z_2,$$

where Z_2 is given by

$$Z_2 = H\left(Z_1\right) - \frac{\|\boldsymbol{\theta}\|^2}{n}\mathbf{I}_n.$$

Based on Lemma 6.9.2, we have

$$\|Z_2\|_{op} \leq 4\left\|\frac{1}{n}\eta\boldsymbol{\theta}^\top W\right\|_{op} + 2\left\|\frac{1}{n}W^\top W - \mathbf{E}\left(\frac{1}{n}W^\top W\right)\right\|_{op} + \frac{\|\boldsymbol{\theta}\|^2}{n}. \tag{6.29}$$

Using the Davis-Kahan $\sin\theta$ Theorem cf .Theorem 4.5.5 in Vershynin (2018), we obtain

$$\min_{\nu\in\{-1,1\}}\left\|\hat{v}-\frac{1}{\sqrt{n}}\nu\eta\right\|^2\leq 8\frac{\|Z_2\|_{op}^2}{\|\boldsymbol{\theta}\|^4}. \tag{6.30}$$

Hence, using Lemma 6.9.5, we get

$$\frac{1}{n}r(\eta^0,\eta)\leq 16\frac{\|Z_2\|_{op}^2}{\|\boldsymbol{\theta}\|^4}\leq\frac{512}{\|\boldsymbol{\theta}\|^4}\left(\left\|\frac{1}{n}\eta\boldsymbol{\theta}^\top W\right\|_{op}^2+\left\|\frac{1}{n}W^\top W-\mathbf{E}\left(\frac{1}{n}W^\top W\right)\right\|_{op}^2\right)+\frac{32}{n^2}. \tag{6.31}$$

Since $\mathbf{r}_n\geq C$, for some C large enough. We may assume that $\|\boldsymbol{\theta}\|^2\geq 1$ so that $1+p/n\leq\|\boldsymbol{\theta}\|^2+p/n$. The inequality in expectation is a consequence of Lemma 6.9.3 and Lemma 6.9.4.

For the inequality in probability, we first observe, using (6.31), that

$$\frac{1}{n}|\eta^\top\eta^0|\geq 1-8\frac{\|Z_2\|_{op}^2}{\|\boldsymbol{\theta}\|^4}.$$

Next, and since $\mathbf{r}_n\geq C$ for some C large enough, observe that

$$\mathbf{P}_{(\boldsymbol{\theta},\eta)}\left(\frac{1}{n}|\eta^\top\eta^0|\leq 1-\frac{\log n}{n}-\frac{C}{\mathbf{r}_n^2}\right)\leq A_1+A_2,$$

where

$$A_1=\mathbf{P}_{(\boldsymbol{\theta},\eta)}\left(\left\|\frac{1}{n}\eta\theta^\top W\right\|_{op}\geq\sqrt{\frac{\log n}{2n}}\|\boldsymbol{\theta}\|^2+2\right),$$

and

$$A_2=\mathbf{P}_{(\boldsymbol{\theta},\eta)}\left(\left\|\frac{1}{n}W^\top W-\mathbf{E}\left(\frac{1}{n}W^\top W\right)\right\|_{op}\geq\sqrt{\frac{\log n}{2n}}\|\boldsymbol{\theta}\|^2+C\left(1\vee\sqrt{p/n}\right)\right),$$

Using Lemma 6.9.3 and Lemma 6.9.4, we get

$$\mathbf{P}_{(\boldsymbol{\theta},\eta)}\left(\frac{1}{n}|\eta^\top\eta^0|\leq 1-\frac{\log n}{n}-\frac{C}{\mathbf{r}_n^2}\right)\leq 2e^{-c\sqrt{n\log n}\|\boldsymbol{\theta}\|^2(1\wedge\frac{\sqrt{n\log n}\|\boldsymbol{\theta}\|^2}{p})}\leq e^{-c\sqrt{\log n}\mathbf{r}_n^2},$$

Using the tail Gaussian function, we conclude easily that

$$e^{-c\sqrt{\log n}\mathbf{r}_n^2}=o(\Phi^{\mathsf{c}}(\mathbf{r}_n)).$$

Proof of Theorem 6.4.1

By the definition of $r(\hat{\eta},\eta)$, we may assume w.l.o.g that $\eta^\top\hat{\eta}^0>0$. Define the random events $\mathbf{A}_i$ for $i=1,\dots,n$, $\mathbf{B}$ and $\mathbf{C}$ such that for all $i=1,\dots,n$

$$\mathbf{A}_i=\left\{\left(\frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top\eta\right)\eta_i\geq\|\boldsymbol{\theta}\|^2\left(\frac{8C}{\mathbf{r}_n}+\frac{C'}{\mathbf{r}_n^2}+8c'\sqrt{\frac{\log n}{n}}+\nu_n\right)\right\},$$

$$\mathbf{C}=\left\{\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\mathbf{A}_i}\leq\frac{C'}{4\mathbf{r}_n^2}\right\}$$

and

$$\mathbf{B} = \left\{ \|Z_2\|_{op} \leq c' \sqrt{\frac{\log n}{n}} \|\boldsymbol{\theta}\|^2 + C\left(1 \vee \sqrt{p/n}\right) \right\},$$

where we use the same notation of the previous proof and c' a positive constant that we may choose large enough.

We first prove, by induction, that on the event $\mathbf{B} \cap \mathbf{C}$, we have

$$\frac{1}{n}\eta^\top \hat{\eta}^k \geq 1 - \frac{C'}{\mathbf{r}_n^2} - \nu_n, \quad \forall k = 0, 1, \ldots$$

For $k = 0$, the result is obvious. Let $k \geq 1$. Assume that the result holds for k, and we prove it for $k+1$. Remember that

$$\frac{1}{n}\mathbf{H}(Y^\top Y) = \frac{1}{n}\|\boldsymbol{\theta}\|^2 \eta\eta^\top + Z_2.$$

A simple calculation leads to

$$\frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top \hat{\eta}^k = (Z_2)_i^\top (\hat{\eta}^k - \eta) + \frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top \eta - \|\boldsymbol{\theta}\|^2 \eta_i \frac{n - \eta^\top \hat{\eta}^k}{n}.$$

Hence if $\eta_i = -1$ and if $\mathbf{A}_i$ is true, then using the induction hypothesis we get

$$\frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top \hat{\eta}^k \leq (Z_2)_i^\top (\hat{\eta}^k - \eta) - \|\boldsymbol{\theta}\|^2 \left(\frac{8C}{\mathbf{r}_n} + 8c' \sqrt{\frac{\log n}{n}} \right).$$

Hence when $\eta_i = -1$ we have

$$\mathbf{1}_{\left\{\frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top \hat{\eta}^k \geq 0\right\}} \mathbf{1}_{\mathbf{A}_i} \leq \mathbf{1}_{\left\{(Z_2)_i^\top (\hat{\eta}^k - \eta) \geq \|\boldsymbol{\theta}\|^2 \left(\frac{8C}{\mathbf{r}_n} + 8c'\sqrt{\frac{\log n}{n}}\right)\right\}} \leq \left(\frac{(Z_2)_i^\top (\hat{\eta}^k - \eta)}{\|\boldsymbol{\theta}\|^2 \left(\frac{8C}{\mathbf{r}_n} + 8c' \sqrt{\frac{\log n}{n}} \right)} \right)^2.$$

similarly we get for $\eta_i = 1$ that

$$\mathbf{1}_{\left\{\frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top \hat{\eta}^k \leq 0\right\}} \mathbf{1}_{\mathbf{A}_i} \leq \left(\frac{(Z_2)_i^\top (\hat{\eta}^k - \eta)}{\|\boldsymbol{\theta}\|^2 \left(\frac{8C}{\mathbf{r}_n} + 8c' \sqrt{\frac{\log n}{n}} \right)} \right)^2.$$

It is clear that

$$\frac{1}{2}|\hat{\eta}^{k+1} - \eta| = \sum_{\eta_i = -1} \mathbf{1}_{\left\{\frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top \hat{\eta}^k \geq 0\right\}} + \sum_{\eta_i = 1} \mathbf{1}_{\left\{\frac{1}{n}\mathbf{H}(Y^\top Y)_i^\top \hat{\eta}^k \leq 0\right\}}.$$

Hence we get using the events $\mathbf{A}_i$ for $i = 1, \ldots, n$, that

$$\frac{1}{2n}|\hat{\eta}^{k+1} - \eta| \leq \frac{\|Z_2\|_{op}^2}{\|\boldsymbol{\theta}\|^4 \left(\frac{8C}{\mathbf{r}_n} + 8c' \sqrt{\frac{\log n}{n}} \right)^2} \frac{\|\hat{\eta}^k - \eta\|^2}{n} + \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^c}. \tag{6.32}$$

Using the induction hypothesis and the events $\mathbf{B}$ and $\mathbf{C}$, we get

$$1 - \frac{1}{n}\eta^\top \hat{\eta}^{k+1} \leq 4 \left(\frac{c'\sqrt{\frac{\log n}{n}}\|\boldsymbol{\theta}\|^2 + C\left(1 \vee \sqrt{p/n}\right)}{\|\boldsymbol{\theta}\|^2 \left(\frac{8C}{\mathbf{r}_n} + 8c'\sqrt{\frac{\log n}{n}} \right)} \right)^2 (C'/\mathbf{r}_n^2 + \nu_n) + \frac{C'}{2\mathbf{r}_n^2}.$$

Since $\mathbf{r}_n > C$ for C large enough, then $(1 \vee \sqrt{p/n}) \leq \|\boldsymbol{\theta}\|^2/\mathbf{r}_n$, it comes that

$$\frac{1}{n}\eta^\top \hat{\eta}^{k+1} \geq 1 - \frac{C'}{\mathbf{r}_n^2} - \nu_n.$$

That concludes that on $\mathbf{B} \cap \mathbf{C}$, for all $k = 0, 1, \ldots$ we get

$$\frac{1}{n}\eta^\top \hat{\eta}^{k} \geq 1 - \frac{C'}{\mathbf{r}_n^2} - \nu_n.$$

Hence, and using (6.32), we obtain

$$\frac{1}{n}|\hat{\eta}^{k+1} - \eta|\mathbf{1}_\mathbf{B}\mathbf{1}_\mathbf{C} \leq \frac{1}{4}\frac{1}{n}|\hat{\eta}^{k} - \eta|\mathbf{1}_\mathbf{B}\mathbf{1}_\mathbf{C} + \frac{2}{n}\sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^\mathbf{c}}.$$

As a consequence we find that for $k = 0, 1, \ldots$

$$\frac{1}{n}|\hat{\eta}^{k} - \eta|\mathbf{1}_\mathbf{B}\mathbf{1}_\mathbf{C} \leq 2\left(\frac{1}{4}\right)^k + \frac{8}{3n}\sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^\mathbf{c}}.$$

Observe that for $k \geq \lfloor 3\log n \rfloor$, we have $k \geq 2\frac{\log n}{\log 4}$ and

$$\left(\frac{1}{4}\right)^k \leq \frac{1}{n^2}.$$

Hence for $k \geq \lfloor 3\log n \rfloor$,

$$\frac{1}{n}|\hat{\eta}^{k} - \eta|\mathbf{1}_\mathbf{B}\mathbf{1}_\mathbf{C} \leq \frac{2}{n^2} + \frac{8}{3n}\sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^\mathbf{c}}.$$

Observe that if $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^\mathbf{c}} = 0$ then $\frac{1}{n}|\hat{\eta}^{k} - \eta|\mathbf{1}_\mathbf{B}\mathbf{1}_\mathbf{C} = 0$. Else, $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^\mathbf{c}} \geq \frac{1}{n}$. This leads to

$$\frac{1}{n}|\hat{\eta}^{k} - \eta|\mathbf{1}_\mathbf{B}\mathbf{1}_\mathbf{C} \leq \frac{14}{3n}\sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^\mathbf{c}}.$$

Finally we get for $k \geq \lfloor 3\log n \rfloor$,

$$\frac{1}{n}\mathbf{E}\left(|\hat{\eta}^{k} - \eta|\right) \leq \frac{14}{3n}\sum_{i=1}^n \mathbf{P}\left(\mathbf{A}_i^\mathbf{c}\right) + \mathbf{P}\left(\mathbf{B}^\mathbf{c}\right) + \mathbf{P}\left(\mathbf{C}^\mathbf{c}\right) \leq \left(\frac{14}{3} + \frac{4\mathbf{r}_n^2}{C'}\right)\frac{1}{n}\sum_{i=1}^n \mathbf{P}\left(\mathbf{A}_i^\mathbf{c}\right) + \mathbf{P}\left(\mathbf{B}^\mathbf{c}\right).$$

The term $\mathbf{P}\left(\mathbf{B}^\mathbf{c}\right)$ is upper bounded exactly as in the previous proof and we have

$$\mathbf{P}(\mathbf{B}^\mathbf{c}) = o(\Phi^\mathbf{c}(\mathbf{r}_n)).$$

For the other term observe that

$$\mathbf{P}\left(\mathbf{A}_i^\mathbf{c}\right) = G_\sigma\left(\frac{C''}{\mathbf{r}_n} + \epsilon_n, \|\boldsymbol{\theta}\|^2\right),$$

for some $C'' > 0$ and $\epsilon_n = o(1)$. That concludes the proof.

Proof of Theorem 6.4.2

Combining Theorem 6.3.1 and Theorem 6.4.1, it is enough to prove that

$$\mathbf{r}_n^2 \sup_{\|\boldsymbol{\theta}\| \geq \Delta} G_\sigma \left(\epsilon_n + \frac{C'}{\mathbf{r}_n}, \boldsymbol{\theta} \right) \leq \Phi^{\mathsf{c}}(\left(\mathbf{r}_n \left(1 - \epsilon_n' - \frac{C'' \log \mathbf{r}_n}{\mathbf{r}_n} \right) \right) + \epsilon_n' \Phi^{\mathsf{c}}(\mathbf{r}_n),$$

for some $\epsilon_n' = o(1)$ and $C'' > 0$. Recall that

$$G_\sigma \left(\epsilon_n + \frac{C'}{\mathbf{r}_n}, \boldsymbol{\theta} \right) = \mathbf{P} \left((\boldsymbol{\theta} + \xi_1)^\top \left(\boldsymbol{\theta} + \frac{\xi_2}{\sqrt{n-1}} \right) \leq \left(\epsilon_n + \frac{C'}{\mathbf{r}_n} \right) \|\boldsymbol{\theta}\|^2 \right),$$

where ξ_1, ξ_2 are two independent Gaussian random vector with i.i.d. standard entries and $\boldsymbol{\theta}$ and independent Gaussian prior. Moreover, using independence, we have

$$G_\sigma \left(\epsilon_n + \frac{C'}{\mathbf{r}_n}, \boldsymbol{\theta} \right) = \mathbf{P} \left(\varepsilon \sqrt{\|\boldsymbol{\theta}\|^2 + \frac{\|\xi_2\|^2}{n-1} + \frac{2}{\sqrt{n-1}} \boldsymbol{\theta}^\top \xi_2} \geq \|\boldsymbol{\theta}\|^2 \left(1 - \epsilon_n - \frac{C'}{\mathbf{r}_n} \right) + \frac{1}{\sqrt{n-1}} \boldsymbol{\theta}^\top \xi_2 \right),$$

where ε is a standard Gaussian random variable. Set the random event

$$\mathcal{A} = \left\{ \frac{\|\xi_2\|^2}{n-1} \leq \frac{p}{n-1} + \zeta_n \|\boldsymbol{\theta}\|^2 \right\} \cap \left\{ |\boldsymbol{\theta}^\top \xi_2| \leq \sqrt{n-1} \beta_n \|\boldsymbol{\theta}\|^2 \right\},$$

where ζ_n and β_n are positive sequences. It is easy to check that

$$\mathbf{P}\left(\mathcal{A}^{\mathsf{c}}\right) \leq e^{-c\|\boldsymbol{\theta}\|^4 n^2 \zeta_n^2 / p} + e^{-c\beta_n^2 n \|\boldsymbol{\theta}\|^2} + e^{-c\zeta_n n \|\boldsymbol{\theta}\|^2},$$

for some $c > 0$. Hence using the event $\mathcal{A}$, we get

$$G_\sigma \left(\epsilon_n + \frac{C'}{\mathbf{r}_n}, \boldsymbol{\theta} \right) \leq \mathbf{P} \left(\varepsilon \sqrt{\|\boldsymbol{\theta}\|^2 (1 + \zeta_n + 2\beta_n) + \frac{p}{n-1}} \geq \|\boldsymbol{\theta}\|^2 \left(1 - \epsilon_n - \frac{C'}{\mathbf{r}_n} - \beta_n \right) \right) + \mathbf{P}(\mathcal{A}^{\mathsf{c}}).$$

By choosing $\beta_n = \zeta_n = \sqrt{\frac{\log n}{n}}$, we get that

$$\mathbf{P}(\mathcal{A}^{\mathsf{c}}) \leq e^{-c\sqrt{\log n} \mathbf{r}_n}.$$

The last fact is due to the condition $\mathbf{r}_n \geq C$ for some $C > 0$. Hence

$$\mathbf{P}(\mathcal{A}^{\mathsf{c}}) = o(\Phi^{\mathsf{c}}(\mathbf{r}_n)).$$

Moreover and since ζ_n and β_n are vanishing sequences as $n \to \infty$, we get that

$$\mathbf{P} \left(\varepsilon \sqrt{\|\boldsymbol{\theta}\|^2 (1 + \zeta_n + 2\beta_n) + \frac{p}{n-1}} \geq \|\boldsymbol{\theta}\|^2 \left(1 - \epsilon_n - \frac{C'}{\mathbf{r}_n} - \beta_n \right) \right) = \Phi^{\mathsf{c}}(\left(\frac{\|\boldsymbol{\theta}\|^2}{\sqrt{\|\boldsymbol{\theta}\| + \frac{p}{n}}} \left(1 - \frac{C'}{\mathbf{r}_n} - \epsilon_n' \right) \right)$$

for some $\epsilon_n' = o(1)$. We conclude using the fact that $x \to \frac{x}{\sqrt{x + \frac{p}{n}}}$ is non-decreasing on $\mathbf{R}^+$ and the fact that for $C < x < y$, we have $x^2 \Phi^{\mathsf{c}}(y) \leq c_1 \Phi^{\mathsf{c}}(y - c_2 \log x)$, for some $c_1, c_2 > 0$.

Proof of Proposition 6.4.1

Set n large enough. According to Theorem 6.2.2, we have

$$\Psi_\Delta \geq \frac{1}{2}\Phi^{\mathsf{c}}(2\mathbf{r}_n). \tag{6.33}$$

For the upper bound. If $\mathbf{r}_n$ is larger than $2C$, then using Theorem 6.4.2, we get

$$\Psi_\Delta \leq C'\Phi^{\mathsf{c}}(\left(\frac{\mathbf{r}_n}{4}\right), \tag{6.34}$$

for some $C' > 0$. Observe that for $\mathbf{r}_n \leq 2C$, we have

$$c_1 \leq \Phi^{\mathsf{c}}(\mathbf{r}_n),$$

for some $c_1 > 0$. Hence, for $\mathbf{r}_n \leq 2C$, we get

$$\Psi_\Delta \leq \frac{\Phi^{\mathsf{c}}(\mathbf{r}_n)}{c_1}. \tag{6.35}$$

We conclude combining (6.34), (6.33) and (6.35).

Proof of Theorem 6.7.1

- *Necessary conditions:*

  According to Theorem 6.2.2, we have

  $$\Psi_\Delta \geq (1-\epsilon_n)\Phi^{\mathsf{c}}(\mathbf{r}_n(1+\epsilon_n)),$$

  for some $\epsilon_n = o(1)$. If for some $\epsilon > 0$,

  $$\liminf_{n\to\infty} n\Phi^{\mathsf{c}}(\mathbf{r}_n(1+\epsilon)) > 0,$$

  then using the monotonicity of $\Phi^{\mathsf{c}}($, we conclude that exact recovery is impossible.

  For Almost full recovery, assume that $\Phi^{\mathsf{c}}(\mathbf{r}_n)$ does not converge to 0, and that almost full recovery is possible. Then using continuity and monotonicity of $\Phi^{\mathsf{c}}($, we get that $\mathbf{r}_n(1+\epsilon_n) \to \infty$. Hence $\mathbf{r}_n \to \infty$ and $\Phi^{\mathsf{c}}(\mathbf{r}_n) \to 0$ which is absurd. That proves that the condition $\Phi^{\mathsf{c}}(\mathbf{r}_n) \to 0$ is necessary to achieve almost full recovery.

- *Sufficient conditions:*

  According to Theorem 6.4.2, we have that, under the condition $\mathbf{r}_n > C$ for some $C > 0$, the estimator $\hat{\eta}^k$ defined in the Theorem satisfies

  $$\sup_{(\boldsymbol{\theta},\eta)\in\Omega_\Delta} \frac{1}{n}\mathbf{E}_{(\boldsymbol{\theta},\eta)} r(\eta^k, \eta) \leq C'\Phi^{\mathsf{c}}(\left(\mathbf{r}_n\left(1-\epsilon_n-\frac{C'\log\mathbf{r}_n}{\mathbf{r}_n}\right)\right),$$

  for some sequence ϵ_n such that $\epsilon_n = o(1)$. If $\Phi^{\mathsf{c}}(\mathbf{r}_n) \to 0$, then $\mathbf{r}_n \to \infty$. Hence for any $\epsilon > 0$, $\mathbf{r}_n(1-\epsilon) \to \infty$. It follows that $\mathbf{r}_n\left(1-\epsilon_n-\frac{C'\log\mathbf{r}_n}{\mathbf{r}_n}\right) \to \infty$. We

conclude that almost full recovery is possible under the condition $\Phi^{\mathsf{c}}(\mathbf{r}_n) \to 0$, and $\hat{\eta}^k$ achieves almost full recovery in that case.

For exact recovery, observe that, if

$$n\Phi^{\mathsf{c}}(\mathbf{r}_n(1-\epsilon)) \to 0,$$

for some $\epsilon > 0$, then $\mathbf{r}_n \to \infty$. It follows that for n large enough

$$\mathbf{r}_n \left(1 - \epsilon_n - \frac{C' \log \mathbf{r}_n}{\mathbf{r}_n}\right) \geq \mathbf{r}_n(1-\epsilon).$$

We conclude by taking the limit that $\hat{\eta}^k$ achieves exact recovery in that case, and that exact recovery is possible.

Proof of Theorem 6.7.2 and 6.5.1

By inverting the function $x \to \frac{x}{\sqrt{x+\frac{p}{n}}}$, we observe that for any $A > 0$,

$$\mathbf{r}_n^2 \geq A \quad \Leftrightarrow \quad \Delta_n^2 \geq A\frac{1+\sqrt{1+\frac{4p}{nA}}}{2}.$$

Using Theorem 6.7.1 and the Gaussian tail function, we get immediately the results for both almost full recovery and exact recovery.

6.9 Appendix: Technical lemmas

Lemma 6.9.1. *Let A be a matrix in $\mathbf{R}^{n\times n}$. Then*

$$\|H(A)\|_{op} \le 2\|A\|_{op}.$$

Proof. From the linearity of H, we have that

$$\|H(A)\|_{op} \le \|A\|_{op} + \|\mathrm{diag}(A)\|_{op},$$

where

$$\|\mathrm{diag}(A)\|_{op} = \max_i |A_{ii}| \le \|A\|_{op}.$$

□

Lemma 6.9.2. *For any random matrix W with independent columns, we have*

$$\|H(W^\top W)\|_{op} \le 2\left\|W^\top W - \mathbf{E}\left(W^\top W\right)\right\|_{op}.$$

Proof. Since $\mathbf{E}\left(W^\top W\right)$ is a diagonal matrix, it follows that

$$H(W^\top W) = H\left(W^\top W - \mathbf{E}\left(W^\top W\right)\right).$$

The result follows from Lemma 6.9.1. □

Lemma 6.9.3. *Let $u \in \mathbf{S}^{p-1}$ and $v \in \mathbf{S}^{n-1}$, and $W \in \mathbf{R}^{p\times n}$ a matrix with i.i.d. centered Gaussian entries of variance at most σ^2. Then, for some $c, C > 0$*

$$\forall t \ge 2\sigma, \quad \mathbf{P}\left(\left\|\frac{1}{\sqrt{n}} W^\top u v^\top\right\|_{op} \ge t\right) \le e^{-cnt/\sigma},$$

and

$$\mathbf{E}\left(\left\|\frac{1}{\sqrt{n}} W^\top u v^\top\right\|_{op}^2\right) \le C\sigma^2.$$

Proof. We can easily check that

$$\left\|\frac{1}{\sqrt{n}} W^\top u v^\top\right\|_{op} \le \frac{1}{\sqrt{n}}\|W^\top u\|_2.$$

Since $\|u\|_2 = 1$, we have that $W^\top u$ is Gaussian with mean 0 and covariance matrix $\sigma^2\mathbf{I}_n$. We conclude using a tail inequality for quadratic forms of sub-Gaussian random variables using the fact that $t \ge 2\sigma$, see, e.g., Hsu et al. (2012). The inequality in expectation is immediate by integration of the tail function. □

Lemma 6.9.4. *Let $W \in \mathbf{R}^{p\times n}$ be a matrix with i.i.d. centered Gaussian entries of variance at most σ^2. For some $c, C, C' > 0$ we have*

$$\forall t \ge C\sigma^2\left(1 \vee \sqrt{\frac{p}{n}}\right), \quad \mathbf{P}\left(\frac{1}{n}\|H(W^\top W)\|_{op} \ge t\right) \le e^{-cnt/\sigma^2\left(1\wedge\frac{tn}{p\sigma^2}\right)},$$

and

$$\mathbf{E}\left(\frac{1}{n}\|H(W^\top W)\|_{op}^2\right) \le C'\sigma^4(1 + p/n).$$

Proof. Using Lemma 6.9.2, we get

$$\mathbf{P}\left(\frac{1}{n}\|H(W^\top W)\|_{op} \geq t\right) \leq \mathbf{P}\left(\frac{1}{n}\|W^\top W - \mathbf{E}(W^\top W)\|_{op} \geq t/2\right).$$

Now based on Theorem 4.6.1 in Vershynin (2018), we get moreover that

$$\mathbf{P}\left(\frac{1}{n}\|H(W^\top W)\|_{op} \geq t\sigma^2\right) \leq 9^n 2e^{-cnt(1\wedge tn/p)},$$

for some $c > 0$. For $t \geq C(1 \vee \sqrt{p/n})\sigma^2$ with C large enough, we get $ct(1 \wedge tn/p\sigma^2) \geq 4\sigma^2 \log 9$, hence

$$\mathbf{P}\left(\frac{1}{n}\|H(W^\top W)\|_{op} \geq t\right) \leq e^{-c'nt/\sigma^2(1\wedge tn/p\sigma^2)},$$

for some $c' > 0$. The result in expectation is immediate by integration. □

Lemma 6.9.5. *For any $x \in \{-1, 1\}^n$ and $y \in \mathbf{R}^n$, we have*

$$\frac{1}{n}\,|x - sign(y)| \leq 2\left\|\frac{x}{\sqrt{n}} - y\right\|^2.$$

Proof. It is enough to observe that if $x_i \in \{-1, 1\}$, then

$$|x_i - \mathrm{sign}(y_i)| = 2\mathbf{1}(x_i \neq \mathrm{sign}(y_i)) \leq 2|x_i - \sqrt{n}y_i|^2.$$

□

Part III

Adaptive robust estimation

Chapter 7

Adaptive robust estimation in sparse vector model

For the sparse vector model, we consider estimation of the target vector, of its ℓ_2-norm and of the noise variance. We construct adaptive estimators and establish the optimal rates of adaptive estimation when adaptation is considered with respect to the triplet "noise level – noise distribution – sparsity". We consider classes of noise distributions with polynomially and exponentially decreasing tails as well as the case of Gaussian noise. The obtained rates turn out to be different from the minimax non-adaptive rates when the triplet is known. A crucial issue is the ignorance of the noise variance. Moreover, knowing or not knowing the noise distribution can also influence the rate. For example, the rates of estimation of the noise variance can differ depending on whether the noise is Gaussian or sub-Gaussian without a precise knowledge of the distribution. Estimation of noise variance in our setting can be viewed as an adaptive variant of robust estimation of scale in the contamination model, where instead of fixing the "nominal" distribution in advance we assume that it belongs to some class of distributions.

Based on Comminges et al. (2018): Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. (2018). Adaptive robust estimation in sparse vector model. *arXiv preprint arXiv:1802.04230v3.*

7.1 Introduction

This paper considers estimation of the unknown sparse vector, of its ℓ_2-norm and of the noise level in the sparse sequence model. The focus is on construction of estimators that are optimally adaptive in a minimax sense with respect to the noise level, to the form of the noise distribution, and to the sparsity.

We consider the model defined as follows. Let the signal $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ be observed with noise of unknown magnitude $\sigma > 0$:

$$Y_i = \theta_i + \sigma\xi_i, \quad i = 1, \dots, d. \tag{7.1}$$

The noise random variables $\xi_1, \dots, \xi_d$ are assumed to be i.i.d. and we denote by P_ξ the unknown distribution of ξ_1. We assume throughout that the noise is zero-mean, $\mathbf{E}(\xi_1) = 0$, and that $\mathbf{E}(\xi_1^2) = 1$, since σ needs to be identifiable. We denote by $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}$ the distribution of $\boldsymbol{Y} = (Y_1, \dots, Y_d)$ when the signal is $\boldsymbol{\theta}$, the noise level is σ and the

distribution of the noise variables is P_ξ. We also denote by $\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}$ the expectation with respect to $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}$.

We assume that the signal $\boldsymbol{\theta}$ is s-sparse, *i.e.,*

$$\|\boldsymbol{\theta}\|_0 = \sum_{i=1}^{d} \mathbb{1}_{\theta_i \neq 0} \leq s,$$

where $s \in \{1, \dots, d\}$ is an integer. Set $\Theta_s = \{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_0 \leq s\}$. We consider the problems of estimating $\boldsymbol{\theta}$ under the ℓ_2 loss, estimating the variance σ^2, and estimating the ℓ_2-norm

$$\|\boldsymbol{\theta}\|_2 = \Big(\sum_{i=1}^{d} \theta_i^2\Big)^{1/2}.$$

The classical Gaussian sequence model corresponds to the case where the noise ξ_i is standard Gaussian ($P_\xi = \mathcal{N}(0,1)$) and the noise level σ is known. Then, the optimal rate of estimation of $\boldsymbol{\theta}$ under the ℓ_2 loss in a minimax sense on the class Θ_s is $\sqrt{s\log(ed/s)}$ and it is attained by thresholding estimators Donoho et al. (1992). Also, for the Gaussian sequence model with known σ, minimax optimal estimator of the norm $\|\boldsymbol{\theta}\|_2$ as well as the corresponding minimax rate are available from Collier et al. (2017) (see Table 1).

In this chapter, we study estimation of the three objects $\boldsymbol{\theta}$, $\|\boldsymbol{\theta}\|_2$, and σ^2 in the following two settings.

(a) *The distribution of ξ_i and the noise level σ are both unknown.* This is the main setting of our interest. For the unknown distribution of ξ_i, we consider two types of assumptions. Either P_ξ belongs to a class $\mathcal{G}_{a,\tau}$, *i.e.,* for some $a, \tau > 0$,

$$P_\xi \in \mathcal{G}_{a,\tau} \quad \text{iff} \quad \mathbf{E}(\xi_1) = 0,\ \mathbf{E}(\xi_1^2) = 1 \text{ and } \quad \forall t \geq 2,\ \mathbf{P}\big(|\xi_1| > t\big) \leq 2e^{-(t/\tau)^a}, \tag{7.2}$$

which includes for example sub-Gaussian distributions ($a = 2$), or to a class of distributions with polynomially decaying tails $\mathcal{P}_{a,\tau}$, *i.e.,* for some $\tau > 0$ and $a \geq 2$,

$$P_\xi \in \mathcal{P}_{a,\tau} \quad \text{iff} \quad \mathbf{E}(\xi_1) = 0,\ \mathbf{E}(\xi_1^2) = 1 \text{ and } \quad \forall t \geq 2,\ \mathbf{P}\big(|\xi_1| > t\big) \leq \Big(\frac{\tau}{t}\Big)^a. \tag{7.3}$$

We propose estimators of $\boldsymbol{\theta}$, $\|\boldsymbol{\theta}\|_2$, and σ^2 that are optimal in non-asymptotic minimax sense on these classes of distributions and the sparsity class Θ_s. We establish the corresponding non-asymptotic minimax rates. They are given in the second and third columns of Table 1. We also provide the minimax optimal estimators.

(b) *Gaussian noise ξ_i and unknown σ.* The results on the non-asymptotic minimax rates are summarized in the first column of Table 1. Notice an interesting effect – the rates of estimation of σ^2 and of the norm $\|\boldsymbol{\theta}\|_2$ when the noise is Gaussian are faster than the optimal rates when the noise is sub-Gaussian. This can be seen by comparing the first column of Table 1 with the particular case $a = 2$ of the second column corresponding to sub-Gaussian noise.

Some comments about Table 1 and additional details are in order.

	Gaussian noise model	Noise in class $\mathcal{G}_{a,\tau}$,	Noise in class $\mathcal{P}_{a,\tau}$,
$\boldsymbol{\theta}$	$\sqrt{s\log(ed/s)}$ known σ Donoho et al. (1992) unknown σ Verzelen (2012)	$\sqrt{s}\log^{\frac{1}{a}}(ed/s)$ unknown σ	$\sqrt{s}(d/s)^{\frac{1}{a}}$ unknown σ
$\|\boldsymbol{\theta}\|_2$	$\sqrt{s\log(1+\frac{\sqrt{d}}{s})} \wedge d^{1/4}$ known σ Collier et al. (2017) $\sqrt{s\log(1+\frac{\sqrt{d}}{s})} \vee \sqrt{\frac{s}{1+\log_+(s^2/d)}}$ unknown σ	$\sqrt{s}\log^{\frac{1}{a}}(ed/s) \wedge d^{1/4}$ known σ $\sqrt{s}\log^{\frac{1}{a}}(ed/s)$ unknown σ	$\sqrt{s}(d/s)^{\frac{1}{a}} \wedge d^{1/4}$ known σ $\sqrt{s}(d/s)^{\frac{1}{a}}$ unknown σ
σ^2	$\frac{1}{\sqrt{d}} \vee \frac{s}{d(1+\log_+(s^2/d))}$	$\frac{1}{\sqrt{d}} \vee \frac{s}{d}\log^{\frac{2}{a}}\left(\frac{ed}{s}\right)$	$\frac{1}{\sqrt{d}} \vee \left(\frac{s}{d}\right)^{1-\frac{2}{a}}$

Table 7.1: Optimal rates of convergence.

- The difference between the minimax rates for estimation of $\boldsymbol{\theta}$ and estimation of the ℓ_2-norm $\|\boldsymbol{\theta}\|_2$ turns out to be specific for the pure Gaussian noise model. It disappears for the classes $\mathcal{G}_{a,\tau}$ and $\mathcal{P}_{a,\tau}$. This is somewhat unexpected since $\mathcal{G}_{2,\tau}$ is the class of sub-Gaussian distributions, and it turns out that $\|\boldsymbol{\theta}\|_2$ is estimated optimally at different rates for sub-Gaussian and pure Gaussian noise. Another conclusion is that if the noise is not Gaussian and σ is unknown, the minimax rate for $\|\boldsymbol{\theta}\|_2$ does not have an elbow between the "dense" ($s > \sqrt{d}$) and the "sparse" ($s \le \sqrt{d}$) zones.

- For the problem of estimation of variance σ^2 with *known* distribution of the noise P_ξ, we consider a more general setting than (b) mentioned above. We show that when the noise distribution is exactly known (and satisfies a rather general assumption, not necessarily Gaussian - can have polynomial tails), then the rate of estimation of σ^2 can be as fast as $\max\left(\frac{1}{\sqrt{d}}, \frac{s}{d}\right)$, which is faster than the optimal rate $\max\left(\frac{1}{\sqrt{d}}, \frac{s}{d}\log\left(\frac{ed}{s}\right)\right)$ for the class of sub-Gaussian noise. In other words, the phenomenon of improved rate is not due to the Gaussian character of the noise but rather to the fact that the noise distribution is known.

- Our findings show that there is a dramatic difference between the behavior of optimal estimators of $\boldsymbol{\theta}$ in the sparse sequence model and in the sparse linear regression model with "well spread" regressors. It is known from Gautier and Tsybakov (2013); Belloni et al. (2014) that in sparse linear regression with "well spread" regressors (that is, having positive variance), the rates of estimating $\boldsymbol{\theta}$ are the same for the noise with sub-Gaussian and polynomial tails. We show that the situation is quite different in the sparse sequence model, where the optimal rates are much slower and depend on the polynomial index of the noise.

- The rates shown in Table 1 for the classes $\mathcal{G}_{a,\tau}$ and $\mathcal{P}_{a,\tau}$ are achieved on estimators that are adaptive to the sparsity index s. Thus, knowing or not knowing s does not influence the optimal rates of estimation when the distribution of ξ and the noise level are unknown.

We conclude this section by a discussion of related work. Chen et al. (2018) explore the problem of robust estimation of variance and of covariance matrix under Hubers's contamination model. As explained in Section 7.4 below, this problem has similarities with estimation of noise level in our setting. The main difference is that instead of fixing in advance the Gaussian nominal distribution of the contamination model we assume that it belongs to a class of distributions, such as (7.2) or (7.3). Therefore, the corresponding results in Section 7.4 can be viewed as results on robust estimation of scale where, in contrast to the classical setting, we are interested in adaptation to the unknown nominal law. Another aspect of robust estimation of scale is analyzed by Wei and Minsker (2017) who consider classes of distributions similar to $\mathcal{P}_{a,\tau}$ rather than the contamination model. The main aim in Wei and Minsker (2017) is to construct estimators having sub-Gaussian deviations under weak moment assumptions. Our setting is different in that we consider the sparsity class Θ_s of vectors $\boldsymbol{\theta}$ and the rates that we obtain depend on s. Estimation of variance in sparse linear model is discussed in Sun and Zhang (2012) where some upper bounds for the rates are given. We also mention the recent paper Golubev and Krymova (2017) that deals with estimation of variance in linear regression in a framework that does not involve sparsity, as well as the work on estimation of signal-to-noise ratio functionals in settings involving sparsity Verzelen and Gassiat (2018); Guo et al. (2018) and not involving sparsity Janson et al. (2017). Papers Collier et al. (2018); Carpentier and Verzelen (2019) discuss estimation of other functionals than the ℓ_2-norm $\|\boldsymbol{\theta}\|_2$ in the sparse vector model when the noise is Gaussian with unknown variance.

Notation. For $x > 0$, let $\lfloor x \rfloor$ denote the maximal integer smaller than x. For a finite set A, we denote by $|A|$ its cardinality. Let $\inf_{\hat{T}}$ denote the infimum over all estimators. The notation C, C',c, c' will be used for positive constants that can depend only a and τ and can vary from line to line.

7.2 Estimation of the sparse vector

In this section, we study the problem of estimating a sparse vector $\boldsymbol{\theta}$ in ℓ_2-norm when the variance of noise σ and the distribution of ξ_i are both unknown. We only assume that the noise distribution belongs a given class, which can be either a class of distributions with polynomial tails $\mathcal{P}_{a,\tau}$, or a class $\mathcal{G}_{a,\tau}$ with exponential decay of the tails.

First, we introduce a preliminary estimator $\tilde{\sigma}^2$ of σ^2 that will be used to define an estimator of $\boldsymbol{\theta}$. Let $\gamma \in (0, 1/2]$ be a constant that will be chosen small enough and depending only on a and τ. Divide $\{1, \dots, d\}$ into $m = \lfloor \gamma d \rfloor$ disjoint subsets $B_1, \dots, B_m$, each of cardinality $|B_i| \geq k := \lfloor d/m \rfloor \geq 1/\gamma - 1$. Consider the median-of-means estimator

$$\tilde{\sigma}^2 = \mathsf{med}(\bar{\sigma}_1^2, \dots, \bar{\sigma}_m^2), \text{ where } \bar{\sigma}_i^2 = \frac{1}{|B_i|} \sum_{j \in B_i} Y_j^2, \quad i = 1, \dots, m. \tag{7.4}$$

Here, $\mathsf{med}(\bar\sigma_1^2,\dots,\bar\sigma_m^2)$ denotes the median of $\bar\sigma_1^2,\dots,\bar\sigma_m^2$. The next proposition shows that the estimator $\tilde\sigma^2$ recovers σ^2 to within a constant factor.

Proposition 7.2.1. *Let $\tau>0, a>2$. There exist constants $\gamma\in(0,1/2]$, $c>0$ and $C>0$ depending only on a and τ such that for any integers s and d satisfying $1\le s<\lfloor\gamma d\rfloor/4$ we have*

$$\inf_{P_\xi\in\mathcal{P}_{a,\tau}}\inf_{\sigma>0}\inf_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(1/2\le\frac{\tilde\sigma^2}{\sigma^2}\le 3/2\Big)\ge 1-\exp(-cd),$$

$$\sup_{P_\xi\in\mathcal{P}_{a,\tau}}\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left|\tilde\sigma^2-\sigma^2\right|\le C\sigma^2,$$

and for $a>4$,

$$\sup_{P_\xi\in\mathcal{P}_{a,\tau}}\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\tilde\sigma^2-\sigma^2\right)^2\le C\sigma^4.$$

Note that the result of Proposition 7.2.1 also holds for the class $\mathcal{G}_{a,\tau}$ for all $a>0$ and $\tau>0$. Indeed, $\mathcal{G}_{a,\tau}\subset\mathcal{P}_{a,\tau}$ for all $a>2$ and $\tau>0$, while for any $0<a\le 2$ and $\tau>0$, there exist $a'>4$ and $\tau'>0$ such that $\mathcal{G}_{a,\tau}\subset\mathcal{P}_{a',\tau'}$.

We further note that assuming $s<cd$ for some $0<c<1$ is natural in the context of variance estimation since σ is not identifiable when $s=d$. In what follows, all upper bounds on the risks of estimators will be obtained under this assumption.

Consider now an estimator $\hat{\boldsymbol{\theta}}$ defined as follows:

$$\hat{\boldsymbol{\theta}}\in\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\Big(\sum_{i=1}^d(Y_i-\theta_i)^2+\tilde\sigma\|\boldsymbol{\theta}\|_*\Big). \tag{7.5}$$

Here, $\|\cdot\|_*$ denotes the sorted ℓ_1-norm:

$$\|\boldsymbol{\theta}\|_*=\sum_{i=1}^d\lambda_i|\theta|_{(d-i+1)}, \tag{7.6}$$

where $|\theta|_{(1)}\le\dots\le|\theta|_{(d)}$ are the order statistics of $|\theta_1|,\dots,|\theta_d|$, and $\lambda_1\ge\dots\ge\lambda_p>0$ are tuning parameters.

Set

$$\phi^*_{\mathsf{exp}}(s,d)=\sqrt{s}\log^{1/a}(ed/s),\qquad \phi^*_{\mathsf{pol}}(s,d)=\sqrt{s}(d/s)^{1/a}. \tag{7.7}$$

The next theorem shows that $\hat{\boldsymbol{\theta}}$ estimates $\boldsymbol{\theta}$ with the rates $\phi^*_{\mathsf{exp}}(s,d)$ and $\phi^*_{\mathsf{pol}}(s,d)$ when the noise distribution belongs to the class $\mathcal{G}_{a,\tau}$ and class $\mathcal{P}_{a,\tau}$, respectively.

Theorem 7.2.1. *Let s and d be integers satisfying $1\le s<\lfloor\gamma d\rfloor/4$ where $\gamma\in(0,1/2]$ is the tuning parameter in the definition of $\tilde\sigma^2$. Then for the estimator $\hat{\boldsymbol{\theta}}$ defined by (7.5) the following holds.*

1. *Let $\tau>0$, $a>0$. There exist constants $c,C>0$ and $\gamma\in(0,1/2]$ depending only on (a,τ) such that if $\lambda_j=c\log^{1/a}(ed/j), j=1,\dots,d$, we have*

$$\sup_{P_\xi\in\mathcal{G}_{a,\tau}}\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2^2\right)\le C\sigma^2\left(\phi^*_{\mathsf{exp}}(s,d)\right)^2.$$

2. *Let $\tau > 0, a > 2$. There exist constants $c, C > 0$ and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\lambda_j = c(d/j)^{1/a}, j = 1, \dots, d$, we have*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right) \le C\sigma^2 \left(\phi^*_{\mathsf{pol}}(s, d) \right)^2 .$$

Furthermore, it follows from the lower bound of Theorem 7.3.1 in Section 7.3 that the rates $\phi^*_{\mathsf{exp}}(s, d)$ and $\phi^*_{\mathsf{pol}}(s, d)$ cannot be improved in a minimax sense. Thus, the estimator $\hat{\boldsymbol{\theta}}$ defined in (7.5) achieves the optimal rates in a minimax sense.

From Theorem 7.2.1, we can conclude that the optimal rate ϕ^*_{pol} under polynomially decaying noise is very different from the optimal rate ϕ^*_{exp} under exponential tails, in particular, from the rate under the sub-Gaussian noise. At first sight, this phenomenon seems to contradict some results in the literature on sparse regression model. Indeed, Gautier and Tsybakov (2013) consider sparse linear regression with unknown noise level σ and show that the Self-Tuned Dantzig estimator can achieve the same rate as in the case of Gaussian noise (up to a logarithmic factor) under the assumption that the noise is symmetric and has only a bounded moment of order $a > 2$. Belloni et al. (2014) show for the same model that a square-root Lasso estimator achieves analogous behavior under the assumption that the noise has a bounded moment of order $a > 2$. However, a crucial condition in Belloni et al. (2014) is that the design is "well spread", that is all components of the design vectors are random with positive variance. The same type of condition is needed in Gautier and Tsybakov (2013) to obtain a sub-Gaussian rate. This condition of "well spreadness" is not satisfied in the sparse sequence model that we are considering here. In this model viewed as a special case of linear regression, the design is deterministic, with only one non-zero component. We see that such a degenerate design turns out to be the least favorable from the point of view of the convergence rate, while the "well spread" design is the best one. An interesting general conclusion of comparing our findings to Gautier and Tsybakov (2013) and Belloni et al. (2014) is that the optimal rate of convergence of estimators under sparsity when the noise level is unknown depends dramatically on the properties of the design. There is a whole spectrum of possibilities between the degenerate and "well spread" designs where a variety of new rates can arise depending on the properties of the design. Studying them remains an open problem.

7.3 Estimation of the norm

In this section, we consider the problem of estimation of the ℓ_2-norm of a sparse vector when the variance of the noise and the form of its distribution are both unknown. We show that the rates $\phi^*_{\mathsf{exp}}(s, d)$ and $\phi^*_{\mathsf{pol}}(s, d)$ are optimal in a minimax sense on the classes $\mathcal{G}_{a,\tau}$ and $\mathcal{P}_{a,\tau}$, respectively. We first provide a lower bound on the risks of any estimators of the ℓ_2-norm when the noise level σ is unknown and the unknown noise distribution P_ξ belongs either to $\mathcal{G}_{a,\tau}$ or $\mathcal{P}_{a,\tau}$. We denote by $\mathcal{L}$ the set of all monotone non-decreasing functions $\ell : [0, \infty) \to [0, \infty)$ such that $\ell(0) = 0$ and $\ell \not\equiv 0$.

Theorem 7.3.1. *Let s, d be integers satisfying $1 \le s \le d$. Let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then, for any $a > 0, \tau > 0$,*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\, \ell\Big(c(\phi^*_{\mathsf{exp}}(s, d))^{-1} \Big| \frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma} \Big| \Big) \ge c', \tag{7.8}$$

and, for any $a \geq 2, \tau > 0$,

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\Big(\bar{c}(\phi^*_{\mathsf{pol}}(s,d))^{-1}\Big|\frac{\hat{T}-\|\boldsymbol{\theta}\|_2}{\sigma}\Big|\Big) \geq \bar{c}'. \tag{7.9}$$

Here, $\inf_{\hat{T}}$ *denotes the infimum over all estimators, and* $c, \bar{c} > 0$, $c', \bar{c}' > 0$ *are constants that can depend only on* $\ell(\cdot)$, τ *and* a.

The lower bound (7.9) implies that the rate of estimation of the ℓ_2-norm of a sparse vector deteriorates dramatically if the bounded moment assumption is imposed on the noise instead, for example, of the sub-Gaussian assumption.

Note also that (7.8) and (7.9) immediately imply lower bounds with the same rates ϕ^*_{exp} and ϕ^*_{pol} for the estimation of the s-sparse vector $\boldsymbol{\theta}$ under the ℓ_2-norm.

Given the upper bounds of Theorem 7.2.1, the lower bounds (7.8) and (7.9) are tight for the quadratic loss, and are achieved by the following plug-in estimator independent of s or σ:

$$\hat{N} = \|\hat{\boldsymbol{\theta}}\|_2 \tag{7.10}$$

where $\hat{\boldsymbol{\theta}}$ is defined in (7.5).

In conclusion, when both P_ξ and σ are unknown the rates ϕ^*_{exp} and ϕ^*_{pol} defined in (7.7) are minimax optimal both for estimation of $\boldsymbol{\theta}$ and of the the norm $\|\boldsymbol{\theta}\|_2$.

We now compare these results with the findings in Collier et al. (2017) regarding the (nonadaptive) estimation of $\|\boldsymbol{\theta}\|_2$ when ξ_i have the standard Gaussian distribution ($P_\xi = \mathcal{N}(0,1)$) and σ is known. It is shown in Collier et al. (2017) that in this case the optimal rate of estimation of $\|\boldsymbol{\theta}\|_2$ has the form

$$\phi_{\mathcal{N}(0,1)}(s,d) = \min\left\{\sqrt{s\log(1+\sqrt{d}/s)}, d^{1/4}\right\}.$$

Namely, the following proposition holds.

Proposition 7.3.1 (Gaussian noise, known σ Collier et al. (2017))**.** *For any* $\sigma > 0$ *and any integers* s, d *satisfying* $1 \leq s \leq d$, *we have*

$$c\sigma^2\phi^2_{\mathcal{N}(0,1)}(s,d) \leq \inf_{\hat{T}} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\big(\hat{T}-\|\boldsymbol{\theta}\|_2\big)^2 \leq C\sigma^2\phi^2_{\mathcal{N}(0,1)}(s,d),$$

where $c > 0$ *and* $C > 0$ *are absolute constants and* $\inf_{\hat{T}}$ *denotes the infimum over all estimators.*

We have seen that, in contrast to this result, in the case of unknown P_ξ and σ the optimal rates (7.7) do not exhibit an elbow at $s = \sqrt{d}$ between the "sparse" and "dense" regimes. Another conclusion is that, in the "dense" zone $s > \sqrt{d}$, adaptation to P_ξ and σ is only possible with a significant deterioration of the rate. On the other hand, for the sub-Gaussian class $\mathcal{G}_{2,\tau}$, in the "sparse" zone $s \leq \sqrt{d}$ the nonadaptive rate $\sqrt{s\log(1+\sqrt{d}/s)}$ differs only slightly from the adaptive sub-Gaussian rate $\sqrt{s\log(ed/s)}$; in fact, this difference in the rate appears only in a vicinity of $s = \sqrt{d}$.

A natural question is whether such a deterioration of the rate is caused by the ignorance of σ or by the ignorance of the distribution of ξ_i within the sub-Gaussian

class $\mathcal{G}_{2,\tau}$. The answer is that both are responsible. It turns out that if only one of the two ingredients (σ or the noise distribution) is unknown, then a rate faster than the adaptive sub-Gaussian rate $\phi^*_{\mathsf{exp}}(s,d) = \sqrt{s\log(ed/s)}$ can be achieved. This is detailed in the next two propositions.

Consider first the case of Gaussian noise and unknown σ. Set

$$\phi^*_{\mathcal{N}(0,1)}(s,d) = \max\left\{\sqrt{s\log(1+\sqrt{d}/s)}, \sqrt{\frac{s}{1+\log_+(s^2/d)}}\right\},$$

where $\log_+(x) = \max(0,\log(x))$ for any $x>0$. We divide the set $\{1,\dots,d\}$ into two disjoint subsets I_1 and I_2 with $\min(|I_1|,|I_2|) \geq \lfloor d/2\rfloor$. Let $\hat\sigma^2$ be the variance estimator defined by (7.15), cf. Section 7.4 below, and let $\hat\sigma^2_{\mathsf{med},1}, \hat\sigma^2_{\mathsf{med},2}$ be the median estimators (7.12) corresponding to the samples $(Y_i)_{i\in I_1}$ and $(Y_i)_{i\in I_2}$, respectively. Consider the estimator

$$\hat N^* = \begin{cases} \sqrt{\left|\sum_{j=1}^d (Y_j^2\, \mathbb{1}_{\{|Y_j|>\rho_j\}}) - d\alpha\hat\sigma^2\right|} & \text{if} \quad s\leq\sqrt{d}, \\ \sqrt{\left|\sum_{j=1}^d Y_j^2 - d\hat\sigma^2\right|} & \text{if} \quad s>\sqrt{d}, \end{cases} \tag{7.11}$$

where $\rho_j = 2\hat\sigma_{\mathsf{med},1}\sqrt{2\log(1+d/s^2)}$ if $j\in I_2$, $\rho_j = 2\hat\sigma_{\mathsf{med},2}\sqrt{2\log(1+d/s^2)}$ if $j\in I_1$ and $\alpha = \mathbf{E}\left(\xi_1^2\, \mathbb{1}_{\{|\xi_1|>2\sqrt{2\log(1+d/s^2)}\}}\right)$. Note that Y_j is independent of ρ_j for every j. Note also that the estimator $\hat N^*$ depends on the preliminary estimator $\tilde\sigma^2$ since $\hat\sigma>0$ defined in (7.15) depends on it.

Proposition 7.3.2 (Gaussian noise, unknown σ)**.** *The following two properties hold.*

(i) Let s and d be integers satisfying $1\leq s<\lfloor\gamma d\rfloor/4$, where $\gamma\in(0,1/2]$ is the tuning parameter in the definition of $\tilde\sigma^2$. There exist absolute constants $C>0$ and $\gamma\in(0,1/2]$ such that

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\leq s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\left(\hat N^* - \|\boldsymbol{\theta}\|_2\right)^2 \leq C\sigma^2\left(\phi^*_{\mathcal{N}(0,1)}(s,d)\right)^2.$$

(ii) Let s and d be integers satisfying $1\leq s\leq d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,

$$\inf_{\hat T}\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\leq s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\,\ell\left(c(\phi^*_{\mathcal{N}(0,1)}(s,d))^{-1}\left|\frac{\hat T - \|\boldsymbol{\theta}\|_2}{\sigma}\right|\right) \geq c',$$

where $\inf_{\hat T}$ denotes the infimum over all estimators, and $c>0$, $c'>0$ are constants that can depend only on $\ell(\cdot)$.

Proposition 7.3.2 establishes the minimax optimality of the rate $\phi^*_{\mathcal{N}(0,1)}(s,d)$. It also shows that if σ is unknown, the knowledge of the Gaussian character of the noise leads to an improvement of the rate compared to the adaptive sub-Gaussian rate $\sqrt{s\log(ed/s)}$. However, the improvement is only in a logarithmic factor.

Consider now the case of unknown noise distribution in $\mathcal{G}_{a,\tau}$ and known σ. We show in the next proposition that in this case the minimax rate is of the form

$$\phi^\circ_{\mathsf{exp}}(s,d) = \min\{\sqrt{s}\log^{\frac{1}{a}}(ed/s), d^{1/4}\}$$

and it is achieved by the estimator

$$\hat{N}^{\circ}_{\mathsf{exp}} = \begin{cases} \|\hat{\boldsymbol{\theta}}\|_2 & \text{if} \quad s \leq \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}, \\ \left|\sum_{j=1}^{d} Y_j^2 - d\sigma^2\right|^{1/2} & \text{if} \quad s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}, \end{cases}$$

where $\hat{\boldsymbol{\theta}}$ is defined in (7.5). Note $\phi^{\circ}_{\mathsf{exp}}(s,d)$ can be written equivalently (up to absolute constants) as $\min\{\sqrt{s}\log^{\frac{1}{a}}(ed), d^{1/4}\}$.

Proposition 7.3.3 (Unknown noise in $\mathcal{G}_{a,\tau}$, known σ)**.** *Let $a, \tau > 0$. The following two properties hold.*

(i) Let s and d be integers satisfying $1 \leq s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. There exist constants $c, C > 0$, and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\hat{\boldsymbol{\theta}}$ is the estimator defined in (7.5) with $\lambda_j = c\log^{\frac{1}{a}}(ed/j)$, $j = 1, \ldots, d$, then

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\hat{N}^{\circ}_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2\right)^2 \leq C\sigma^2 \left(\phi^{\circ}_{\mathsf{exp}}(s,d)\right)^2.$$

(ii) Let s and d be integers satisfying $1 \leq s \leq d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then, there exist constants $c > 0$, $c' > 0$ depending only on $\ell(\cdot)$, a and τ such that

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\, \ell\left(c(\phi^{\circ}_{\mathsf{exp}}(s,d))^{-1} \left|\frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma}\right|\right) \geq c',$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators.

Proposition 7.3.3 establishes the minimax optimality of the rate $\phi^{\circ}_{\mathsf{exp}}(s,d)$. It also shows that if the noise distribution is unknown and belongs to $\mathcal{G}_{a,\tau}$, the knowledge of σ leads to an improvement of the rate compared to the case when σ is unknown. In contrast to the case of Proposition 7.3.2 (Gaussian noise), the improvement here is substantial; it results not only in a logarithmic but in a polynomial factor in the dense zone $s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$.

We end this section by considering the case of unknown polynomial noise and known σ. The next proposition shows that in this case the minimax rate, for a given $a > 4$, is of the form

$$\phi^{\circ}_{\mathsf{pol}}(s,d) = \min\{\sqrt{s}(d/s)^{\frac{1}{a}}, d^{1/4}\}$$

and it is achieved by the estimator

$$\hat{N}^{\circ}_{\mathsf{pol}} = \begin{cases} \|\hat{\boldsymbol{\theta}}\|_2 & \text{if} \quad s \leq d^{\frac{1}{2}-\frac{1}{a-2}}, \\ \left|\sum_{j=1}^{d} Y_j^2 - d\sigma^2\right|^{1/2} & \text{if} \quad s > d^{\frac{1}{2}-\frac{1}{a-2}}, \end{cases}$$

where $\hat{\boldsymbol{\theta}}$ is defined in (7.5).

Proposition 7.3.4 (Unknown noise in $\mathcal{P}_{a,\tau}$, known σ)**.** *Let $\tau > 0, a > 4$. The following two properties hold.*

(i) Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. There exist constants $c, C > 0$, and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\hat{\boldsymbol{\theta}}$ is the estimator defined in (7.5) with $\lambda_j = c(d/j)^{\frac{1}{a}}$, $j = 1, \dots, d$, then

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\hat{N}^\circ_{\mathsf{pol}} - \|\boldsymbol{\theta}\|_2 \right)^2 \le C\sigma^2 \left(\phi^\circ_{\mathsf{pol}}(s, d) \right)^2 .$$

(ii) Let s and d be integers satisfying $1 \le s \le d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then, there exist constants $c > 0$, $c' > 0$ depending only on $\ell(\cdot)$, a and τ such that

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \, \ell\left(c(\phi^\circ_{\mathsf{pol}}(s, d))^{-1} \left| \frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma} \right| \right) \ge c',$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators.

Note that here, similarly to Proposition 7.3.3, the improvement over the case of unknown σ is in a polynomial factor in the dense zone $s > d^{\frac{1}{2} - \frac{1}{a-2}}$.

7.4 Estimating the variance of the noise

Estimating σ^2 when the distribution P_ξ is known

In the sparse setting when $\|\boldsymbol{\theta}\|_0$ is small, estimation of the noise level can be viewed as a problem of robust estimation of scale. Indeed, our aim is to recover the second moment of $\sigma\xi_1$ but the sample second moment cannot be used as an estimator because of the presence of a small number of outliers $\theta_i \neq 0$. Thus, the models in robustness and sparsity problems are quite similar but the questions of interest are different. When robust estimation of σ^2 is considered, the object of interest is the pure noise component of the sparsity model while the non-zero components θ_i that are of major interest in the sparsity model play a role of nuisance.

In the context of robustness, it is known that the estimator based on sample median can be successfully applied. Recall that, when $\boldsymbol{\theta} = 0$, the median M-estimator of scale (*cf.* Huber (2011)) is defined as

$$\hat{\sigma}^2_{\mathsf{med}} = \frac{\hat{M}}{\beta} \tag{7.12}$$

where $\hat{M}$ is the sample median of $(Y_1^2, \dots, Y_d^2)$, that is

$$\hat{M} \in \arg\min_{x > 0} \left| F_d(x) - 1/2 \right|,$$

and β is the median of the distribution of ξ_1^2. Here, F_d denotes the empirical c.d.f. of $(Y_1^2, \dots, Y_d^2)$. When F denotes the c.d.f. of ξ_1^2, it is easy to see that

$$\beta = F^{-1}(1/2). \tag{7.13}$$

The following proposition specifies the rate of convergence of the estimator $\hat{\sigma}^2_{\mathsf{med}}$.

Proposition 7.4.1. *Let ξ_1^2 have a c.d.f. F with positive density, and let β be given by (7.13). There exist constants $\gamma \in (0, 1/8)$, $c > 0$, $c_* > 0$ and $C > 0$ depending only on F such that for any integers s and d satisfying $1 \le s < \gamma d$ and any $t > 0$ we have*

$$\sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},F,\sigma} \left(\left| \frac{\hat{\sigma}^2_{\mathsf{med}}}{\sigma^2} - 1 \right| \ge c_* \left(\sqrt{\frac{t}{d}} + \frac{s}{d} \right) \right) \le 2(e^{-t} + e^{-cd}),$$

and if $\mathbf{E}|\xi_1|^{2+\epsilon} < \infty$ for some $\epsilon > 0$. Then,

$$\sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},F,\sigma} |\hat{\sigma}^2_{\mathsf{med}} - \sigma^2|}{\sigma^2} \le C \max\left(\frac{1}{\sqrt{d}}, \frac{s}{d} \right).$$

The main message of Proposition 7.4.1 is that the rate of convergence of $\hat{\sigma}^2_{\mathsf{med}}$ in probability and in expectation is as fast as

$$\max\left(\frac{1}{\sqrt{d}}, \frac{s}{d} \right) \tag{7.14}$$

and it does not depend on F when F varies in a large class. The role of Proposition 7.4.1 is to contrast the subsequent results of this section dealing with unknown distribution of noise and providing slower rates. It emphasizes the fact that the knowledge of the noise distribution is crucial as it leads to an improvement of the rate of estimating the variance.

However, the rate (7.14) achieved by the median estimator is not necessarily optimal. As shown in the next proposition, in the case of Gaussian noise the optimal rate is even better:

$$\phi_{\mathcal{N}(0,1)}(s, d) = \max\left\{ \frac{1}{\sqrt{d}}, \frac{s}{d(1 + \log_+(s^2/d))} \right\}.$$

This rate is attained by an estimator that we are going to define now. We use the observation that, in the Gaussian case, the modulus of the empirical characteristic function $\varphi_d(t) = \frac{1}{d}\sum_{i=1}^d e^{itY_j}$ is to within a constant factor of the Gaussian characteristic function $\exp(-\frac{t^2\sigma^2}{2})$ for any t. This suggests the estimator

$$\tilde{v}^2 = -\frac{2\log(|\varphi_d(\hat{t}_1)|)}{\hat{t}_1^2},$$

with a suitable choice of $t = \hat{t}_1$ that we further set as follows:

$$\hat{t}_1 = \frac{1}{\tilde{\sigma}} \sqrt{\log\left(4(es/\sqrt{d} + 1)\right)},$$

where $\tilde{\sigma}$ is the preliminary estimator (7.4) with some tuning parameter $\gamma \in (0, 1/2]$. The final variance estimator is defined as a truncated version of $\tilde{v}^2$:

$$\hat{\sigma}^2 = \begin{cases} \tilde{v}^2 & \text{if } |\varphi_d(\hat{t}_1)| > (es/\sqrt{d} + 1)^{-1}/4, \\ \tilde{\sigma}^2 & \text{otherwise.} \end{cases} \tag{7.15}$$

Proposition 7.4.2 (Gaussian noise)**.** *The following two properties hold.*

(i) Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. There exist absolute constants $C > 0$ and $\gamma \in (0, 1/2]$ such that the estimator $\hat{\sigma}^2$ defined in (7.15) satisfies

$$\sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma} |\hat{\sigma}^2 - \sigma^2|}{\sigma^2} \le C \phi_{\mathcal{N}(0,1)}(s, d).$$

(ii) Let s and d be integers satisfying $1 \le s \le d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,

$$\inf_{\hat{T}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\, \ell\left(c(\phi_{\mathcal{N}(0,1)}(s, d))^{-1} \left| \frac{\hat{T}}{\sigma^2} - 1 \right| \right) \ge c',$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators, and $c > 0$, $c' > 0$ are constants that can depend only on $\ell(\cdot)$.

Estimators of variance or covariance matrix based on the empirical characteristic function have been studied in several papers Butucea et al. (2005); Cai and Jin (2010); Belomestny et al. (2017); Carpentier and Verzelen (2019). The setting in Butucea et al. (2005); Cai and Jin (2010); Belomestny et al. (2017) is different from the ours as those papers deal with the model where the non-zero components of $\boldsymbol{\theta}$ are random with a smooth distribution density. The estimators in Butucea et al. (2005); Cai and Jin (2010) are also quite different. On the other hand, Belomestny et al. (2017); Carpentier and Verzelen (2019) consider estimators close to $\tilde{v}^2$. In particular, Carpentier and Verzelen (2019) uses a similar pilot estimator for testing in the sparse vector model where it is assumed that $\sigma \in [\sigma_-, \sigma_+]$, $0 < \sigma_- < \sigma_+ < \infty$, and the estimator depends on σ_+. Although Carpentier and Verzelen (2019) does not provide explicitly stated result about the rate of this estimator, the proofs in Carpentier and Verzelen (2019) come close to it and we believe that it satisfies an upper bound as in item (i) of Proposition 7.4.2 with $\sup_{\sigma>0}$ replaced by $\sup_{\sigma\in[\sigma_-,\sigma_+]}$.

Distribution-free variance estimators

The main drawback of the estimator $\hat{\sigma}^2_{\mathsf{med}}$ is the dependence on the parameter β. It reflects the fact that the estimator is tailored for a given and known distribution of noise F. Furthermore, as shown below, the rate (7.14) cannot be achieved if it is only known that F belongs to one of the classes of distributions that we consider in this chapter.

Instead of using one particular quantile, like the median in Section 7.4, one can estimate σ^2 by an integral over all quantiles, which allows one to avoid considering distribution-dependent quantities like (7.13).

Indeed, with the notation $q_\alpha = G^{-1}(1-\alpha)$ where G is the c.d.f. of $(\sigma\xi_1)^2$ and $0 < \alpha < 1$, the variance of the noise can be expressed as

$$\sigma^2 = \mathbf{E}(\sigma\xi_1)^2 = \int_0^1 q_\alpha \, d\alpha.$$

Discarding the higher order quantiles that are dubious in the presence of outliers and replacing q_α by the empirical quantile $\hat{q}_\alpha$ of level α we obtain the following estimator

$$\hat{\sigma}^2 = \int_0^{1-s/d} \hat{q}_\alpha \, d\alpha = \frac{1}{d} \sum_{k=1}^{d-s} Y_{(k)}^2, \tag{7.16}$$

where $Y_{(1)}^2 \le \ldots \le Y_{(d)}^2$ are the ordered values of the squared observations $Y_1^2, \ldots, Y_d^2$. Note that $\hat{\sigma}^2$ is an L-estimator, *cf.* Huber (2011). Also, up to a constant factor, $\hat{\sigma}^2$ coincides with the statistic used in Collier et al. (2017) .

The following theorem provides an upper bound on the risk of the estimator $\hat{\sigma}^2$ under the assumption that the noise belongs to the class $\mathcal{G}_{a,\tau}$. Set

$$\phi_{\mathsf{exp}}(s,d) = \max\left(\frac{1}{\sqrt{d}}, \frac{s}{d}\log^{2/a}\left(\frac{ed}{s}\right)\right).$$

Theorem 7.4.1. *Let $\tau > 0$, $a > 0$, and let s, d be integers satisfying $1 \le s < d/2$. Then, the estimator $\hat{\sigma}^2$ defined in (7.16) satisfies*

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\hat{\sigma}^2 - \sigma^2\right)^2}{\sigma^4} \le C\phi_{\mathsf{exp}}^2(s,d) \tag{7.17}$$

where $C > 0$ is a constant depending only on a and τ.

The next theorem establishes the performance of variance estimation in the case of distributions with polynomially decaying tails. Set

$$\phi_{\mathsf{pol}}(s,d) = \max\left(\frac{1}{\sqrt{d}}, \left(\frac{s}{d}\right)^{1-\frac{2}{a}}\right).$$

Theorem 7.4.2. *Let $\tau > 0, a > 4$, and let s, d be integers satisfying $1 \le s < d/2$. Then, the estimator $\hat{\sigma}^2$ defined in (7.16) satisfies*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\hat{\sigma}^2 - \sigma^2\right)^2}{\sigma^4} \le C\phi_{\mathsf{pol}}^2(s,d), \tag{7.18}$$

where $C > 0$ is a constant depending only on a and τ.

We assume here that the noise distribution has a moment of order greater than 4, which is close to the minimum requirement since we deal with the expected squared error of a quadratic function of the observations.

We now state the lower bounds matching the results of Theorems 7.4.1 and 7.4.2.

Theorem 7.4.3. *Let $\tau > 0$, $a > 0$, and let s, d be integers satisfying $1 \le s \le d$. Let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\left(c(\phi_{\mathsf{exp}}(s,d))^{-1}\left|\frac{\hat{T}}{\sigma^2} - 1\right|\right) \ge c', \tag{7.19}$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators and $c > 0$, $c' > 0$ are constants depending only on $\ell(\cdot)$, a and τ.

Theorems 7.4.1 and 7.4.3 imply that the estimator $\hat{\sigma}^2$ is rate optimal in a minimax sense when the noise belongs to $\mathcal{G}_{a,\tau}$, in particular when it is sub-Gaussian. Interestingly, an extra logarithmic factor appears in the optimal rate when passing from the pure Gaussian distribution of ξ_i's (*cf.* Proposition 7.4.2) to the class of all sub-Gaussian distributions. This factor can be seen as a price to pay for the lack of information regarding the exact form of the distribution. Also note that this logarithmic factor vanishes as $a \to \infty$.

Under polynomial tail assumption on the noise, we have the following minimax lower bound.

Theorem 7.4.4. *Let $\tau > 0$, $a \geq 2$, and let s, d be integers satisfying $1 \leq s \leq d$. Let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\Big(c(\phi_{\mathsf{pol}}(s,d))^{-1}\Big|\frac{\hat{T}}{\sigma^2} - 1\Big|\Big) \geq c' \tag{7.20}$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators and $c > 0$, $c' > 0$ are constants depending only on $\ell(\cdot)$, a and τ.

This theorem shows that the rate $\phi_{\mathsf{pol}}(s,d)$ obtained in Theorem 7.4.2 cannot be improved in a minimax sense.

A drawback of the estimator defined in (7.16) is in the lack of adaptivity to the sparsity parameter s. At first sight, it may seem that the estimator

$$\hat{\sigma}_*^2 = \frac{2}{d} \sum_{1 \leq k \leq d/2} Y_{(k)}^2 \tag{7.21}$$

could be taken as its adaptive version. However, $\hat{\sigma}_*^2$ is not a good estimator of σ^2 as can be seen from the following proposition.

Proposition 7.4.3. *Define $\hat{\sigma}_*^2$ as in (7.21). Let $\tau > 0$, $a \geq 2$, and let s, d be integers satisfying $1 \leq s \leq d$, and $d = 4k$ for an integer k. Then,*

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\big(\hat{\sigma}_*^2 - \sigma^2\big)^2}{\sigma^4} \geq \frac{1}{64}.$$

On the other hand, it turns out that a simple plug-in estimator

$$\hat{\sigma}^2 = \frac{1}{d}\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|_2^2 \tag{7.22}$$

with $\hat{\boldsymbol{\theta}}$ chosen as in Section 7.2 achieves rate optimality adaptively to the noise distribution and to the sparsity parameter s. This is detailed in the next theorem.

Theorem 7.4.5. *Let s and d be integers satisfying $1 \leq s < \lfloor \gamma d \rfloor / 4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. Let $\hat{\sigma}^2$ be the estimator defined by (7.22) where $\hat{\boldsymbol{\theta}}$ is defined in (7.5). Then the following properties hold.*

1. *Let $\tau > 0, a > 0$. There exist constants $c, C > 0$ and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\lambda_j = c \log^{1/a}(ed/j), j = 1, \ldots, d$, we have*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\big|\hat{\sigma}^2 - \sigma^2\big| \leq C\sigma^2 \phi_{\mathsf{exp}}(s,d).$$

2. *Let $\tau > 0, a > 4$. There exist constants $c, C > 0$ and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\lambda_j = c(d/j)^{1/a}, j = 1, \dots, d$, we have*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left| \hat{\sigma}^2 - \sigma^2 \right| \le C \sigma^2 \phi_{\mathsf{pol}}(s, d).$$

7.5 Appendix: Proofs of the upper bounds

Proof of Proposition 7.2.1

Fix $\boldsymbol{\theta} \in \Theta_s$ and let S be the support of $\boldsymbol{\theta}$. We will call outliers the observations Y_i with $i \in S$. There are at least $m - s$ blocks B_i that do not contain outliers. Denote by J a set of $m - s$ indices i, for which B_i contains no outliers.

As $a > 2$, there exist constants $L = L(a, \tau)$ and $r = r(a, \tau) \in (1, 2]$ such that $\mathbf{E}|\xi_1^2 - 1|^r \le L$. Using von Bahr-Esseen inequality (*cf.* Petrov (1995)) and the fact that $|B_i| \ge k$ we get

$$\mathbf{P}\Big(\Big| \frac{1}{|B_i|} \sum_{j \in B_i} \xi_j^2 - 1 \Big| > 1/2 \Big) \le \frac{2^{r+1} L}{k^{r-1}}, \quad i = 1, \dots, m.$$

Hence, there exists a constant $C_1 = C_1(a, \tau)$ such that if $k \ge C_1$ (i.e., if γ is small enough depending on a and τ), then

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\bar{\sigma}_i^2 \notin I) \le \frac{1}{4}, \quad i = 1, \dots, m, \tag{7.23}$$

where $I = [\frac{\sigma^2}{2}, \frac{3\sigma^2}{2}]$. Next, by the definition of the median, for any interval $I \subseteq \mathbb{R}$ we have

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\tilde{\sigma}^2 \notin I) \le \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\Big(\sum_{i=1}^m \mathbb{1}_{\bar{\sigma}_i^2 \notin I} \ge \frac{m}{2} \Big) \le \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\Big(\sum_{i \in J} \mathbb{1}_{\bar{\sigma}_i^2 \notin I} \ge \frac{m}{2} - s \Big). \tag{7.24}$$

Now, $s \le \frac{\lfloor \gamma d \rfloor}{4} = \frac{m}{4}$, so that $\frac{m}{2} - s \ge \frac{m-s}{3}$. Set $\eta_i = \mathbb{1}_{\bar{\sigma}_i^2 \notin I}$, $i \in J$. Due to (7.23) we have $\mathbf{E}(\eta_i) \le 1/4$, and $(\eta_i, i \in J)$ are independent. Using these remarks and Hoeffding's inequality we find

$$\mathbf{P}\Big(\sum_{i \in J} \eta_i \ge \frac{m}{2} - s \Big) \le \mathbf{P}\Big(\sum_{i \in J} (\eta_i - \mathbf{E}(\eta_i)) \ge \frac{m - s}{12} \Big) \le \exp(-C(m - s)).$$

Note that $|J| = m - s \ge 3m/4 = 3\lfloor \gamma d \rfloor / 4$. Thus, if γ is chosen small enough depending only on a and τ then

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\tilde{\sigma}^2 \notin I) \le \exp(-Cd).$$

This proves the desired bound in probability. To obtain the bounds in expectation, set $Z = |\tilde{\sigma}^2 - \sigma^2|$. Let first $a > 4$ and take some $r \in (1, a/4)$. Then

$$\begin{aligned}
\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(Z^2\right) &\le \frac{\sigma^4}{4} + \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(Z^2 \mathbb{1}_{Z \ge \frac{\sigma^2}{2}} \right) \\
&\le \frac{9\sigma^4}{4} + 2\left(\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\tilde{\sigma}^{4r} \right) \right)^{1/r} \left(\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(Z \ge \sigma^2/2 \right) \right)^{1 - 1/r} \\
&\le \frac{9\sigma^4}{4} + 2\left(\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\tilde{\sigma}^{4r} \right) \right)^{1/r} \exp(-Cd).
\end{aligned}$$

Since $m \geq 4s$, we can easily argue that $\tilde{\sigma}^{4r} \leq \sum_{i \in J} \bar{\sigma}_i^{4r}$. It follows that

$$\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\tilde{\sigma}^{4r}\right) \leq C\sigma^{4r} d^2.$$

Hence $\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(Z^2\right) \leq C\sigma^4$. Similarly, if $a > 2$, then $\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(Z\right) \leq C\sigma^2$.

Proof of Theorem 7.2.1

Set $\boldsymbol{u} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$. It follows from Lemma A.2 in Bellec et al. (2018) that

$$2\|\boldsymbol{u}\|_2^2 \leq 2\sigma \sum_{i=1}^{d} \xi_i u_i + \tilde{\sigma}\|\boldsymbol{\theta}\|_* - \tilde{\sigma}\|\hat{\boldsymbol{\theta}}\|_*,$$

where u_i are the components of $\boldsymbol{u}$. Next, Lemma A.1 in Bellec et al. (2018) yields

$$\|\boldsymbol{\theta}\|_* - \|\hat{\boldsymbol{\theta}}\|_* \leq \Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} \|\boldsymbol{u}\|_2 - \sum_{j=s+1}^{d} \lambda_j |u|_{(d-j+1)}$$

where $|u|_{(k)}$ is the kth order statistic of $|u_1|, \ldots, |u_d|$. Combining these two inequalities we get

$$2\|\boldsymbol{u}\|_2^2 \leq 2\sigma \sum_{j=1}^{d} \xi_j u_j + \tilde{\sigma}\Big\{\Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} \|\boldsymbol{u}\|_2 - \sum_{j=s+1}^{d} \lambda_j |u|_{(d-j+1)}\Big\}. \tag{7.25}$$

For some permutation $(\varphi(1), \ldots, \varphi(d))$ of $(1, \ldots, d)$, we have

$$\Big|\sum_{i=1}^{d} \xi_j u_j\Big| \leq \sum_{j=1}^{d} |\xi|_{(d-j+1)} |u_{\varphi(j)}| \leq \sum_{j=1}^{d} |\xi|_{(d-j+1)} |u|_{(d-j+1)}, \tag{7.26}$$

where the last inequality is due to the fact that the sequence $|\xi|_{(d-j+1)}$ is non-increasing. Hence

$$\begin{aligned} 2\|\boldsymbol{u}\|_2^2 &\leq 2\sigma \sum_{j=1}^{s} |\xi|_{(d-j+1)} |u|_{(d-j+1)} + \tilde{\sigma}\Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} \|\boldsymbol{u}\|_2 + \sum_{j=s+1}^{d} \left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right) |u|_{(d-j+1)} \\ &\leq \left\{2\sigma\Big(\sum_{j=1}^{s} |\xi|_{(d-j+1)}^2\Big)^{1/2} + \tilde{\sigma}\Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} + \Big(\sum_{j=s+1}^{d} \left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\Big)^{1/2}\right\} \|\boldsymbol{u}\|_2. \end{aligned}$$

This implies

$$\|\boldsymbol{u}\|_2^2 \leq C\left\{\sigma^2 \sum_{j=1}^{s} |\xi|_{(d-j+1)}^2 + \tilde{\sigma}^2 \sum_{j=1}^{s} \lambda_j^2 + \sum_{j=s+1}^{d} \left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right\}.$$

From Lemmas 7.7.1 and 7.7.2 we have $\mathbf{E}(|\xi|_{(d-j+1)}^2) \leq C\lambda_j^2$. Using this and Proposition 7.2.1 we obtain

$$\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\|\boldsymbol{u}\|_2^2\right) \leq C\left(\sigma^2 \sum_{j=1}^{s} \lambda_j^2 + \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\sum_{j=s+1}^{d} \left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right)\right). \tag{7.27}$$

Define the events $\mathcal{A}_j = \left\{|\xi|_{(d-j+1)} \le \lambda_j/4\right\} \cap \left\{1/2 \le \tilde{\sigma}^2/\sigma^2 \le 3/2\right\}$ for $j = s+1, \dots, d$. Then

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right) \le 4\sigma^2\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}|\xi|_{(d-j+1)}^2\mathbb{1}_{\mathcal{A}_j^c}\right).$$

Fixing some $1 < r < a/2$ we get

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right) \le 4\sigma^2\sum_{j=s+1}^{d}\mathbf{E}\left(|\xi|_{(d-j+1)}^{2r}\right)^{1/r}\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\mathcal{A}_j^c\right)^{1-1/r}.$$

Lemmas 7.7.1, 7.7.2 and the definitions of parameters λ_j imply that

$$\mathbf{E}\left(|\xi|_{(d-j+1)}^{2r}\right)^{1/r} \le C\lambda_s^2, \quad j = s+1, \dots, d.$$

Furthermore, it follows from the proofs of Lemmas 7.7.1 and 7.7.2 that if the constant c in the definition of λ_j is chosen large enough, then $\mathbf{P}(|\xi|_{(d-j+1)} > \lambda_j/4) \le q^j$ for some $q < 1/2$ depending only on a and τ. This and Proposition 7.2.1 imply that $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}(\mathcal{A}_j^c) \le e^{-cd} + q^j$. Hence,

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right) \le C\sigma^2\lambda_s^2\sum_{j=s+1}^{d}(e^{-cd}+q^j)^{1-1/r} \le C'\sigma^2\sum_{j=1}^{s}\lambda_j^2.$$

Combining this inequality with (7.27) we obtain

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\|\boldsymbol{u}\|_2^2\right) \le C\sigma^2\sum_{j=1}^{s}\lambda_j^2. \tag{7.28}$$

To complete the proof, it remains to note that $\sum_{j=1}^s \lambda_j^2 \le C(\phi^*_{\mathsf{pol}}(s,d))^2$ in the polynomial case and $\sum_{j=1}^s \lambda_j^2 \le C(\phi^*_{\mathsf{exp}}(s,d))^2$ in the exponential case, cf. Lemma 7.7.3.

Proof of part (i) of Proposition 7.3.2

We consider separately the "dense" zone $s > \sqrt{d}$ and the "sparse" zone $s \le \sqrt{d}$. Let first $s > \sqrt{d}$. Then the rate $\phi^*_{\mathcal{N}(0,1)}(s,d)$ is of order $\sqrt{\frac{s}{1+\log_+(s^2/d)}}$. Thus, for $s > \sqrt{d}$ we need to prove that

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\left(\left|\frac{\hat{N}^* - \|\boldsymbol{\theta}\|_2}{\sigma}\right|^2\right) \le \frac{Cs}{1+\log_+(s^2/d)}. \tag{7.29}$$

Denoting $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d)$ we have

$$\begin{aligned}
\left|\hat{N}^* - \|\boldsymbol{\theta}\|_2\right| &= \left|\left|\sum_{j=1}^{d}Y_j^2 - d\hat{\sigma}^2\right|^{1/2} - \|\boldsymbol{\theta}\|_2\right| \qquad (7.30)\\
&= \left|\sqrt{\left|\|\boldsymbol{\theta}\|_2^2 + 2\sigma(\boldsymbol{\theta},\boldsymbol{\xi}) + \sigma^2\|\boldsymbol{\xi}\|_2^2 - d\hat{\sigma}^2\right|} - \|\boldsymbol{\theta}\|_2\right|\\
&\le \left|\sqrt{\left|\|\boldsymbol{\theta}\|_2^2 + 2\sigma(\boldsymbol{\theta},\boldsymbol{\xi})\right|} - \|\boldsymbol{\theta}\|_2\right| + \sigma\sqrt{\left|\|\boldsymbol{\xi}\|_2^2 - d\right|} + \sqrt{d|\sigma^2 - \hat{\sigma}^2|}.
\end{aligned}$$

The first term in the last line vanishes if $\boldsymbol{\theta} = 0$, while for $\boldsymbol{\theta} \neq 0$ it is bounded as follows:

$$\left| \sqrt{\left| \|\boldsymbol{\theta}\|_2^2 + 2\sigma(\boldsymbol{\theta}, \boldsymbol{\xi}) \right|} - \|\boldsymbol{\theta}\|_2 \right| = \|\boldsymbol{\theta}\|_2 \left| \sqrt{\left| 1 + \frac{2\sigma(\boldsymbol{\theta}, \boldsymbol{\xi})}{\|\boldsymbol{\theta}\|_2^2} \right|} - 1 \right| \leq \frac{2\sigma |(\boldsymbol{\theta}, \boldsymbol{\xi})|}{\|\boldsymbol{\theta}\|_2} \tag{7.31}$$

where we have used the inequality $|\sqrt{|1+x|} - 1| \leq |x|$, $\forall x \in \mathbb{R}$. Since here $|(\boldsymbol{\theta}, \boldsymbol{\xi})|/\|\boldsymbol{\theta}\|_2 \sim \mathcal{N}(0,1)$ we have, for all $\boldsymbol{\theta}$,

$$\mathbf{E} \left(\left| \sqrt{\left| \|\boldsymbol{\theta}\|_2^2 + 2\sigma(\boldsymbol{\theta}, \boldsymbol{\xi}) \right|} - \|\boldsymbol{\theta}\|_2 \right|^2 \right) \leq 4\sigma^2, \tag{7.32}$$

and since $\|\boldsymbol{\xi}\|_2^2$ has a chi-square distribution with d degrees of freedom we have

$$\mathbf{E}\Big(\big| \|\boldsymbol{\xi}\|_2^2 - d \big| \Big) \;\; \leq \;\; \Big(\mathbf{E}\Big(\big| \|\boldsymbol{\xi}\|_2^2 - d \big|^2 \Big) \Big)^{1/2} = \sqrt{2d}.$$

Next, by Proposition 7.4.2 we have that, for $s > \sqrt{d}$,

$$\sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta}, \mathcal{N}(0,1), \sigma} \left(\left| \frac{\hat{\sigma}^2}{\sigma^2} - 1 \right| \right) \leq \frac{Cs}{d(1 + \log_+(s^2/d))} \tag{7.33}$$

for some absolute constant $C > 0$. Combining (7.30) – (7.33) yields (7.29).

Let now $s \leq \sqrt{d}$. Then the rate $\phi^*_{\mathcal{N}(0,1)}(s,d)$ is of order $\sqrt{s \log(1 + d/s^2)}$. Thus, for $s \leq \sqrt{d}$ we need to prove that

$$\sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta}, \mathcal{N}(0,1), \sigma} \left(\left| \frac{\hat{N}^* - \|\boldsymbol{\theta}\|_2}{\sigma} \right|^2 \right) \leq Cs \log(1 + d/s^2). \tag{7.34}$$

We have

$$\begin{aligned} \left| \hat{N}^* - \|\boldsymbol{\theta}\|_2 \right| &= \left| \left| \sum_{j=1}^{d} (Y_j^2 \; \mathbb{1}_{\{|Y_j| > \rho_j\}}) - d\alpha\hat{\sigma}^2 \right|^{1/2} - \|\boldsymbol{\theta}\|_2 \right| \qquad (7.35) \\ &= \left| \left| \sum_{j \in S} (Y_j^2 \; \mathbb{1}_{\{|Y_j| > \rho_j\}}) + \sigma^2 \sum_{j \notin S} (\xi_j^2 \; \mathbb{1}_{\{\sigma|\xi_j| > \rho_j\}}) - d\alpha\hat{\sigma}^2 \right|^{1/2} - \|\boldsymbol{\theta}\|_2 \right| \\ &\leq \left| \sqrt{\sum_{j \in S} (Y_j^2 \; \mathbb{1}_{\{|Y_j| > \rho_j\}})} - \|\boldsymbol{\theta}\|_2 \right| + \left| \sigma^2 \sum_{j \notin S} (\xi_j^2 \; \mathbb{1}_{\{\sigma|\xi_j| > \rho_j\}}) - d\alpha\hat{\sigma}^2 \right|^{1/2}. \end{aligned}$$

Here,

$$\begin{aligned} \left| \sqrt{\sum_{j \in S} (Y_j^2 \; \mathbb{1}_{\{|Y_j| > \rho_j\}})} - \|\boldsymbol{\theta}\|_2 \right| &\leq \left| \sqrt{\sum_{j \in S} (Y_j \; \mathbb{1}_{\{|Y_j| > \rho_j\}} - \theta_j)^2} \right| \qquad (7.36) \\ &\leq \sqrt{\sum_{j \in S} \rho_j^2} + \sigma \sqrt{\sum_{j \in S} \xi_j^2}. \end{aligned}$$

Hence, writing for brevity $\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma} = \mathbf{E}$, we get

$$\mathbf{E}\left(\left|\sqrt{\sum_{j\in S}(Y_j^2\ \mathbb{1}_{\{|Y_j|>\rho_j\}})} - \|\boldsymbol{\theta}\|_2\right|^2\right) \leq 16\mathbf{E}\left(\hat{\sigma}^2_{\mathsf{med},1} + \hat{\sigma}^2_{\mathsf{med},2}\right) s\log\left(1+d/s^2\right) + 2\sigma^2 s$$
$$\leq C\sigma^2 s\log(1+d/s^2),$$

where we have used the fact that $\mathbf{E}\left(|\hat{\sigma}^2_{\mathsf{med},\mathsf{k}} - \sigma^2|\right) \leq C\sigma^2$, $\mathsf{k} = 1, 2$, by Proposition 7.4.1. Next, we study the term $\Gamma = \left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}}) - d\alpha\hat{\sigma}^2\right|$. We first write

$$\Gamma \leq \left|\sigma^2\sum_{j\notin S}\xi_j^2(\ \mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}} - \ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}})\right| + \left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat{\sigma}^2\right|, \quad (7.37)$$

where $t_* = 2\sigma\sqrt{2\log(1+d/s^2)}$. For the second summand on the right hand side of (7.37) we have

$$\left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat{\sigma}^2\right| \leq \sigma^2\left|\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - (d-|S|)\alpha\right| + \left|\sigma^2-\hat{\sigma}^2\right| d\alpha + |S|\alpha\sigma^2,$$

where $|S|$ denotes the cardinality of S. By Proposition 7.4.2 we have $\mathbf{E}(|\hat{\sigma}^2 - \sigma^2|) \leq C/\sqrt{d}$ for $s \leq \sqrt{d}$. Hence,

$$\mathbf{E}\left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat{\sigma}^2\right| \leq \sigma^2\sqrt{d\mathbf{E}\left(\xi_1^4\ \mathbb{1}_{\{|\xi_1|>\sqrt{2\log(1+d/s^2)}\}}\right)} + C\alpha\sigma^2\left(\sqrt{d}+s\right).$$

It is not hard to check (cf., e.g., (Collier et al., 2017, Lemma 4)) that, for $s \leq \sqrt{d}$,

$$\alpha \leq C(\log\left(1+d/s^2\right))^{1/2}\frac{s^2}{d},$$

and

$$\mathbf{E}\left(\xi_1^4\ \mathbb{1}_{\{|\xi_1|>\sqrt{2\log(1+d/s^2)}\}}\right) \leq C(\log\left(1+d/s^2\right))^{3/2}\frac{s^2}{d},$$

so that

$$\mathbf{E}\left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat{\sigma}^2\right| \leq C\sigma^2 s\log(1+d/s^2).$$

Thus, to complete the proof it remains to show that

$$\sigma^2\sum_{j\notin S}\mathbf{E}\left|\xi_j^2(\mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}} - \mathbb{1}_{\{\sigma|\xi_j|>t_*\}})\right| \leq C\sigma^2 s\log(1+d/s^2). \quad (7.38)$$

Recall that ρ_j is independent from ξ_j. Hence, conditioning on ρ_j we obtain

$$\sigma^2\mathbf{E}\left(\left|\xi_j^2(\mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}} - \mathbb{1}_{\{\sigma|\xi_j|>t_*\}})\right|\rho_j\right) \leq |\rho_j^2 - t_*^2|e^{-t_*^2/(8\sigma^2)} + \sigma^2\mathbb{1}_{\{\rho_j<t_*/2\}}, \quad (7.39)$$

where we have used the fact that, for $b > a > 0$,

$$\int_a^b x^2 e^{-x^2/2}dx \leq \int_a^b xe^{-x^2/4}dx \leq |b^2-a^2|e^{-\min(a^2,b^2)/4}/2.$$

Using Proposition 7.4.1 and definitions of ρ_j and t_*, we get that, for $s \le \sqrt{d}$,

$$\begin{aligned}\mathbf{E}\left(|\rho_j^2 - t_*^2|\right) e^{-t_*^2/(8\sigma^2)} &\le 8 \max_{\mathsf{k}=1,2} \mathbf{E}(|\hat{\sigma}^2_{\mathsf{med},\mathsf{k}} - \sigma^2|) \frac{s^2}{d} \log(1 + d/s^2) \\ &\le C\sigma^2 \frac{s}{d} \log(1 + d/s^2).\end{aligned} \tag{7.40}$$

Next, it follows from Proposition 7.4.1 that there exists $\gamma \in (0, 1/8)$ small enough such that for $s \le \gamma d$ we have $\max_{\mathsf{k}=1,2} \mathbf{P}(\hat{\sigma}^2_{\mathsf{med},\mathsf{k}} < \sigma^2/2) \le 2e^{-c_\gamma d}$ where $c_\gamma > 0$ is a constant. Thus, $\sigma^2 \mathbf{P}(\rho_j < t_*/2) \le 2\sigma^2 e^{-c_\gamma d} \le C\sigma^2 (s/d) \log(1 + d/s^2)$. Combining this with (7.39) and (7.40) proves (7.38).

Proof of part (i) of Proposition 7.3.3 and part (i) of Proposition 7.3.4

We only prove Proposition 7.3.3 since the proof of Proposition 7.3.4 is similar taking into account that $\mathbf{E}(\xi_1^4) < \infty$. We consider separately the "dense" zone $s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$ and the "sparse" zone $s \le \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$. Let first $s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$. Then the rate $\phi^\circ_{\mathsf{exp}}(s, d)$ is of order $d^{1/4}$ and thus we need to prove that

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\big(|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2\big) \le C\sigma^2\sqrt{d}.$$

Since σ is known, arguing similarly to (7.30) - (7.31) we find

$$|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2| \le \left|\frac{2\sigma|(\boldsymbol{\theta}, \boldsymbol{\xi})|}{\|\boldsymbol{\theta}\|_2}\right| \mathbb{1}_{\boldsymbol{\theta}\neq 0} + \sigma\sqrt{\big|\|\boldsymbol{\xi}\|_2^2 - d\big|}.$$

As $\mathbf{E}(\xi_1^4) < \infty$, this implies

$$\mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\big(|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2\big) \le 8\sigma^2 + C\sigma^2\sqrt{d},$$

which proves the result in the dense case. Next, in the sparse cas $s \le \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$, we need to prove that

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\big(|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2\big) \le C\sigma^2 s \log^{\frac{2}{a}}(ed).$$

This is immediate by Theorem 7.2.1 and the fact that $|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2 \le \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2$ for the plug-in estimator $\hat{N}^\circ_{\mathsf{exp}} = \|\hat{\boldsymbol{\theta}}\|_2$.

Proof of Proposition 7.4.1

Denote by G the cdf of $(\sigma\xi_1)^2$ and by G_d the empirical cdf of $((\sigma\xi_i)^2 : i \notin S)$, where S is the support of $\boldsymbol{\theta}$. Let M be the median of G, that is $G(M) = 1/2$. By the definition of $\hat{M}$,

$$|F_d(\hat{M}) - 1/2| \le |F_d(M) - 1/2|. \tag{7.41}$$

It is easy to check that $|F_d(x) - G_d(x)| \le s/d$ for all $x > 0$. Therefore,

$$|G_d(\hat{M}) - 1/2| \le |G_d(M) - 1/2| + 2s/d. \tag{7.42}$$

The DKW inequality (Wasserman, 2013, page 99), yields that $\mathbf{P}(\sup_{x\in\mathbb{R}} |G_d(x) - G(x)| \ge u) \le 2e^{-2u^2(d-s)}$ for all $u > 0$. Fix $t > 0$ such that $\sqrt{\frac{t}{d}} + \frac{s}{d} \le 1/8$, and consider the event

$$\mathcal{A} := \left\{ \sup_{x\in\mathbb{R}} |G_d(x) - G(x)| \le \sqrt{\frac{t}{2(d-s)}} \right\}.$$

Then, $\mathbf{P}(\mathcal{A}) \ge 1 - 2e^{-t}$. On the event $\mathcal{A}$, we have

$$|G(\hat{M}) - 1/2| \le |G(M) - 1/2| + 2\left(\sqrt{\frac{t}{2(d-s)}} + \frac{s}{d}\right) \le 2\left(\sqrt{\frac{t}{d}} + \frac{s}{d}\right) \le \frac{1}{4}, \tag{7.43}$$

where the last two inequalities are due to the fact that $G(M) = 1/2$ and to the assumption about t. Notice that

$$|G(\hat{M}) - 1/2| = |G(\hat{M}) - G(M)| = \left|F(\hat{M}/\sigma^2) - F(M/\sigma^2)\right|. \tag{7.44}$$

Using (7.43), (7.44) and the fact that $M = \sigma^2 F^{-1}(1/2)$ we obtain that, on the event $\mathcal{A}$,

$$F^{-1}(1/4) \le \hat{M}/\sigma^2 \le F^{-1}(3/4). \tag{7.45}$$

This and (7.44) imply

$$|G(\hat{M}) - 1/2| \ge c_{**}\left|\hat{M}/\sigma^2 - M/\sigma^2\right| = c_{**}\beta\left|\hat{\sigma}^2_{\mathsf{med}}/\sigma^2 - 1\right|. \tag{7.46}$$

where $c_{**} = \min_{x\in[F^{-1}(1/4),F^{-1}(3/4)]} F'(x) > 0$, and $\beta = F^{-1}(1/2)$. Combining the last inequality with (7.43) we get that, on the event $\mathcal{A}$,

$$\left|\hat{\sigma}^2_{\mathsf{med}}/\sigma^2 - 1\right| \le c_{**}^{-1}\beta\left(\sqrt{\frac{t}{d}} + \frac{s}{d}\right).$$

Recall that we assumed that $\sqrt{\frac{t}{d}} + \frac{s}{d} \le 1/8$. Thus, there exists a constant $c_* > 0$ depending only on F such that for $t > 0$ and integers s, d satisfying $\sqrt{\frac{t}{d}} + \frac{s}{d} \le 1/8$ we have

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s} \mathbf{P}_{\boldsymbol{\theta},F,\sigma}\left(\left|\frac{\hat{\sigma}^2_{\mathsf{med}}}{\sigma^2} - 1\right| \ge c_*\left(\sqrt{\frac{t}{d}} + \frac{s}{d}\right)\right) \le 2e^{-t}. \tag{7.47}$$

This and the assumption that $\frac{s}{d} \le \gamma < 1/8$ imply the result of the proposition in probability. We now prove the result in expectation. Set $Z = \left|\hat{\sigma}^2_{\mathsf{med}} - \sigma^2\right|/\sigma^2$. We have

$$\mathbf{E}_{\boldsymbol{\theta},F,\sigma}(Z) \le c_* s/d + \int_{c_* s/d}^{c_*/8} \mathbf{P}_{\boldsymbol{\theta},F,\sigma}(Z > u)\, du + \mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z\mathbb{1}_{Z\ge c_*/8}\right).$$

Using (7.47), we get

$$\int_{c_* s/d}^{c_*/8} \mathbf{P}_{\boldsymbol{\theta},F,\sigma}(Z > u)\, du \le \frac{2c_*}{\sqrt{d}}\int_0^\infty e^{-t^2} dt \le \frac{C}{\sqrt{d}}.$$

As $s < d/2$, one may check that $\hat{\sigma}_{\mathsf{med}}^{2+\epsilon} \le \big(\max_{i\notin S}(\sigma\xi_i)^2/\beta\big)^{1+\epsilon/2} \le (\sigma^2/\beta)^{1+\epsilon/2}\sum_{i=1}^d |\xi_i|^{2+\epsilon}$. Since $\mathbf{E}|\xi_1|^{2+\epsilon} < \infty$ this yields $\mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z^{1+\epsilon}\right) \le Cd$. It follows that

$$\mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z\mathbb{1}_{Z\ge c_*/8}\right) \le \left(\mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z^{1+\epsilon}\right)\right)^{1/(1+\epsilon)}\mathbf{P}_{\boldsymbol{\theta},F,\sigma}\left(Z \ge c_*/8\right)^{\epsilon/(1+\epsilon)} \le Cde^{-d/C}.$$

Combining the last three displays yields the desired bound in expectation.

Proof of part (i) of Proposition 7.4.2

In this proof, we write for brevity $\mathbf{E} = \mathbf{E}_{\boldsymbol{\theta},\sigma,\mathcal{N}(0,1)}$ and $\mathbf{P} = \mathbf{P}_{\boldsymbol{\theta},\sigma,\mathcal{N}(0,1)}$. Set

$$\varphi_d(t) = \frac{1}{d}\sum_{i=1}^d e^{itY_j}, \quad \varphi(t) = \mathbf{E}(\varphi_d(t)), \quad \varphi_0(t) = e^{-\frac{t^2\sigma^2}{2}}.$$

Since $s/d < 1/8$ and $\varphi(t) = \varphi_0(t)\big(1 - \frac{|S|}{d} + \frac{1}{d}\sum_{j\in S}\exp(i\theta_j t)\big)$, we have

$$\frac{3}{4}\varphi_0(t) \le \Big(1 - \frac{2s}{d}\Big)\varphi_0(t) \le |\varphi(t)| \le \varphi_0(t). \tag{7.48}$$

Consider the events

$$\mathcal{B}_1 = \Big\{\sigma^2/2 \le \tilde{\sigma}^2 \le 3\sigma^2/2\Big\} \quad \text{and} \quad \mathcal{A}_u = \Big\{\sup_{v\in\mathbb{R}}|\varphi_d(v) - \varphi(v)| \le \sqrt{\frac{u}{d}}\Big\}, \quad u > 0.$$

By Proposition 7.2.1, $\mathcal{B}_1$ holds with probability at least $1 - e^{-cd}$ if the tuning parameter γ in the definition of $\tilde{\sigma}^2$ is small enough. Using Hoeffding's inequality, it is not hard to check that $\mathcal{A}_u$ holds with probability at least $1 - 4e^{-u}$. Moreover,

$$\mathbf{E}\Big(\sqrt{d}\sup_{v\in\mathbb{R}}|\varphi_d(v) - \varphi(v)|\Big) \le C. \tag{7.49}$$

Notice that on the event $\mathcal{D} = \{|\varphi_d(\hat{t}_1)| > (es/\sqrt{d}+1)^{-1}/4\}$ we have $\hat{\sigma}^2 = \tilde{v}^2 \le 2\tilde{\sigma}^2$. First, we bound the risk restricted to $\mathcal{D}\cap\mathcal{B}_1^c$. We have

$$\mathbf{E}\big(|\hat{\sigma}^2 - \sigma^2|\mathbb{1}_{\mathcal{D}\cap\mathcal{B}_1^c}\big) \le \mathbf{E}\big(|2\tilde{\sigma}^2 + \sigma^2|\mathbb{1}_{\mathcal{B}_1^c}\big).$$

Thus, using the Cauchy-Schwarz inequality and Proposition 7.2.1 we find

$$\mathbf{E}\big(|\hat{\sigma}^2 - \sigma^2|\mathbb{1}_{\mathcal{D}\cap\mathcal{B}_1^c}\big) \le C\sigma^2 e^{-d/C} \le \frac{C'\sigma^2}{\sqrt{d}}. \tag{7.50}$$

Next, we bound the risk restricted to $\mathcal{D}^c$. It will be useful to note that $\mathcal{A}_{\log d}\cap\mathcal{B}_1 \subset \mathcal{D}$. Indeed, on $\mathcal{A}_{\log d}\cap\mathcal{B}_1$, using the assumption $s < d/8$ we have

$$|\varphi_d(\hat{t}_1)| \ge \frac{3}{4}\varphi_0(\hat{t}_1) - \sqrt{\frac{\log d}{d}} \ge \frac{3}{4(es/\sqrt{d}+1)^{1/3}} - \sqrt{\frac{\log d}{d}} > \frac{1}{4(es/\sqrt{d}+1)}.$$

Thus, applying again the Cauchy-Schwarz inequality and Proposition 7.2.1 we find

$$\begin{aligned}\mathbf{E}\big(|\hat{\sigma}^2 - \sigma^2|\mathbb{1}_{\mathcal{D}^c}\big) = \mathbf{E}\big(|\tilde{\sigma}^2 - \sigma^2|\mathbb{1}_{\mathcal{D}^c}\big) &\le \big(\mathbf{E}\big(|\tilde{\sigma}^2 - \sigma^2|^2\big)\big)^{1/2}\big(\mathbf{P}(\mathcal{D}^c)\big)^{1/2} \\ &\le C\sigma^2\sqrt{\mathbf{P}(\mathcal{A}_{\log d}^c) + \mathbf{P}(\mathcal{B}_1^c)} \le C\sigma^2\sqrt{\frac{4}{d} + e^{-cd}} \le \frac{C'\sigma^2}{\sqrt{d}}.\end{aligned} \tag{7.51}$$

To complete the proof, it remains to handle the risk restricted to the event $\mathcal{C} = \mathcal{D} \cap \mathcal{B}_1$. We will use the following decomposition

$$|\hat{\sigma}^2 - \sigma^2| \le \Big|\frac{2\log(|\varphi_d(\hat{t}_1)|)}{\hat{t}_1^2} - \frac{2\log(|\varphi(\hat{t}_1)|)}{\hat{t}_1^2}\Big| + \Big| - \frac{2\log(|\varphi(\hat{t}_1)|)}{\hat{t}_1^2} - \sigma^2\Big|. \tag{7.52}$$

Since $-2\log(|\varphi_0(\hat{t}_1)|)/\hat{t}_1^2 = \sigma^2$, it follows from (7.48) that

$$\Big| - \frac{2\log(|\varphi(\hat{t}_1)|)}{\hat{t}_1^2} - \sigma^2\Big| \le \frac{Cs}{d\,\hat{t}_1^2} = \frac{Cs\tilde{\sigma}^2}{d\log(4(es/\sqrt{d}+1))}.$$

Therefore,

$$\mathbf{E}\Big(\Big| - \frac{2\log(|\varphi(\hat{t}_1)|)}{\hat{t}_1^2} - \sigma^2\Big|\mathbb{1}_{\mathcal{C}}\Big) \le \frac{Cs\sigma^2}{d\log(es/\sqrt{d}+1)}. \tag{7.53}$$

Next, using the inequality

$$\big|\log(|\varphi_d(t)|) - \log(|\varphi(t)|)\big| \le \frac{|\varphi_d(t) - \varphi(t)|}{|\varphi(t)| \wedge |\varphi_d(t)|}, \quad \forall t \in \mathbb{R},$$

we find

$$\begin{aligned}\Big|\frac{\log(|\varphi_d(\hat{t}_1)|)}{\hat{t}_1^2} - \frac{\log(|\varphi(\hat{t}_1)|)}{\hat{t}_1^2}\Big|\mathbb{1}_{\mathcal{C}} &\le \frac{\sup_{v\in\mathbb{R}}|\varphi_d(v) - \varphi(v)|}{\hat{t}_1^2|\varphi(\hat{t}_1)| \wedge |\varphi_d(\hat{t}_1)|}\mathbb{1}_{\mathcal{C}} \\ &\le \frac{C\sigma^2 U}{\sqrt{d}\,\log(es/\sqrt{d}+1)}\left(\frac{es}{\sqrt{d}}+1\right),\end{aligned}$$

where $U = \sqrt{d}\,\sup_{v\in\mathbb{R}}|\varphi_d(v) - \varphi(v)|$. Bounding $\mathbf{E}(U)$ by (7.49) we finally get

$$\mathbf{E}\left[\Big|\frac{\log(|\varphi_d(\hat{t}_1)|)}{\hat{t}_1^2} - \frac{\log(|\varphi(\hat{t}_1)|)}{\hat{t}_1^2}\Big|\mathbb{1}_{\mathcal{C}}\right] \le C\sigma^2\max\left(\frac{1}{\sqrt{d}}, \frac{s}{d\log(es/\sqrt{d}+1)}\right). \tag{7.54}$$

We conclude by combining inequalities (7.50) - (7.54).

Proof of Theorems 7.4.1 and 7.4.2

Let $\|\boldsymbol{\theta}\|_0 \le s$ and denote by S the support of $\boldsymbol{\theta}$. Note first that, by the definition of $\hat{\sigma}^2$,

$$\frac{\sigma^2}{d}\sum_{i=1}^{d-2s}\xi_{(i)}^2 \le \hat{\sigma}^2 \le \frac{\sigma^2}{d}\sum_{i\in S^c}\xi_i^2, \tag{7.55}$$

where $\xi_{(1)}^2 \le \cdots \le \xi_{(d)}^2$ are the ordered values of $\xi_1^2, \ldots, \xi_d^2$. Indeed, the right hand inequality in (7.55) follows from the relations

$$\sum_{k=1}^{d-s} Y_{(k)}^2 = \min_{J:|J|=d-s}\sum_{i\in J} Y_{(i)}^2 \le \sum_{i\in S^c} Y_{(i)}^2 = \sum_{i\in S^c}\sigma^2\xi_i^2.$$

To show the left hand inequality in (7.55), notice that at least $d-2s$ among the $d-s$ order statistics $Y_{(1)}^2, \ldots, Y_{(d-s)}^2$ correspond to observations Y_k of pure noise, *i.e.*, $Y_k = \sigma\xi_k$. The

sum of squares of such observations is bounded from below by the sum of the smallest $d-2s$ values $\sigma^2\xi^2_{(1)},\dots,\sigma^2\xi^2_{(d-2s)}$ among $\sigma^2\xi_1^2,\dots,\sigma^2\xi_d^2$.

Using (7.55) we get

$$\Big(\hat\sigma^2-\frac{\sigma^2}{d}\sum_{i=1}^d\xi_i^2\Big)^2\le\frac{\sigma^4}{d^2}\Big(\sum_{i=d-2s+1}^d\xi^2_{(i)}\Big)^2,$$

so that

$$\mathbf{E}_{\boldsymbol\theta,P_\xi,\sigma}\Big(\hat\sigma^2-\frac{\sigma^2}{d}\sum_{i=1}^d\xi_i^2\Big)^2\le\frac{\sigma^4}{d^2}\Big(\sum_{i=1}^{2s}\sqrt{\mathbf{E}\xi^4_{(d-i+1)}}\Big)^2.$$

Then

$$\begin{aligned}\mathbf{E}_{\boldsymbol\theta,P_\xi,\sigma}(\hat\sigma^2-\sigma^2)^2 &\le 2\mathbf{E}_{\boldsymbol\theta,P_\xi,\sigma}\Big(\hat\sigma^2-\frac{\sigma^2}{d}\sum_{i=1}^d\xi_i^2\Big)^2+2\mathbf{E}_{\boldsymbol\theta,P_\xi,\sigma}\Big(\frac{\sigma^2}{d}\sum_{i=1}^d\xi_i^2-\sigma^2\Big)^2\\ &\le \frac{2\sigma^4}{d^2}\Big(\sum_{i=1}^{2s}\sqrt{\mathbf{E}\xi^4_{(d-i+1)}}\Big)^2+\frac{2\sigma^4\mathbf{E}(\xi_1^4)}{d}.\end{aligned}$$

Note that under assumption (7.2) we have $\mathbf{E}(\xi_1^4)<\infty$ and Lemmas 7.7.1 and 7.7.3 yield

$$\sum_{i=1}^{2s}\sqrt{\mathbf{E}\xi^4_{(d-i+1)}}\le\sqrt{C}\sum_{i=1}^{2s}\log^{2/a}\left(ed/i\right)\le C'\sqrt{C}s\log^{2/a}\left(\frac{ed}{2s}\right).$$

This proves Theorem 7.4.1. To prove Theorem 7.4.2, we act analogously by using Lemma 7.7.2 and the fact that $\mathbf{E}(\xi_1^4)<\infty$ under assumption (7.3) with $a>4$.

Proof of Theorem 7.4.5

With the same notation as in the proof of Theorem 7.2.1, we have

$$\hat\sigma^2-\sigma^2=\frac{\sigma^2}{d}\left(\|\boldsymbol\xi\|_2^2-d\right)+\frac{1}{d}\left(\|\boldsymbol u\|_2^2-2\sigma\boldsymbol u^T\boldsymbol\xi\right).\tag{7.56}$$

It follows from (7.25) that

$$\|\boldsymbol u\|_2^2+2\sigma|\boldsymbol u^T\boldsymbol\xi|\le 3\sigma|\boldsymbol u^T\boldsymbol\xi|+\frac{\tilde\sigma}{2}\Big\{\Big(\sum_{j=1}^s\lambda_j^2\Big)^{1/2}\|\boldsymbol u\|_2-\sum_{j=s+1}^d\lambda_j|u|_{(d-j+1)}\Big\}.$$

Arguing as in the proof of Theorem 7.2.1, we obtain

$$\|\boldsymbol u\|_2^2+2\sigma|\boldsymbol u^T\boldsymbol\xi|\le\Big(U_1+\frac{\tilde\sigma}{2}\Big(\sum_{j=1}^s\lambda_j^2\Big)^{1/2}+U_2\Big)\|\boldsymbol u\|_2,$$

where

$$U_1=3\sigma\Big(\sum_{j=1}^s|\xi|^2_{(d-j+1)}\Big)^{1/2},\quad U_2=\Big(\sum_{j=s+1}^d\Big(3\sigma|\xi|_{(d-j+1)}-\frac{\tilde\sigma}{2}\lambda_j\Big)_+^2\Big)^{1/2}$$

Using the Cauchy-Schwarz inequality, Proposition 7.2.1 and (7.28) and writing for brevity $\mathbf{E} = \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}$ we find

$$\mathbf{E}\Big(\tilde{\sigma}\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\|\boldsymbol{u}\|_2\Big) \le \Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\sqrt{\mathbf{E}(\tilde{\sigma}^2)}\sqrt{\mathbf{E}(\|\boldsymbol{u}\|_2^2)} \le C\sigma^2\sum_{j=1}^s \lambda_j^2.$$

Since $\mathbf{E}(\xi_1^4) < \infty$ we also have $\mathbf{E}\big|\|\boldsymbol{\xi}\|_2^2 - d\big| \le C\sqrt{d}$. Finally, using again (7.28) we get, for $k = 1, 2$,

$$\mathbf{E}(U_k\|\boldsymbol{u}\|_2) \le \sqrt{\mathbf{E}(\|\boldsymbol{u}\|_2^2)}\sqrt{\mathbf{E}(U_k^2)} \le \sigma\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\sqrt{\mathbf{E}(U_k^2)} \le C\sigma^2\sum_{j=1}^s \lambda_j^2,$$

where the last inequality follows from the same argument as in the proof of Theorem 7.2.1. These remarks together with (7.56) imply

$$\mathbf{E}\left(|\hat{\sigma}^2 - \sigma^2|\right) \le \frac{C}{d}\Big(\sigma^2\sqrt{d} + \sigma^2\sum_{j=1}^s \lambda_j^2\Big).$$

We conclude the proof by bounding $\sum_{j=1}^s \lambda_j^2$ in the same way as in the end of the proof of Theorem 7.2.1.

7.6 Appendix: Proofs of the lower bounds

Proof of Theorems 7.4.3 and 7.4.4 and part (ii) of Proposition 7.4.2

Since we have $\ell(t) \ge \ell(A)\mathbb{1}_{t>A}$ for any $A > 0$, it is enough to prove the theorems for the indicator loss $\ell(t) = \mathbb{1}_{t>A}$. This remark is valid for all the proofs of this section and will not be further repeated.

(i) We first prove the lower bounds with the rate $1/\sqrt{d}$ in Theorems 7.4.3 and 7.4.4. Let $f_0 : \mathbb{R} \to [0, \infty)$ be a probability density with the following properties: f_0 is continuously differentiable, symmetric about 0, supported on $[-3/2, 3/2]$, with variance 1 and finite Fisher information $I_{f_0} = \int (f_0'(x))^2(f_0(x))^{-1}dx$. The existence of such f_0 is shown in Lemma 7.7.7. Denote by F_0 the probability distribution corresponding to f_0. Since F_0 is zero-mean, with variance 1 and supported on $[-3/2, 3/2]$ it belongs to $\mathcal{G}_{a,\tau}$ with any $\tau > 0$, $a > 0$, and to $\mathcal{P}_{a,\tau}$ with any $\tau > 0$, $a \ge 2$. Define $\mathbf{P}_0 = \mathbf{P}_{0,F_0,1}$, $\mathbf{P}_1 = \mathbf{P}_{0,F_0,\sigma_1}$ where $\sigma_1^2 = 1 + c_0/\sqrt{d}$ and $c_0 > 0$ is a small constant to be fixed later. Denote by $H(\mathbf{P}_1, \mathbf{P}_0)$ the Hellinger distance between $\mathbf{P}_1$ and $\mathbf{P}_0$. We have

$$H^2(\mathbf{P}_1, \mathbf{P}_0) = 2\big(1 - (1 - h^2/2)^d\big) \tag{7.57}$$

where $h^2 = \int(\sqrt{f_0(x)} - \sqrt{f_0(x/\sigma_1)/\sigma_1})^2 dx$. By Theorem 7.6. in Ibragimov and Has'Minskii (2013),

$$h^2 \le \frac{(1-\sigma_1)^2}{4}\sup_{t\in[1,\sigma_1]} I(t)$$

where $I(t)$ is the Fisher information corresponding to the density $f_0(x/t)/t$, that is $I(t) = t^{-2}I_{f_0}$. It follows that $h^2 \le \bar{c}c_0^2/d$ where $\bar{c} > 0$ is a constant. This and (7.57)

imply that for c_0 small enough we have $H(\mathbf{P}_1, \mathbf{P}_0) \le 1/2$. Finally, choosing such a small c_0 and using Theorem 2.2(ii) in Tsybakov (2008) we obtain

$$\inf_{\hat{T}} \max\Big\{\mathbf{P}_0\Big(\Big|\hat{T}-1\Big| > \frac{c_0}{2(1+c_0)\sqrt{d}}\Big), \mathbf{P}_1\Big(\Big|\frac{\hat{T}}{\sigma_1^2}-1\Big| > \frac{c_0}{2(1+c_0)\sqrt{d}}\Big)\Big\}$$
$$\ge \inf_{\hat{T}} \max\Big\{\mathbf{P}_0\Big(|\hat{T}-1| > \frac{c_0}{2\sqrt{d}}\Big), \mathbf{P}_1\Big(|\hat{T}-\sigma_1^2| > \frac{c_0}{2\sqrt{d}}\Big)\Big\} \ge \frac{1-H(\mathbf{P}_1,\mathbf{P}_0)}{2} \ge \frac{1}{4}.$$

(ii) We now prove the lower bound with the rate $\frac{s}{d}\log^{2/a}(ed/s)$ in Theorem 7.4.3. It is enough to conduct the proof for $s \ge s_0$ where $s_0 > 0$ is an arbitrary absolute constant. Indeed, for $s \le s_0$ we have $\frac{s}{d}\log^{2/a}(ed/s) \le C/\sqrt{d}$ where $C > 0$ is an absolute constant and thus Theorem 7.4.3 follows already from the lower bound with the rate $1/\sqrt{d}$ proved in item (i). Therefore, in the rest of this proof we assume without loss of generality that $s \ge 32$.

We take $P_\xi = U$ where U is the Rademacher distribution, that is the uniform distribution on $\{-1, 1\}$. Clearly, $U \in \mathcal{G}_{a,\tau}$. Let $\delta_1, \dots, \delta_d$ be i.i.d. Bernoulli random variables with probability of success $\mathbf{P}(\delta_1 = 1) = \frac{s}{2d}$, and let $\epsilon_1, \dots, \epsilon_d$ be i.i.d. Rademacher random variables that are independent of $(\delta_1, \dots, \delta_d)$. Denote by μ the distribution of $(\alpha\delta_1\epsilon_1, \dots, \alpha\delta_d\epsilon_d)$ where $\alpha = (\tau/2)\log^{1/a}(ed/s)$. Note that μ is not necessarily supported on $\Theta_s = \{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_0 \le s\}$ as the number of nonzero components of a vector drawn from μ can be larger than s. Therefore, we consider a restricted to Θ_s version of μ defined by

$$\bar{\mu}(A) = \frac{\mu(A \cap \Theta_s)}{\mu(\Theta_s)} \tag{7.58}$$

for all Borel subsets A of $\mathbb{R}^d$. Finally, we introduce two mixture probability measures

$$\mathbb{P}_\mu = \int \mathbf{P}_{\boldsymbol{\theta},U,1}\,\mu(d\boldsymbol{\theta}) \quad \text{and} \quad \mathbb{P}_{\bar{\mu}} = \int \mathbf{P}_{\boldsymbol{\theta},U,1}\,\bar{\mu}(d\boldsymbol{\theta}). \tag{7.59}$$

Notice that there exists a probability measure $\tilde{P} \in \mathcal{G}_{a,\tau}$ such that

$$\mathbb{P}_\mu = \mathbf{P}_{0,\tilde{P},\sigma_0} \tag{7.60}$$

where $\sigma_0 > 0$ is defined by

$$\sigma_0^2 = 1 + \frac{\tau^2 s}{8d}\log^{2/a}(ed/s) \le 1 + \frac{\tau^2}{8}. \tag{7.61}$$

Indeed, $\sigma_0^2 = 1 + \frac{\alpha^2 s}{2d}$ is the variance of zero-mean random variable $\alpha\delta\epsilon + \xi$, where $\xi \sim U$, $\epsilon \sim U$, $\delta \sim \mathcal{B}\big(\frac{s}{2d}\big)$ and ϵ, ξ, δ are jointly independent. Thus, to prove (7.60) it is enough to show that, for all $t \ge 2$,

$$\mathbf{P}\big((\tau/2)\log^{1/a}(ed/s)\,\delta\epsilon + \xi > t\sigma_0\big) \le e^{-(t/\tau)^a}. \tag{7.62}$$

But this inequality immediately follows from the fact that for $t \ge 2$ the probability in (7.62) is smaller than

$$\mathbf{P}(\epsilon = 1, \delta = 1)\,\mathbb{1}_{(\tau/2)\log^{1/a}(ed/s) > t-1} \le \frac{s}{4d}\mathbb{1}_{\tau\log^{1/a}(ed/s) > t} \le e^{-(t/\tau)^a}. \tag{7.63}$$

Now, for any estimator $\hat{T}$ and any $u > 0$ we have

$$\begin{aligned}
&\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(\Big|\frac{\hat{T}}{\sigma^2} - 1\Big| \ge u\Big) \\
&\quad \ge \max\Big\{\mathbf{P}_{0,\tilde{P},\sigma_0}(|\hat{T} - \sigma_0^2| \ge \sigma_0^2 u), \int \mathbf{P}_{\boldsymbol{\theta},U,1}(|\hat{T} - 1| \ge u)\bar{\mu}(d\boldsymbol{\theta})\Big\} \\
&\quad \ge \max\Big\{\mathbb{P}_\mu(|\hat{T} - \sigma_0^2| \ge \sigma_0^2 u), \mathbb{P}_{\bar{\mu}}(|\hat{T} - 1| \ge \sigma_0^2 u)\Big\} \qquad (7.64)
\end{aligned}$$

where the last inequality uses (7.60). Write $\sigma_0^2 = 1 + 2\phi$ where $\phi = \frac{\tau^2 s}{16d}\log^{2/a}(ed/s)$ and choose $u = \phi/\sigma_0^2 \ge \phi/(1 + \tau^2/8)$. Then, the expression in (7.64) is bounded from below by the probability of error in the problem of distinguishing between two simple hypotheses $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar{\mu}}$, for which Theorem 2.2 in Tsybakov (2008) yields

$$\max\Big\{\mathbb{P}_\mu(|\hat{T} - \sigma_0^2| \ge \phi), \mathbb{P}_{\bar{\mu}}(|\hat{T} - 1| \ge \phi)\Big\} \ge \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}})}{2} \qquad (7.65)$$

where $V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}})$ is the total variation distance between $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar{\mu}}$. The desired lower bound follows from (7.65) and Lemma 7.7.5 for any $s \ge 32$.

(iii) Finally, we prove the lower bound with the rate $\tau^2(s/d)^{1-2/a}$ in Theorem 7.4.4. Again, we do not consider the case $s \le 32$ since in this case the rate $1/\sqrt{d}$ is dominating and Theorem 7.4.4 follows from item (i) above. For $s \ge 32$, the proof uses the same argument as in item (ii) above but we choose $\alpha = (\tau/2)(d/s)^{1/a}$. Then the variance of $\alpha\delta\epsilon + \xi$ is equal to

$$\sigma_0^2 = 1 + \frac{\tau^2(s/d)^{1-2/a}}{8}.$$

Furthermore, with this definition of σ_0^2 there exists $\tilde{P} \in \mathcal{P}_{a,\tau}$ such that (7.60) holds. Indeed, analogously to (7.62) we now have, for all $t \ge 2$,

$$\mathbf{P}\big(\alpha\,\delta\epsilon + \xi > t\sigma_0\big) \le \mathbf{P}(\epsilon = 1, \delta = 1)\,\mathbb{1}_{(\tau/2)(d/s)^{1/a} > t-1} \le \frac{s}{4d}\mathbb{1}_{\tau(d/s)^{1/a} > t} \le (t/\tau)^a. \quad (7.66)$$

To finish the proof, it remains to repeat the argument of (7.64) and (7.65) with $\phi = \frac{\tau^2(s/d)^{1-2/a}}{16}$.

Proof of Theorem 7.3.1

We argue similarly to the proof of Theorems 7.4.3 and 7.4.4, in particular, we set $\alpha = (\tau/2)\log^{1/a}(ed/s)$ when proving the bound on the class $\mathcal{G}_{a,\tau}$, and $\alpha = (\tau/2)(d/s)^{1/a}$ when proving the bound on $\mathcal{P}_{a,\tau}$. In what follows, we only deal with the class $\mathcal{G}_{a,\tau}$ since the proof for $\mathcal{P}_{a,\tau}$ is analogous. Consider the measures μ $\bar{\mu}$, $\mathbb{P}_\mu$, $\mathbb{P}_{\bar{\mu}}$ and $\tilde{P}$ defined in

Section 7.6. Similarly to (7.64), for any estimator $\hat{T}$ and any $u > 0$ we have

$$\begin{aligned}
&\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat{T} - \|\boldsymbol{\theta}\|_2| \ge \sigma u\big) \\
&\quad \ge \max\Big\{\mathbf{P}_{0,\tilde{P},\sigma_0}(|\hat{T}| \ge \sigma_0 u), \int \mathbf{P}_{\boldsymbol{\theta},U,1}(|\hat{T} - \|\boldsymbol{\theta}\|_2| \ge u)\bar{\mu}(d\boldsymbol{\theta})\Big\} \\
&\quad \ge \max\Big\{\mathbb{P}_\mu(|\hat{T}| \ge \sigma_0 u), \mathbb{P}_{\bar{\mu}}(|\hat{T} - \|\boldsymbol{\theta}\|_2| \ge \sigma_0 u)\Big\} \\
&\quad \ge \max\Big\{\mathbb{P}_\mu(|\hat{T}| \ge \sigma_0 u), \mathbb{P}_{\bar{\mu}}(|\hat{T}| < \sigma_0 u, \|\boldsymbol{\theta}\|_2 \ge 2\sigma_0 u)\Big\} \\
&\quad \ge \min_B \max\big\{\mathbb{P}_\mu(B), \mathbb{P}_{\bar{\mu}}(B^c) - \bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u)\big\} \\
&\quad \ge \min_B \frac{\mathbb{P}_\mu(B) + \mathbb{P}_{\bar{\mu}}(B^c)}{2} - \frac{\bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u)}{2}
\end{aligned} \tag{7.67}$$

where σ_0 is defined in (7.61), U denotes the Rademacher law and $\min_B$ is the minimum over all Borel sets. The third line in the last display is due to (7.60) and to the inequality $\sigma_0 \ge 1$. Since $\min_B \big\{\mathbb{P}_\mu(B) + \mathbb{P}_{\bar{\mu}}(B^c)\big\} = 1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}})$, we get

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat{T} - \|\boldsymbol{\theta}\|_2|/\sigma \ge u\big) \ge \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}}) - \bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u)}{2} \tag{7.68}$$

Consider first the case $s \ge 32$. Set $u = \frac{\alpha\sqrt{s}}{4\sigma_0}$. Then (7.94) and (7.97) imply that

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}}) \le e^{-\frac{3s}{16}}, \quad \bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u) \le 2e^{-\frac{s}{16}},$$

which, together with (7.68) and the fact that $s \ge 32$ yields the result.

Let now $s < 32$. Then we set $u = \frac{\alpha\sqrt{s}}{8\sqrt{2}\sigma_0}$. It follows from (7.95) and (7.98) that

$$1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}}) - \bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u) \ge \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) = 1\Big) = \frac{s}{2}\Big(1 - \frac{s}{2d}\Big)^{d-1}.$$

It is not hard to check that the minimum of the last expression over all integers s, d such that $1 \le s < 32$, $s \le d$, is bounded from below by a positive number independent of d. We conclude by combining these remarks with (7.68).

Proof of part (ii) of Proposition 7.3.2

The lower bound corresponding to the sparse regime (i.e. $s \le \sqrt{d}$) is already proven in Collier et al. (2017) for known variance σ. Hence, we only focus on the dense regime where we may assume without loss of generality that $s \ge \sqrt{d}$ for d large enough. The proof is inspired by ideas from Cai and Jin (2010) even if their original proof does not apply in our setting. In what follows,we will use the Fourier transform defined for any integrable function f as

$$\hat{f}(t) = \int_{\mathbb{R}} e^{-itx} f(x)\, dx. \tag{7.69}$$

In the following, C is an absolute constant whose value may change from line to line. We denote by ϕ_{σ^2} the density of $\mathcal{N}(0, \sigma^2)$. Moreover, we set $\epsilon = \frac{s}{2d} \le 1/2$, $\tau = \sqrt{\alpha \log(es^2/d)}$ with α large enough, and φ, c_0 are defined in Lemma 7.7.9.

First, we build some probability distributions on Θ_s. If $\delta_1, \dots, \delta_d \overset{iid}{\sim} \mathcal{B}(\epsilon)$, we define $\mu_i,\ i = 1, 2$ respectively the distribution of $(\delta_1 X_1^{(i)}, \dots, \delta_d X_d^{(i)})$ where

$$X_1^{(1)}, \dots, X_d^{(1)} \overset{iid}{\sim} (\phi_\varphi * g_1 d\lambda)^d, \quad X_1^{(2)}, \dots, X_d^{(2)} \overset{iid}{\sim} (g_2 d\lambda)^d \tag{7.70}$$

where g_1, g_2 are density functions given by Lemma 7.7.9 and λ is the Lebesgue measure. Then, we consider the probability distributions $\mathbf{P}_1 := \mathbf{P}_{\mu_1, \mathcal{N}(0,1), 1}$ and $\mathbf{P}_2 := \mathbf{P}_{\mu_2, \mathcal{N}(0,1), \sqrt{1+\varphi}}$ whose density functions are respectively

$$f_1 = (1-\epsilon)\phi_1 + \epsilon \phi_{1+\varphi} * g_1, \quad f_2 = (1-\epsilon)\phi_{1+\varphi} + \epsilon \phi_{1+\varphi} * g_2. \tag{7.71}$$

Then, $\bar{\mu}_1$ and $\bar{\mu}_2$ defined by

$$\bar{\mu}_i(A) = \frac{\mu_i\big(A \cap \Theta_s\big)}{\mu_i\big(\Theta_s\big)} \tag{7.72}$$

are supported on Θ_s. Now, using Theorem 2.15 in Tsybakov (2008) and the fact that $\ell(t) \geq l(a)\mathbb{1}_{t>a}$ for any $a > 0$, we get

$$\inf_{\hat{T}} \sup_{\boldsymbol{\theta} \in \Theta_s, \sigma > 0} \mathbf{E}_{\boldsymbol{\theta}, \mathcal{N}(0,1), \sigma}\, \ell\Big(\big(\phi^*_{\mathcal{N}(0,1)}(s,d)\big)^{-1} \Big| \frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma} \Big|\Big) \geq \frac{\ell(v)}{2}(1 - V'), \tag{7.73}$$

where for some $v, w \in \mathbb{R}$,

$$V' = V(\mathbf{P}_1, \mathbf{P}_2) + \bar{\mu}_1(\|\boldsymbol{\theta}\|_2 \leq w + 2\sigma \phi^*_{\mathcal{N}(0,1)} v) + \bar{\mu}_2(\|\boldsymbol{\theta}\|_2 \geq w). \tag{7.74}$$

Decomposing in particular the total-variation distance, we have

$$V' \leq 2V(\bar{\mu}_1, \mu_1) + 2V(\bar{\mu}_2, \mu_2) + \sqrt{\chi^2(\mathbf{P}_1, \mathbf{P}_2)} \tag{7.75}$$

$$+ \mu_1\big[\|\boldsymbol{\theta}\|_2 \leq w + 2\phi^*_{\mathcal{N}(0,1)} v\big] + \mu_2\big[\|\boldsymbol{\theta}\|_2 \geq w\big]. \tag{7.76}$$

The first two terms in the right-hand side are upper bounded using Lemma 7.7.5 by $4e^{-\frac{3s}{16}}$. Then, we choose

$$v = \frac{\sqrt{c + 2u} - \sqrt{c}}{2\phi^*_{\mathcal{N}(0,1)}}, \quad w = \sqrt{c}, \tag{7.77}$$

where

$$c = m_2 + \frac{m_1 - m_2}{4}, \quad u = \frac{m_1 - m_2}{4}, \quad m_i = \mu_i\big(\|\boldsymbol{\theta}\|_2^2\big). \tag{7.78}$$

Moreover, since by definition,

$$m_1 = \frac{s}{2} \int x^2 g_1 * \phi_\varphi(x)\, dx, \quad m_2 = \frac{s}{2} \int x^2 g_2(x)\, dx, \tag{7.79}$$

Lemma 7.7.9 implies that

$$m_1 + m_2 \leq \frac{\beta_1 s}{\tau^2}, \quad m_1 - m_2 = \frac{c_0 s}{2\tau^2}, \tag{7.80}$$

so that

$$\sqrt{c + 2u} - \sqrt{c} = \frac{2u}{\sqrt{c + 2u} + \sqrt{c}} \geq C \frac{m_1 - m_2}{\sqrt{m_1 + m_2}} \geq C \frac{\sqrt{s}}{\tau}, \tag{7.81}$$

and v, and thus $\ell(v)$, is lower bounded by a positive absolute constant. Finally, by Markov's and von Bahr-Esseen's inequalities von Bahr and Esseen (1965), we have

$$\mu_1\big(\|\boldsymbol{\theta}\|_2^2 \le c+2u\big) = \mu_1\big(\|\boldsymbol{\theta}\|_2^2 - m_1 \le -u\big) \le \frac{\mu_1\big\{\big|\|\boldsymbol{\theta}\|_2^2 - m_1\big|^{5/4}\big\}}{u^{5/4}} \tag{7.82}$$

$$\le \frac{s}{2u^{5/4}} \int \Big|x^2 - \epsilon \int x^2 g_1 * \phi_\varphi(x)\,dx\Big|^{5/4} g_1 * \phi_\varphi(x)\; dx \tag{7.83}$$

$$\le \frac{Cs}{u^{5/4}}\Big[\int |x|^{5/2} g_1 * \phi_\varphi(x)\,dx + \Big(\epsilon \int x^2 g_1 * \phi_\varphi(x)\,dx\Big)^{5/4}\Big], \tag{7.84}$$

and using Lemma 7.7.9 again, this is smaller than $C\beta_2 s/(\tau^{3/4}u^{5/4}) \le C\tau^{7/4}s^{-1/4}$. Applying similar arguments for the second probability and because $s \ge \sqrt{d}$ with d large enough, we get that

$$\mu_1\big(\|\boldsymbol{\theta}\|_2 \le w + 2\phi^*_{\mathcal{N}(0,1)}v\big) + \mu_2\big(\|\boldsymbol{\theta}\|_2 \ge w\big) \le \frac{C}{s^{1/5}}. \tag{7.85}$$

We conclude by applying Lemma 7.7.10 to bound $\chi^2(\mathbf{P}_{\bar\mu_1}, \mathbf{P}_{\bar\mu_2}) = (1+\chi^2(f_1,f_2))^d - 1$, so that for d large enough, $V' \le 1/2$.

Proof of part (ii) of Proposition 7.3.3 and part (ii) of Proposition 7.3.4

We argue similarly to the proof of Theorems 7.4.3 and 7.4.4, in particular, we set $\alpha = (\tau/2)\log^{1/a}(ed/s)$ when proving the bound on the class $\mathcal{G}_{a,\tau}$, and $\alpha = (\tau/2)(d/s)^{1/a}$ when proving the bound on $\mathcal{P}_{a,\tau}$. In what follows, we only deal with the class $\mathcal{G}_{a,\tau}$ since the proof for $\mathcal{P}_{a,\tau}$ is analogous. Without loss of generality we assume that $\sigma = 1$.

To prove the lower bound with the rate $\phi^\circ_{\mathsf{exp}}(s,d)$, we only need to prove it for s such that $(\phi^\circ_{\mathsf{exp}}(s,d))^2 \le c_0\sqrt{d}/\log^{2/a}(ed)$ with any small absolute constant $c_0 > 0$, since the rate is increasing with s.

Consider the measures μ $\bar\mu$, $\mathbb{P}_\mu$, $\mathbb{P}_{\bar\mu}$ defined in Section 7.6 with $\sigma_0 = 1$. Let ξ_1 be distributed with c.d.f. F_0 defined in item (i) of the proof of Theorems 7.4.3 and 7.4.4. Using the notation as in the proof of Theorems 7.4.3 and 7.4.4, we define $\tilde P$ as the distribution of $\tilde\xi_1 = \sigma_1\xi_1 + \alpha\delta_1\epsilon_1$ with $\sigma_1^2 = (1+\alpha^2 s/(2d))^{-1}$ where now δ_1 is the Bernoulli random variable with $\mathbf{P}(\delta_1 = 1) = \frac{s}{2d}(1+\alpha^2 s/(2d))^{-1}$. By construction, $\mathbf{E}\tilde\xi_1 = 0$ and $\mathbf{E}\tilde\xi_1^2 = 1$. Since the support of F_0 is in $[-3/2, 3/2]$ one can check as in item (ii) of the proof of Theorems 7.4.3 and 7.4.4 that $\tilde P \in \mathcal{G}_{a,\tau}$. Next, analogously to (7.67) - (7.68) we obtain that, for any $u > 0$,

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\theta\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,1}\big(|\hat T - \|\boldsymbol{\theta}\|_2| \ge u\big) \ge \frac{1 - V(\mathbb{P}_{\bar\mu}, P_{0,\tilde P,1}) - \bar\mu(\|\boldsymbol{\theta}\|_2 < 2u)}{2}.$$

Let $\mathbf{P}_0$ and $\mathbf{P}_1$ denote the distributions of $(\xi_1,\dots,\xi_d)$ and of $(\sigma_1\xi_1,\dots,\sigma_1\xi_d)$, respectively. Acting as in item (i) of the proof of Theorems 7.4.3 and 7.4.4 and using the bound

$$|1-\sigma_1| \le \alpha^2 s/d = \frac{\tau^2}{4}\frac{s}{d}\log^{2/a}(ed/s) \le Cc_0/\sqrt{d}$$

we find that $V(\mathbf{P}_0, \mathbf{P}_1) \le H(\mathbf{P}_0, \mathbf{P}_1) \le 2\kappa c_0^2$ for some $\kappa > 0$. Therefore, $V(\mathbb{P}_\mu, P_{0,\tilde{P},1}) = V(\mathbf{P}_0 * \mathbf{Q}, \mathbf{P}_1 * \mathbf{Q}) \le V(\mathbf{P}_0, \mathbf{P}_1) \le 2\kappa c_0^2$ where $\mathbf{Q}$ denotes the distribution of $(\alpha\delta_1\epsilon_1, \dots, \alpha\delta_d\epsilon_d)$. This bound and the fact that $V(\mathbb{P}_{\bar{\mu}}, P_{0,\tilde{P},1}) \le V(\mathbb{P}_{\bar{\mu}}, \mathbb{P}_\mu) + V(\mathbb{P}_\mu, P_{0,\tilde{P},1})$ imply

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\theta\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, 1}\big(|\hat{T} - \|\boldsymbol{\theta}\|_2| \ge u\big) \ge \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}}) - \bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2u)}{2} - \kappa c_0^2.$$

We conclude by repeating the argument after (7.68) in the proof of Theorem 7.3.1 and choosing $c_0 > 0$ small enough to guarantee that the right hand side of the last display is positive.

Proof of part (ii) of Proposition 7.4.2

The lower bound with the rate $1/\sqrt{d}$ follows from the argument as in item (i) of the proof of Theorems 7.4.3 and 7.4.4 if we replace there F_0 by the standard Gaussian distribution. The lower bound with the rate $\frac{s}{d(1+\log_+(s^2/d))}$ follows from Lemma 7.7.8 and the lower bound for estimation of $\|\boldsymbol{\theta}\|_2$ in Proposition 7.3.2.

Proof of Proposition 7.4.3

Assume that $\boldsymbol{\theta} = 0$, $\sigma = 1$ and set

$$\xi_i = \sqrt{3}\epsilon_i u_i,$$

where the ϵ_i's and the u_i are independent, with Rademacher and uniform distribution on $[0, 1]$ respectively. Then note that

$$\mathbf{E}_{0,P_\xi,1}\big(\hat{\sigma}_*^2 - 1\big)^2 \ge \big(\mathbf{E}_{0,P_\xi,1}(\hat{\sigma}_*^2) - 1\big)^2 = \Big(\mathbf{E}_{0,P_\xi,1}\Big\{\hat{\sigma}_*^2 - \frac{3}{d}\sum_{i=1}^d u_i^2\Big\}\Big)^2, \tag{7.86}$$

since $\mathbf{E}(u_i^2) = 1/3$. Note also that $\hat{\sigma}_*^2 = \frac{3}{d/2}\sum_{i=1}^{d/2} u_{(i)}^2$. Now,

$$\begin{aligned}
\frac{1}{d/2}\sum_{i=1}^{d/2} u_{(i)}^2 - \frac{1}{d}\sum_{i=1}^{d} u_i^2 &= \frac{1}{d}\sum_{i=1}^{d/2} u_{(i)}^2 - \frac{1}{d}\sum_{i=d/2+1}^{d} u_{(i)}^2 \\
&\le \frac{1}{d}\sum_{i=1}^{d/4} u_{(i)}^2 - \frac{1}{d}\sum_{i=3d/4+1}^{d} u_{(i)}^2 \\
&\le \frac{1}{4}(u_{(d/4)}^2 - u_{(3d/4)}^2).
\end{aligned}$$

Since $u_{(i)}$ follows a Beta distribution with parameters $(i, d-i+1)$ we have $\mathbf{E}(u_{(i)}^2) = \frac{i(i+1)}{(d+1)(d+2)}$, and

$$\mathbf{E}_{0,P_\xi,1}\Big(\frac{1}{d/2}\sum_{i=1}^{d/2} u_{(i)}^2 - \frac{1}{d}\sum_{i=1}^{d} u_i^2\Big) \le \frac{1}{4}\mathbf{E}_{0,P_\xi,1}(u_{(d/4)}^2 - u_{(3d/4)}^2) = -\frac{d}{8(d+2)} \le -\frac{1}{24}.$$

This and (7.86) prove the proposition.

7.7 Appendix: Technical lemmas

Lemmas for the upper bounds

Lemma 7.7.1. *Let $z_1, \dots, z_d \overset{iid}{\sim} P$ with $P \in \mathcal{G}_{a,\tau}$ for some $a, \tau > 0$ and let $z_{(1)} \le \dots \le z_{(d)}$ be the order statistics of $|z_1|, \dots, |z_d|$. Then for $u > 2^{1/a}\tau \vee 2$, we have*

$$\mathbf{P}\Big(z_{(d-j+1)} \le u \log^{1/a}\big(ed/j\big), \forall\, j = 1, \dots, d\Big) \ge 1 - 4e^{-u^a/2}, \tag{7.87}$$

and, for any $r > 0$,

$$\mathbf{E}\big(z^r_{(d-j+1)}\big) \le C \log^{r/a}\big(ed/j\big), \qquad j = 1, \dots, d, \tag{7.88}$$

where $C > 0$ is a constant depending only on τ, a and r.

Proof. Using the definition of $\mathcal{G}_{a,\tau}$ we get that, for any $t \ge 2$,

$$\mathbf{P}\big(z_{(d-j+1)} \ge t\big) \le \binom{d}{j}\mathbf{P}^j(|z_1| \ge t) \le 2\Big(\frac{ed}{j}\Big)^j e^{-j(t/\tau)^a}, \qquad j = 1, \dots, d.$$

Thus, for $v \ge 2^{1/a} \vee (2/\tau)$ we have

$$\mathbf{P}(z_{(d-j+1)} \ge v\tau \log^{1/a}(ed/j)) \le 2\Big(\frac{ed}{j}\Big)^{j(1-v^a)} \le 2e^{-jv^a/2}, \qquad j = 1, \dots, d, \tag{7.89}$$

and

$$\mathbf{P}\Big(\exists\, j \in \{1, \dots, d\} : z_{(d-j+1)} \ge v\tau \log^{1/a}(ed/j)\Big) \le 2\sum_{j=1}^{d} e^{-jv^a/2} \le 4e^{-v^a/2}$$

implying (7.87). Finally, (7.88) follows by integrating (7.89). □

Lemma 7.7.2. *Let $z_1, \dots, z_d \overset{iid}{\sim} P$ with $P \in \mathcal{P}_{a,\tau}$ for some $a, \tau > 0$ and let $z_{(1)} \le \dots \le z_{(d)}$ be the order statistics of $|z_1|, \dots, |z_d|$. Then for $u > (2e)^{1/a}\tau \vee 2$, we have*

$$\mathbf{P}\Big(z_{(d-j+1)} \le u\Big(\frac{d}{j}\Big)^{1/a}, \forall\, j = 1, \dots, d\Big) \ge 1 - \frac{2e\tau^a}{u^a} \tag{7.90}$$

and, for any $r \in (0, a)$,

$$\mathbf{E}\big(z^r_{(d-j+1)}\big) \le C\Big(\frac{d}{j}\Big)^{r/a}, \qquad j = 1, \dots, d, \tag{7.91}$$

where $C > 0$ is a constant depending only on τ, a and r.

Proof. Using the definition of $\mathcal{P}_{a,\tau}$ we get that, for any $t \ge 2$,

$$\mathbf{P}\big(z_{(d-j+1)} \ge t\big) \le \Big(\frac{ed}{j}\Big)^j \Big(\frac{\tau}{t}\Big)^{ja}.$$

Set $t_j = u\Big(\frac{d}{j}\Big)^{1/a}$ and $q = e(\tau/u)^a$. The assumption on u yields that $q < 1/2$, so that

$$\mathbf{P}\Big(\exists\, j \in \{1, \dots, d\} : z_{(d-j+1)} \ge u\Big(\frac{d}{j}\Big)^{1/a}\Big) \le \sum_{j=1}^{d} \Big(\frac{ed}{j}\Big)^j \Big(\frac{\tau}{t_j}\Big)^{ja} = \sum_{j=1}^{d} q^j \le 2q.$$

This proves (7.90). The proof of (7.91) is analoguous to that of (7.88). □

Lemma 7.7.3. *For all $a > 0$ and all integers $1 \le s \le d$,*

$$\sum_{i=1}^{s} \log^{2/a}\left(ed/i\right) \le Cs \log^{2/a}\left(\frac{ed}{s}\right)$$

where $C > 0$ depends only on a.

The proof is simple and we omit it.

Lemmas for the lower bounds

For two probability measures P_1 and P_2 on a measurable space $(\Omega, \mathcal{U})$, we denote by $V(\mathrm{P}_1, \mathrm{P}_2)$ the total variation distance between P_1 and P_2:

$$V(\mathrm{P}_1, \mathrm{P}_2) = \sup_{B \in \mathcal{U}} |\mathrm{P}_1(B) - \mathrm{P}_2(B)|\,.$$

Lemma 7.7.4 (Deviations of the binomial distribution)**.** *Let $\mathcal{B}(d,p)$ denote the binomial random variable with parameters d and $p \in (0,1)$. Then, for any $\lambda > 0$,*

$$\mathbf{P}\big(\mathcal{B}(d,p) \ge \lambda\sqrt{d} + dp\big) \le \exp\left(-\frac{\lambda^2}{2p(1-p)\big(1+\frac{\lambda}{3p\sqrt{d}}\big)}\right), \tag{7.92}$$

$$\mathbf{P}\big(\mathcal{B}(d,p) \le -\lambda\sqrt{d} + dp\big) \le \exp\left(-\frac{\lambda^2}{2p(1-p)}\right). \tag{7.93}$$

Inequality (7.92) is a combination of formulas (3) and (10) on pages 440–441 in Shorack and Wellner (2009). Inequality (7.93) is formula (6) on page 440 in Shorack and Wellner (2009).

Lemma 7.7.5. *Let $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar\mu}$ be the probability measures defined in (7.59). The total variation distance between these two measures satisfies*

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) \le \mathbf{P}\Big(\mathcal{B}\Big(d, \frac{s}{2d}\Big) > s\Big) \le e^{-\frac{3s}{16}}, \tag{7.94}$$

and

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) \le 1 - \mathbf{P}\Big(\mathcal{B}\Big(d, \frac{s}{2d}\Big) = 0\Big) - \mathbf{P}\Big(\mathcal{B}\Big(d, \frac{s}{2d}\Big) = 1\Big). \tag{7.95}$$

Proof. We have

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) = \sup_B \left| \int \mathbf{P}_{\boldsymbol{\theta},U,1}(B) d\mu(\boldsymbol{\theta}) - \int \mathbf{P}_{\boldsymbol{\theta},U,1}(B) d\bar\mu(\boldsymbol{\theta}) \right| \le \sup_{|f|\le 1} \left| \int f d\mu - \int f d\bar\mu \right| = V(\mu, \bar\mu).$$

Furthermore, $V(\mu,\bar\mu) \le \mu(\Theta_s^c)$ since for any Borel subset B of $\mathbb{R}^d$ we have $\big|\mu(B) - \bar\mu(B)\big| \le \mu(B \cap \Theta_s^c)$. Indeed,

$$\mu(B) - \bar\mu(B) \le \mu(B) - \mu(B \cap \Theta) = \mu(B \cap \Theta^c)$$

and

$$\bar\mu(B) - \mu(B) = \frac{\mu(B\cap\Theta)}{\mu(\Theta)} - \mu(B\cap\Theta) - \mu(B\cap\Theta^c) \ge -\mu(B\cap\Theta^c).$$

Thus,

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) \le \mu(\Theta_s^c) = \mathbf{P}\Big(\mathcal{B}\Big(d, \frac{s}{2d}\Big) > s\Big). \tag{7.96}$$

Combining this inequality with (7.92) we obtain (7.94). To prove (7.95), we use again (7.96) and notice that $\mathbf{P}\Big(\mathcal{B}\Big(d, \frac{s}{2d}\Big) > s\Big) \le \mathbf{P}\Big(\mathcal{B}\Big(d, \frac{s}{2d}\Big) \ge 2\Big)$ for any integer $s \ge 1$. □

Lemma 7.7.6. *Let $\bar{\mu}$ be defined in (7.58) with some $\alpha > 0$. Then*

$$\bar{\mu}\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha}{2}\sqrt{s}\Big) \le 2e^{-\frac{s}{16}}, \tag{7.97}$$

and, for all $s \le 32$,

$$\bar{\mu}\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha\sqrt{s}}{4\sqrt{2}}\Big) = \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) = 0\Big). \tag{7.98}$$

Proof. First, note that

$$\mu\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha}{2}\sqrt{s}\Big) = \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) < \frac{s}{4}\Big) \le e^{-\frac{s}{16}} \tag{7.99}$$

where the last inequality follows from (7.93). Next, inspection of the proof of Lemma 7.7.5 yields that $\bar{\mu}(B) \le \mu(B) + e^{-\frac{3s}{16}}$ for any Borel set B. Taking here $B = \{\|\boldsymbol{\theta}\|_2 \le \alpha\sqrt{s}/2\}$ and using (7.99) proves (7.97). To prove (7.98), it suffices to note that $\mu\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha\sqrt{s}}{4\sqrt{2}}\Big) = \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) < \frac{s}{32}\Big)$. □

Lemma 7.7.7. *There exists a probability density $f_0 : \mathbb{R} \to [0, \infty)$ with the following properties: f_0 is continuously differentiable, symmetric about 0, supported on $[-3/2, 3/2]$, with variance 1 and finite Fisher information $I_{f_0} = \int (f_0'(x))^2 (f_0(x))^{-1} dx$.*

Proof. Let $K : \mathbb{R} \to [0, \infty)$ be any probability density, which is continuously differentiable, symmetric about 0, supported on $[-1, 1]$, and has finite Fisher information I_K, for example, the density $K(x) = \cos^2(\pi x/2)\mathbb{1}_{|x|\le 1}$. Define $f_0(x) = [K_h(x+(1-\varepsilon)) + K_h(x-(1-\varepsilon))]/2$ where $h > 0$ and $\varepsilon \in (0, 1)$ are constants to be chosen, and $K_h(u) = K(u/h)/h$. Clearly, we have $I_{f_0} < \infty$ since $I_K < \infty$. It is straightforward to check that the variance of f_0 satisfies $\int x^2 f_0(x) dx = (1-\varepsilon)^2 + h^2\sigma_K^2$ where $\sigma_K^2 = \int u^2 K(u) du$. Choosing $h = \sqrt{2\varepsilon - \varepsilon^2}/\sigma_K$ and $\varepsilon \le \sigma_K^2/8$ guarantees that $\int x^2 f_0(x) dx = 1$ and the support of f_0 is contained in $[-3/2, 3/2]$. □

Lemma 7.7.8. *Let $\tau > 0$, $a > 4$ and let s, d be integers satisfying $1 \le s \le d$. Let $\mathcal{P}$ be any subset of $\mathcal{P}_{a,\tau}$. Assume that for some function $\phi(s, d)$ of s and d and for some positive constants c_1, c_2, c_1', c_2' we have*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\left|\frac{\hat{T}}{\sigma^2} - 1\right| \ge \frac{c_1}{\sqrt{d}}\right) \ge c_1', \tag{7.100}$$

and

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\left|\frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma}\right| \ge c_2\phi(s, d)\right) \ge c_2'. \tag{7.101}$$

Then

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\left(\left|\frac{\hat{T}}{\sigma^2} - 1\right| \ge c_3 \max\left(\frac{1}{\sqrt{d}}, \frac{\phi^2(s, d)}{d}\right)\right) \ge c_3'$$

for some constants $c_3, c_3' > 0$.

Proof. Let $\hat{\sigma}^2$ be an arbitrary estimator of σ^2. Based on $\hat{\sigma}^2$, we can construct an estimator $\hat{T} = \hat{N}^*$ of $\|\boldsymbol{\theta}\|_2$ defined by formula (7.11), case $s > \sqrt{d}$. It follows from (7.30), (7.31) and (7.101) that

$$c_2' \le \mathbf{P}\left(2|(\boldsymbol{\theta}, \boldsymbol{\xi})| \ge c_2\|\boldsymbol{\theta}\|_2\phi(s,d)/3\right) + \mathbf{P}\left(\sqrt{|\|\boldsymbol{\xi}\|_2^2 - d|} \ge c_2\phi(s,d)/3\right) + \mathbf{P}\left(\sqrt{d\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right|} \ge c_2\phi(s,d)/3\right),$$

where we write for brevity $\mathbf{P} = \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}$. Hence

$$\mathbf{P}\left(\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| \ge c_2^2\phi^2(s,d)/(9d)\right) \ge c_2' - c^* \max\left(\frac{d}{\phi^4(s,d)}, \frac{1}{\phi^2(s,d)}\right)$$

for some constant $c^* > 0$ depending only on a and τ. If $\phi^2(s,d) > \max\left(\sqrt{\frac{2c^*d}{c_2'}}, \frac{2c^*}{c_2'}\right)$, then

$$\mathbf{P}\left(\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| \ge C\max\left(\frac{1}{\sqrt{d}}, \frac{\phi^2(s,d)}{d}\right)\right) \ge c_2'/2.$$

If $\phi^2(s,d) \le \max\left(\sqrt{\frac{2c^*d}{c_2'}}, \frac{2c^*}{c_2'}\right)$, then $\max\left(\frac{1}{\sqrt{d}}, \frac{\phi^2(s,d)}{d}\right)$ is of order $\frac{1}{\sqrt{d}}$ and the result follows from (7.100). □

Lemma 7.7.9. *If c_0 is small enough, then there exist two density functions such that*

1. $\max\left\{\int_{\mathbb{R}} x^2 g_1 * \phi_\varphi(x)\,dx,\ \int_{\mathbb{R}} x^2 g_2(x)\,dx\right\} \le \frac{\beta_1}{\tau^2}$,
2. $\int_{\mathbb{R}} x^2 g_1 * \phi_\varphi(x)\,dx - \int_{\mathbb{R}} x^2 g_2(x)\,dx = \frac{c_0}{\tau^2}$,
3. $\max\left\{\int_{\mathbb{R}} |x|^{5/2} g_2(x)\,dx,\ \int_{\mathbb{R}} |x|^{5/2} g_1 * \phi_\varphi(x)\,dx\right\} \le \frac{\beta_2}{\tau^{3/4}}$

where β_1 and β_2 are absolute constants and $\varphi = c_0\epsilon/\tau^2$.

Proof. In the following, C denotes an absolute constant whose value may change from line to line. We define

$$g_1(x) = \begin{cases} 0 & \text{if } |x| \le \frac{\pi}{10\tau} \\ \frac{c}{\tau^3 x^4} & \text{if } |x| > \frac{\pi}{10\tau} \end{cases}, \quad c = \frac{3\pi^3}{16000}, \quad g_2 = g_1 + g \tag{7.102}$$

with

$$g = \frac{1}{2}\hat{h}, \quad h(t) = \begin{cases} \ell(t) & \text{if } |t| \le \tau \\ j(t) & \text{if } \tau \le |t| \le 2\tau \\ 0 & \text{if } 2\tau \le |t| \end{cases}, \quad \ell(t) = \frac{1-\epsilon}{\epsilon}(e^{\frac{\varphi t^2}{2}} - 1), \tag{7.103}$$

and

$$j(t) = (1-\epsilon)\sum_{n\ge 1} \frac{\epsilon^{n-1} c_0^n}{2^n n!}\left[c_{1,n}\frac{(2\tau - t)^2}{\tau^2} + c_{2,n}\frac{(2\tau - t)^3}{\tau^3} + c_{3,n}\frac{(2\tau - t)^4}{\tau^4}\right] \tag{7.104}$$

where

$$c_{1,n} = 2n^2 + 5n + 6, \quad c_{2,n} = -4n^2 - 8n - 8, \quad c_{3,n} = 2n^2 + 3n + 3. \tag{7.105}$$

Direct computations show that g_1 is a density function, and the first part of this proof is dedicated to proving that g_2 is a density too, if c_0 is small enough.

First note that j is bounded on $[-2\tau, 2\tau]$ so that $\hat{h}$ is well defined. Then, we can write $g(x)$ as

$$g(x) = \int_0^\tau \ell(t)\cos(tx)\,dt + \int_\tau^{2\tau} j(t)\cos(tx)\,dt. \tag{7.106}$$

Using integration by part, we have

$$\int_\tau^{2\tau} \frac{(2\tau - t)^n}{\tau^n}\cos(tx)\,dt = -\frac{\sin(\tau x)}{x} + n\frac{\cos(\tau x)}{x^2\tau} + \frac{n(n-1)}{\tau^2 x^3}\sin(\tau x) + a_n(x) \tag{7.107}$$

and

$$\int_0^\tau \frac{t^{2n}}{\tau^{2n}}\cos(tx)\,dt = \frac{\sin(\tau x)}{x} + 2n\frac{\cos(\tau x)}{x^2\tau} - 2n(2n-1)\frac{\sin(\tau x)}{x^3\tau^2} + b_n(x) \tag{7.108}$$

with

$$|a_n(x)| \le \frac{2n^3}{x^4\tau^3}, \quad |b_n(x)| \le \frac{16n^3}{\tau^3 x^4}, \tag{7.109}$$

so that

$$g(x) = (1-\epsilon)\sum_{n\ge 1}\frac{c_0^n \epsilon^{n-1}}{2^n n!}\cdot\Big[\frac{\sin(\tau x)}{x}(1 - c_{1,n} - c_{2,n} - c_{3,n}) \tag{7.110}$$

$$+ \frac{\cos(\tau x)}{\tau x^2}(2n + 2c_{1,n} + 3c_{2,n} + 4c_{3,n}) \tag{7.111}$$

$$+ \frac{\sin(\tau x)}{\tau^2 x^3}(-2n(2n-1) + 2c_{1,n} + 6c_{2,n} + 12c_{3,n}) \tag{7.112}$$

$$+ \big(a_n(x) + b_n(x)\big)\Big]. \tag{7.113}$$

The choices of the $c_{i,n}$'s make the first three parts vanish, hence

$$|g(x)| \le \frac{Cc_0}{\tau^3 x^4}. \tag{7.114}$$

On the other hand, if $0 \le x \le \pi/(10\tau)$ and $0 \le t \le 2\tau$, then $0 \le xt \le \pi/5$, so that

$$I := \int_0^\tau \frac{t^2}{\tau^2}\cos(tx)\,dt + \int_\tau^{2\tau}\Big(13\frac{(2\tau-t)^2}{\tau^2} - 20\frac{(2\tau-t)^3}{\tau^3} + 8\frac{(2\tau-t)^4}{\tau^4}\Big)\cos(tx)\,dt \tag{7.115}$$

$$\ge \cos(\pi/5)\,\Big[\int_0^\tau \frac{t^2}{\tau^2}\,dt + 13\int_\tau^{2\tau}\frac{(2\tau-t)^2}{\tau^2}\,dt + 8\int_\tau^{2\tau}\frac{(2\tau-t)^3}{\tau^3}\,dt\Big] - 20\int_\tau^{2\tau}\frac{(2\tau-t)^4}{\tau^4}\,dt \tag{7.116}$$

$$\ge \Big(\frac{94}{15}\cos(\pi/5) - 5\Big)\tau, \tag{7.117}$$

and using in particular the elementary inequality $|e^x - 1 - x| \leq ex^2/2$ for $x \in [0,1]$ and the fact that $|c_{i,n}| \leq 20n^2$,

$$\big|g(x) - (1-\epsilon)c_0 I/2\big| \leq \frac{e}{2}\int_0^\tau \Big(\frac{c_0\epsilon t^2}{2\tau^2}\Big)^2 dt + \sum_{n\geq 2}\frac{20c_0^n n^2}{2^n n!}\big[\frac{1}{3}+\frac{1}{4}+\frac{1}{5}\big] \tag{7.118}$$

$$\leq \Big(\frac{94}{15}\cos(\pi/5) - 5\Big)\tau \tag{7.119}$$

for c_0 small enough. Finally, combining (7.114), (7.117) and (7.119) yields that g_2 is positive on $\mathbb{R}$. Furthermore,

$$\int_{\mathbb{R}} g = \hat{g}(0) = \pi h(0) = 0, \tag{7.120}$$

so that $\int g_1 = \int g_2 = 1$, and g_1, g_2 both are density functions.

We now turn to the properties of g_1, g_2. We first have

$$\int x^2 g_1(x) * \phi_\varphi(x)\,dx = -(\hat{g}_1\hat{\phi}_\varphi)''(0) = -\hat{g}_1''(0)\hat{\phi}_\varphi(0) - \hat{\phi}_\varphi''(0)\hat{g}_1(0) - 2\hat{g}_1'(0)\hat{\phi}_\varphi'(0) \tag{7.121}$$

$$= \int x^2 g_1(x)\,dx + \varphi \leq \frac{C}{\tau^2} \tag{7.122}$$

and similarly

$$\int x^2 g_2(x)\,dx = \int x^2 g_1(x)\,dx - h''(0) = \int x^2 g_1(x)\,dx - \frac{(1-\epsilon)\varphi}{\epsilon} \leq \frac{C}{\tau^2}, \tag{7.123}$$

which yields the first desired property. From the same computations, we get

$$\int x^2 g_1 * \phi_\varphi(x)\,dx - \int x^2 g_2(x)\,dx = \frac{\varphi}{\epsilon} = \frac{c_0}{\tau^2}, \tag{7.124}$$

which is exactly the second property. Furthermore, we have in particular from (7.114) and the fact that $g \leq C\tau$ on $[-\tau^{-1}, \tau^{-1}]$ that

$$\int_{\mathbb{R}} |x|^{5/2} g_1(x)\,dx \leq \frac{C}{\tau^{5/2}}, \quad \int_{\mathbb{R}} |x|^{5/2} g(x)\,dx \leq \frac{C}{\tau^{5/2}}, \tag{7.125}$$

and thus $\int_{\mathbb{R}} |x|^{5/2} g_2(x)\,dx \leq C\tau^{-3/4}$. To finish establishing the last property, we note that

$$\Big|\int_{|y|\leq x/2} (g_1(x-y) - g_1(x))\phi_\varphi(y)\,dy\Big| \leq \int_{\mathbb{R}} \frac{C}{x^4\tau^3}\phi_\varphi(y)\,dy = \frac{C}{x^4\tau^3}, \tag{7.126}$$

and that, denoting $Z \sim \mathcal{N}(0,1)$ and using the fact that $g_1 \leq C\tau$,

$$\Big|\int_{|y|>x/2} (g_1(x-y) - g_1(x))\phi_\varphi(y)\,dy\Big| \leq C\tau \mathbf{P}\big(|Z| > x/2\sqrt{\phi}\big) \tag{7.127}$$

$$\leq C\tau\sqrt{\phi}x^{-1}e^{-x^2/8\varphi} \tag{7.128}$$

$$\leq \frac{C}{x^4\tau^2} \tag{7.129}$$

in the case when $x \geq \tau^{-1/2}$. Consequently, if $x \geq \tau^{-1/2}$, then

$$|g_1 * \phi_\varphi(x)| \leq |g_1 * \phi_\varphi(x) - g_1(x)| + |g_1(x)| \leq \frac{C}{x^4\tau^2} \tag{7.130}$$

Finally, using the inequality $\|g_1 * \phi_\varphi\|_\infty \leq \|g_1\|_\infty \|\phi_\varphi\|_1 \leq C\tau$ for the first integral, we get

$$\int_0^{+\infty} |x|^{5/2} g_1 * \varphi(x)\, dx \leq \int_0^{\tau^{-1/2}} |x|^{5/2} g_1 * \varphi(x)\, dx + \int_{\tau^{-1/2}}^{+\infty} |x|^{5/2} g_1 * \varphi(x)\, dx \leq \frac{C}{\tau^{3/4}}. \tag{7.131}$$

This completes the proof. □

Lemma 7.7.10. *Let f_1 and f_2 defined in (7.71) with g_1, g_2 as in Lemma 7.7.9. Then there exists an absolute constant $\beta_3 > 0$ such that, if c_0 is small enough and α is large enough, then*

$$\chi^2(f_1, f_2) \leq \frac{\beta_3 c_0}{d}. \tag{7.132}$$

Proof. Note that $f_1 \geq (1-\epsilon)\phi_1$. Since $\phi_1^{-1}(t) = \sqrt{2\pi}\sum_{n\geq 0} \frac{t^{2n}}{2^n n!}$ and $\epsilon \leq 1/2$, we get

$$\chi^2(f_1, f_2) = \int_{\mathbb{R}} \frac{(f_1 - f_2)^2}{f_1} \leq 2\sqrt{2\pi} \sum_{n\geq 0} \int_{\mathbb{R}} \frac{t^{2n}}{2^n n!} (f_1 - f_2)^2(t)\, dt. \tag{7.133}$$

But as $\hat{f}_2 - \hat{f}_1 = (1-\epsilon)(\widehat{\phi_{1+\varphi}} - \widehat{\phi_1}) + \epsilon\widehat{\phi_{1+\varphi}}(\hat{g}_2 - \hat{g}_1)$ with $\hat{g}_2 - \hat{g}_1 = h$ defined in Lemma 7.7.9, it holds that $\hat{f}_1 - \hat{f}_2$ is infinitely differentiable everywhere except in $\pm\tau$ and in $\pm 2\tau$. Thus

$$\chi^2(f_1, f_2) \leq C \sum_{n\geq 0} \frac{1}{2^n n!} \int_{\mathbb{R}} [\hat{f}_2^{(n)}(t) - \hat{f}_1^{(n)}(t)]^2\, dt = C \sum_{n\geq 0} \frac{1}{2^{n-1} n!} \int_{\tau}^{+\infty} [\hat{f}_2^{(n)}(t) - \hat{f}_1^{(n)}(t)]^2\, dt, \tag{7.134}$$

since by construction $(1-\epsilon)(\widehat{\phi_{1+\varphi}} - \widehat{\phi_1}) + \epsilon\widehat{\phi_{1+\varphi}}\ell = 0$ (*cf.* Lemma 7.7.9). Furthermore, for every $n \geq 0$,

$$\int_{\tau}^{+\infty} [\hat{f}_2^{(n)}(t) - \hat{f}_1^{(n)}(t)]^2\, dt \leq 2\epsilon^2 \int_{\tau}^{+\infty} \big([\widehat{\phi_{1+\varphi}}(t)(\hat{g}_2 - \hat{g}_1)]^{(n)}\big)^2(t)\, dt \tag{7.135}$$

$$+ 2\int_{\tau}^{+\infty} \big[(\widehat{\phi_{1+\varphi}} - \widehat{\phi_1})^{(n)}(t)\big]^2\, dt. \tag{7.136}$$

Then, note that on $[\tau, 2\tau]$, $|j^{(m)}(t)| \leq C\sum_{n\geq 1} \frac{\epsilon^{n-1} c_0^n n^2}{2^n n!} \leq C c_0$ so that

$$\int_{\tau}^{+\infty} \big([\widehat{\phi_{1+\varphi}}(\hat{g}_2 - \hat{g}_1)]^{(n)}\big)^2 = \int_{\tau}^{2\tau} \big([\widehat{\phi_{1+\varphi}}\hat{j}]^{(n)}\big)^2 \leq C c_0 \sup_{n-4\leq m\leq n} \binom{n}{m}^2 \int_{\tau}^{2\tau} \big([\widehat{\phi_{1+\varphi}}]^{(m)}\big)^2. \tag{7.137}$$

Recall that the Hermite polynomials H_m are defined by

$$H_m(x) = (-1)^m e^{x^2/2} \frac{d^m}{dx^m}\Big(e^{-x^2/2}\Big), \tag{7.138}$$

so that if $n-4 \le m \le n$,

$$\int_\tau^{2\tau} \big(\widehat{[\phi_{1+\varphi}]}^{(m)}\big)^2 \le (1+\varphi)^n \int_\tau^{2\tau} H_m^2(t\sqrt{1+\varphi}) e^{-t^2(1+\varphi)}\, dt \le (1+\varphi)^n n!\, e^{-\tau^2/2}. \tag{7.139}$$

Therefore, if α is large enough and c_0 small enough,

$$\epsilon^2 \sum_{n\ge 0} \frac{1}{2^n n!} \int_\tau^{+\infty} \big(\widehat{[\phi_{1+\varphi}(\hat g_2 - \hat g_1)]}^{(n)}\big)^2 \le C c_0 \epsilon^2 e^{-\tau^2/2} \le \frac{C c_0}{d}. \tag{7.140}$$

Coming back to the second integral in (7.136), we can apply the mean-value theorem to the k-th derivative of $f(t) = \exp(-t^2/2)$ so that for $t \ge 0$

$$\big|(\widehat{\phi_{1+\varphi}} - \widehat{\phi_1})^{(n)}(t)\big| \le [(1+\varphi)^{n/2} - 1].\big|f^{(n)}(t\sqrt{1+\varphi})\big| + t(\sqrt{1+\varphi} - 1) \sup_{u\in[t,t\sqrt{1+\varphi}]} |f^{(n+1)}(u)| \tag{7.141}$$

$$\le [(1+\varphi)^n - 1].\big|H_n(t\sqrt{1+\varphi}) e^{-\frac{(1+\varphi)t^2}{2}}\big| + t\varphi \sup_{u\in[t,t\sqrt{1+\varphi}]} |H_{n+1}(u) e^{-u^2/2}|. \tag{7.142}$$

But, integrating the square of the first term in the right-hand side, we get

$$\int_\tau^{+\infty} [(1+\varphi)^n - 1]^2 H_n^2(t\sqrt{1+\varphi}) e^{-t^2(1+\varphi)}\, dt \le (1+\varphi)^{2n} e^{-\tau^2/2} \int_0^{+\infty} H_n^2(t\sqrt{1+\varphi}) e^{-\frac{t^2(1+\varphi)}{2}}\, dt \tag{7.143}$$

$$\le C(1+\varphi)^{2n} n!\, e^{-\tau^2/2}. \tag{7.144}$$

On the other hand, using the fact that $H_n(u) = \sum_{l=0}^{\lfloor n/2\rfloor} (-1)^l \frac{n!}{2^l l!(n-2l)!} u^{n-2l}$, we have

$$\int_\tau^{+\infty} t^2 \sup_{u\in[t,t\sqrt{1+\varphi}]} |H_n(u)|^2 e^{-t^2}\, dt \le (1+\varphi)^n e^{-\tau^2/2} n \sum_{l=0}^{\lfloor n/2\rfloor} \Big(\frac{n!}{2^l l!(n-2l)!}\Big)^2 \int_0^{+\infty} t^{2n-4l+2} e^{-t^2/2}\, dt \tag{7.145}$$

$$\le C(1+\varphi)^n e^{-\tau^2/2} n \sum_{l=0}^{\lfloor n/2\rfloor} \big(\frac{n!}{2^l l!(n-2l)!}\big)^2 2^{n-2l}(n-2l+1)! \tag{7.146}$$

$$\le C(1+\varphi)^n e^{-\tau^2/2} 2^n n^3 \sup_{0\le l\le \lfloor n/2\rfloor} \frac{(n!)^2}{(l!)^2(n-2l)!} \tag{7.147}$$

$$\le C(1+\varphi)^n e^{-\tau^2/2} (2e)^n n^5 n! \tag{7.148}$$

and

$$\sum_{0\le n\le \lfloor \log\left(\frac{es^2}{d}\right)\rfloor} \frac{1}{2^n n!} \int_\tau^{+\infty} \big[(\widehat{\phi_{1+\varphi}} - \widehat{\phi_1})^{(n)}(t)\big]^2\, dt \le C\varphi^2 (1+\varphi)^{2\log\left(\frac{es^2}{d}\right)} \frac{s^2}{d} \log^5\big(\frac{es^2}{d}\big) e^{-\tau^2/2} \le \frac{C c_0}{d}, \tag{7.149}$$

if α is large enough and c_0 small enough. Furthermore, using in particular the mean-value theorem,

$$\int_\tau^{+\infty} \left[(\widehat{\phi_{1+\varphi}} - \widehat{\phi}_1)^{(n)}(t)\right]^2 dt \le 2\pi \int_{\mathbb{R}} t^{2n}(\phi_{1+\varphi} - \phi_1)^2(t)\, dt \tag{7.150}$$

$$\le C\varphi^2 \int_{\mathbb{R}} t^{2n+2} e^{-t^2/(1+\varphi)}\, dt \tag{7.151}$$

$$\le C\varphi^2 (1+\varphi)^{n+1}(n+1)!\,, \tag{7.152}$$

so that if c_0 is such that $\varphi \le 1/4$,

$$\sum_{n \ge \lfloor \log\left(\frac{es^2}{d}\right) \rfloor} \frac{1}{2^n n!} \int_\tau^\infty \left[(\widehat{\phi_{1+\varphi}} - \widehat{\phi}_1)^{(n)}(t)\right]^2 dt \le C\varphi^2 \Big(\frac{3(1+\varphi)}{4}\Big)^{\log\left(\frac{es^2}{d}\right)} \le \frac{Cc_0}{d}. \tag{7.153}$$

The result follows from the last formula, (7.140) and (7.149).

□

Part IV

Simulation of Gaussian processes under stationarity

Chapter 8

Harmonic analysis meets stationarity: A general framework for series expansions of special Gaussian processes

In this chapter, we present a new approach to derive series expansions for some Gaussian processes based on harmonic analysis of their covariance function. In particular, a new simple rate-optimal series expansion is derived for fractional Brownian motion. The convergence of the latter series holds in mean square and uniformly almost surely, with a rate-optimal decay of the remainder of the series. We also develop a general framework of convergent series expansion for certain classes of Gaussian processes. Finally, an application to functional quantization is described.

Based on Ndaoud (2018a): Ndaoud, M. (2018a). Harmonic analysis meets stationarity: A general framework for series expansions of special Gaussian processes. *arXiv preprint arXiv:1810.11850*.

8.1 Introduction

Let $B = (B_t)_{t\in\mathbb{R}^+}$ be a centered Gaussian process. B is called fractional Brownian motion (fBm) with Hurst exponent $H \in (0,1)$ if it has the following covariance structure

$$\forall t, s \in \mathbb{R}^+, \quad \mathbf{E}B_sB_t = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right).$$

Fractional Brownian motion is a self-similar process i.e. $\forall c, t > 0,\ B_{ct} \overset{\mathbf{d}}{=} c^H B_t$, with stationary increments i.e. $\forall s, t > 0,\ B_t - B_s \overset{\mathbf{d}}{=} B_{t-s}$, where $\overset{\mathbf{d}}{=}$ denotes equality in distribution. When $H = 1/2$, fractional Brownian motion coincides with the standard Brownian motion. Sample paths of fBm are almost surely Hölder-continuous of any order strictly less than H, and hence are almost surely everywhere continuous.

One of the main challenges with fBm is its simulation, as it is the case for Gaussian processes with a complex covariance structure in general. The circulant embedding method, described in Dietrich and Newsam (1997), is one of the most efficient algorithms to simulate either stationary Gaussian processes or Gaussian processes with stationary

increments on a finite interval $[0,T]$ for some $T>0$. In particular, the latter algorithm has an $N\log N$ complexity, where N is the number of time steps discretizing $[0,T]$. This complexity is to be compared with linear complexity for the standard Brownian motion due to the independence of its increments. Besides, circulant embedding does not allow local refinement.

Alternative approximation methods to simulate a Gaussian process involve its Karhunen-Loève expansion. The latter expansion is explicitly known for some processes such as the Brownian motion, the Brownian bridge Deheuvels (2007) and the Ornstein-Uhlenbeck process Corlay (2010), to name a few. Unfortunately, this expansion is not explicit for fBm.

Notation. In the rest of this paper we use the following notation. For given sequences a_n and b_n, we say that $a_n=\mathcal{O}(b_n)$ (resp $a_n=\Omega(b_n)$) when $a_n\leq cb_n$ (resp $a_n\geq cb_n$) for some absolute constant $c>0$. We write $a_n\asymp b_n$ when $a_n=\mathcal{O}(b_n)$ and $a_n=\Omega(b_n)$. For $X\in\mathbb{R}^p$, we denote by $\|X\|$ the Euclidean norm of X. For $x,y\in\mathbb{R}$, we denote by $x\vee y$ the maximum value between x and y. In particular $x\vee 0$ will be denoted by x_+. $\mathrm{Re}(z)$ denotes the real part of complex variable z. Finally, we denote by $C[0,T]$ the space of continuous functions on $[0,T]$ endowed with the sup-norm.

Related literature

In Ayache and Taqqu (2003), one of the first rate-optimal series expansion of fBm based on Wavelet series approximations is presented. For the sake of brevity, we only present , in what follows, trigonometric series, since they can be compared to the framework we expose further.

The first trigonometric series expansion for fBm on $[0,1]$ was discovered in Dzhaparidze and Van Zanten (2004). For $0<H<1$, the series $(B_t^H)_{t\in[0,1]}$ is given by

$$B_t^H=\sum_{n=1}^{\infty}\frac{\sin x_n t}{x_n}X_n+\sum_{n=1}^{\infty}\frac{1-\cos y_n t}{y_n}Y_n,\quad t\in[0,1],$$

where $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are i.i.d centered Gaussian random variables, $(x_n)_{n\geq 1}$ is the sequence of positive roots of the Bessel function J_{-H}, and $(y_n)_{n\geq 1}$ the sequence of positive roots of the Bessel function J_{1-H}. The variance of the Gaussian variables involved in the series is given by

$$\forall n\geq 1,\quad VarX_n=2c_H^2x_n^{-2H}J_{1-H}^{-2}(x_n)\quad\text{and}\quad VarY_n=2c_H^2y_n^{-2H}J_{-H}^{-2}(y_n),$$

where $c_H^2=\pi^{-1}\Gamma(1+2H)\sin(\pi H)$, and Γ is the gamma function. Dzhaparidze and Van Zanten (2004) prove rate-optimality of the above series expansion in the following sense.

Definition 8.1.1. *Let $H\in(0,1)$ and B^H a fBm with Hurst exponent H on $[0,T]$ for some $T>0$. Assume that B^H is given by the series expansion*

$$\forall t\in[0,T],\quad B_t^H=\sum_{i=0}^{\infty}Z_ie_i(t),$$

where $(Z_i)_{i\in\mathbb{N}}$ is a sequence of independent Gaussian random variables and $(e_i)_{i\in\mathbb{N}}$ a sequence of continuous deterministic functions. B^H is said to be uniformly rate-optimal if

$$\mathbf{E}\sup_{t\in[0,T]}\left|\sum_{i=N}^{\infty}Z_ie_i(t)\right|\asymp N^{-H}\sqrt{\log N}.$$

In Kühn and Linde (2002), the rate $N^{-H}\sqrt{\log N}$ is shown to be optimal. Rate-optimality also means that no other series expansion of fBm has a faster rate of convergence. We show later how rate-optimality implies uniform convergence of the series, almost surely.

Another rate-optimal trigonometric series expansion for fBm, in the case $1/2 < H < 1$, is derived in Iglói (2005), that is close to our representation. For $1/2 < H < 1$, this expansion takes the form

$$B_t = a_0 t X_0 + \sum_{k=1}^{\infty} a_k \Big(\sin(k\pi t) X_k + (1 - \cos(k\pi t)) X_{-k} \Big), \quad t \in [0,1],$$

where

$$a_0 = \sqrt{\frac{\Gamma(2-2H)}{B(H-\frac{1}{2}, \frac{3}{2}-H)(2H-1)}},$$

$$\forall k \in \mathbb{N}^*, \quad a_k = \sqrt{\frac{\Gamma(2-2H)}{B(H-\frac{1}{2}, \frac{3}{2}-H)(2H-1)} 2\,\mathrm{Re}(i \exp^{-i\pi H} \gamma(2H-1, ik\pi))} (k\pi)^{-H-\frac{1}{2}},$$

and $(X_k)_{k\in\mathbb{Z}}$ is a sequence of independent standard Gaussian random variables. The functions Γ, B, and γ are the gamma, beta, and complementary (lower) incomplete gamma functions, respectively. Even if this representation is easier to evaluate than the previous one, it still requires computation of special functions.

Main contribution

In this chapter, we give a constructive representation of fBm for all $0 < H < 1$ which is only based on harmonic analysis of its covariance function. Our approach is inspired by the Karhunen-Loève expansion. The latter expansion is obtained through an interesting application of the spectral theorem for compact normal operators, in conjunction with Mercer's theorem. We give here a sketch of its proof. Let $K_B(.,.)$ be the covariance function of the process B of interest on $[0,1]^2$. Mercer's theorem is a series representation of K_B based on the diagonalization of the following linear operator

$$\begin{array}{rccc} T_{K_B} & : & L^2[0,1] & \to & L^2[0,1] \\ & & f & \to & \int_0^1 K_B(s,.) f(s) \mathrm{d}s, \end{array}$$

where $L^2[0,1]$ is the space of square-integrable real-valued functions on $[0,1]$. In particular, it states that there is an orthonormal basis $(e_i)_{i\in\mathbb{N}}$ of $L^2[0,1]$ consisting of eigenfunctions of T_{K_B} such that the corresponding sequence of eigenvalues $(\lambda_i)_{i\in\mathbb{N}}$ is nonnegative. Moreover K_B has the following representation

$$\forall s,t \in [0,1], \quad K(s,t) = \sum_{i=0}^{\infty} \lambda_i e_i(t) e_i(s),$$

where the series converges in L^2. Considering $(Z_i)_{i\in\mathbb{N}}$ a sequence of independent standard Gaussian random variables, and assuming the uniform convergence of the following series

$$\forall t \in [0,1], \quad X_t = \sum_{i=0}^{\infty} \sqrt{\lambda_i} Z_i e_i(t),$$

one may observe that X is a centered Gaussian process on $[0,1]$ with a covariance function equal to K_B. Hence X and B have the same distribution on $[0,1]$. Since the corresponding eigenfunction sequence $(e_i)_{i\in\mathbb{N}}$ is not explicit for fBm, we follow an alternative approach.

In the case where B is a stationary process or has stationary increments, T_{K_B} becomes similar to a convolution operator. It is well known that the Fourier basis is a basis of eigenfunctions for the convolution operator, when the convolution kernel is periodic. In general, we may extend the kernel to an even periodic kernel. Since this modification applies to the covariance function K_B, there is no guarantee that the new symmetric function $\tilde{K}_B(.,.)$ is positive. In particular, the corresponding eigenvalues are not necessary positive. The last condition is crucial in our approach, since we need to take the square root of the Fourier coefficients, as described in the Karhunen-Loève proof.

One of the main contributions, is to exhibit a new class Γ of functions such that the Fourier coefficients are all negative. Let $T>0$ and γ be a real valued function on $(0,T]$. We say that γ satisfies property $(\star)$ if

- γ is continuously differentiable, increasing and concave.
- $x^{\delta}\gamma'(x) = \mathcal{O}(1)$ as $x \to 0^+$, for some $\delta \in [0,2)$.

We denote by Γ the class of such functions. As a consequence, we derive a new series expansion for fBm given by

$$B_t^H = \sqrt{c_0}tZ_0 + \sum_{k=1}^{\infty}\sqrt{\frac{-c_k}{2}}\left(Z_k \sin\frac{k\pi t}{T} + Z_{-k}\left(1-\cos\frac{k\pi t}{T}\right)\right), \quad t\in[0,T],$$

where

$$\begin{cases} c_0 := 0, & H < 1/2 \\ c_0 := HT^{2H-2}, & H > 1/2, \end{cases} \tag{8.1}$$

$$\forall k \geq 1, \quad \begin{cases} c_k := \frac{2}{T}\int_0^T t^{2H}\cos\frac{k\pi t}{T}\mathrm{d}t, & H < 1/2 \\ c_k := -\frac{4H(2H-1)T}{(k\pi)^2}\int_0^T t^{2H-2}\cos\frac{k\pi t}{T}\mathrm{d}t, & H > 1/2, \end{cases} \tag{8.2}$$

and $(Z_k)_{k\in\mathbb{Z}}$ is a sequence of independent standard Gaussian random variables. This series is rate-optimal and its convergence holds uniformly almost surely. More generally, we also derive series expansion for a general class of Gaussian processes with covariance operator linked with the class Γ.

Section 8.2 is devoted to the study of harmonic properties of the class Γ. In Section 8.3, we present our series expansion for fBm, where we prove both uniform convergence and rate-optimality. Next, we generalize this series expansion to a large class of Gaussian processes, before applying it to functional quantization.

8.2 On harmonic properties of the class Gamma

The present section is devoted to general harmonic properties of the class Γ. Let $T>0$ and $\gamma \in \Gamma$. Consider the corresponding Fourier sequence

$$\forall k \in \mathbb{N}, \quad c_k(\gamma) := \frac{2}{T}\int_0^T \gamma(t)\cos\frac{k\pi t}{T}\mathrm{d}t. \tag{8.3}$$

The next Proposition states some important properties of $c(\gamma) = (c_k(\gamma))_{k\in\mathbb{N}}$ when γ satisfies $(\star)$. In what follows, we write for brevity $c_k = c_k(\gamma)$, as long as there is no ambiguity.

Proposition 8.2.1. *Let γ be a function satisfying $(\star)$ and $c(\gamma)$ the sequence defined in (8.3), then*

- *$c(\gamma)$ is well defined.*
- $\forall k \in \mathbb{N}^*,\ \ c_k \leq 0.$
- $|c_k| = \mathcal{O}\left(\frac{1}{k^{2-\delta}}\right).$

The proof is given in Appendix 8.7. It is usually not easy to reveal the sign of the Fourier coefficients of a given function. For the class of functions satisfying $(\star)$, it turns out that the previous question can be answered, based on Proposition 8.2.1. The more general question of characterizing the class of functions with negative Fourier coefficients is beyond the scope of this paper. It is also interesting to notice that for $\gamma \in \Gamma$, the singularity around 0^+, captures the asymptotic behaviour of $c(\gamma)$ that is not trivial in general.

Inspection of the proof of Proposition 8.2.1, shows that γ has a finite limit at 0^+, if $\delta \in [0, 1)$.We will use in that case the notation $\gamma(0) := \lim_{x\to 0^+} \gamma(x)$. The next lemma gives a useful Fourier expansion for functions in Γ.

Lemma 8.2.1. *Let γ be a function satisfying $(\star)$ for some $\delta \in [0, 1)$. Then*

$$\forall t \in [-T, T], \quad \gamma(|t|) = \gamma(0) + \sum_{k=1}^{\infty} c_k \left(\cos \frac{k\pi t}{T} - 1 \right),$$

where the series converges uniformly.

Proof. Let $g : \mathbb{R} \to \mathbb{R}$ be a function such that

$$\forall t \in [-T, T], \quad g(t) = \gamma(|t|).$$

Extending g into a $2T$-periodic function, it can be defined on $\mathbb{R}$. Since g is an even function, its Fourier expansion is given by

$$\forall t \in [-T, T], \quad g(t) = \sum_{k=0}^{\infty} c_k \cos \frac{k\pi t}{T}, \tag{8.4}$$

where $c_0 = \frac{1}{T}\int_0^T \gamma(t)\mathrm{d}t$ and $c_k = \frac{2}{T}\int_0^T \gamma(t)\cos\frac{k\pi t}{T}\mathrm{d}t$, $k = 1, 2, \ldots$ Using Proposition 8.2.1, we have

$$|c_k| = \mathcal{O}\left(\frac{1}{k^{2-\delta}}\right).$$

Since $0 \leq \delta < 1$, the Fourier expansion of g is normally convergent and hence converges uniformly. Replacing t by 0 in (8.4) we get that

$$c_0 = \gamma(0) - \sum_{k=1}^{\infty} c_k.$$

It follows that

$$g(t) = \gamma(0) + \sum_{k=1}^{\infty} c_k \left(\cos \frac{k\pi t}{T} - 1 \right).$$

□

Let $\lambda := (\lambda_k)_{k\in\mathbb{N}}$ be a sequence of real numbers and $e := (e_k)_{k\in\mathbb{N}}$ a family of uniformly bounded and continuous functions on $[0,T]$. We say that (λ, e) satisfies $(\star\star)$ if

- $\exists H > 0$, such that $|\lambda_k| = \mathcal{O}\left(\frac{1}{k^{H+1/2}}\right)$.
- $\exists L > 0$, such that

$$\forall k \in \mathbb{N}, \forall s,t \in [0,T], \quad |e_k(t) - e_k(s)| \leq L\,|t-s|\,.$$

Notice that for $\gamma \in \Gamma$, and setting $e = (\cos(.), \sin(.), 1-\cos(.))$, it is easy to check that $(\sqrt{-c(\gamma)}, e)$ satisfies $(\star\star)$ for $\delta \in [0,1)$.

Theorem 8.2.1. *Let $(\lambda_k)_{k\in\mathbb{N}}$ be a sequence of real numbers, $(Z_k)_{k\in\mathbb{N}}$ a sequence of centered standard Gaussian random variables, and $(e_k)_{k\in\mathbb{N}}$ a family of continuous functions on $[0,T]$. Assume that (λ, e) satisfies $(\star\star)$ for some $H > 0$. Then the series $\sum_{k=0}^{N} \lambda_k e_k\left(\frac{k\pi .}{T}\right) Z_k$ converges almost surely in $C[0,T]$. Moreover, we have*

$$\mathbf{E} \sup_{t\in[0,T]} \left| \sum_{k=N}^{\infty} \lambda_k e_k \left(\frac{k\pi t}{T} \right) Z_k \right| = \underset{N\to\infty}{\mathcal{O}} \left(N^{-H} \sqrt{\log N} \right).$$

The proof is deferred to Appendix 8.7. In what follows, we will repeatedly use Proposition 8.2.1 along with Theorem 8.2.1. In fact, Proposition 8.2.1 describes the asymptotic behaviour of $c(\gamma)$ for $\gamma \in \Gamma$, while Theorem 8.2.1 characterizes the rate of convergence of given series expansions based on the asymptotic behaviour of $c(\gamma)$.

8.3 Constructing the fractional Brownian motion

In this section, we present our first series expansion for fBm and prove its convergence. The construction is based on harmonic decomposition of the auto-covariance function γ on $[0,T]$ such that $\gamma(t) = |t|^{2H}$ for some $0 < H < 1$. As described in the Introduction, the diagonalization of the operator T_{K_X} is not explicit for fbm. In order to benefit from the diagonalization of the convolution operator, we need to extend the auto-covariance function to a periodic function. The resulting function $\tilde{K}(.,.)$ is not guaranteed to be a covariance function. Luckily, harmonic properties of the class Γ will be useful to get around this drawback.

Since our approach does not hold for both cases, we give results separately for both fBm with $0 < H < 1/2$ and $1/2 < H < 1$, assuming that the series converge. We prove later the convergence and rate-optimality of these series.

The series expansion

The following theorem gives an explicit series expansion for fBm when $0 < H < 1/2$, assuming the series convergence.

Theorem 8.3.1. *Let $H \in \left(0, \frac{1}{2}\right)$. Consider the function γ given by $\gamma(t) = t^{2H}$, $\forall t \in [0,T]$. Denote by $c(\gamma)$ the sequence of its Fourier coefficients. Let B be a stochastic process given by the series expansion*

$$\forall t \in [0,T], \quad B_t = \sum_{k=1}^{\infty} \sqrt{-\frac{c_k}{2}} \left(Z_k \sin \frac{k\pi t}{T} + Z_{-k} \left(1 - \cos \frac{k\pi t}{T} \right) \right),$$

where $(Z_k)_{k\in\mathbb{Z}}$ denotes a sequence of independent standard Gaussian random variables. Then

$$\forall (s,t) \in [0,T]^2, \ \mathbf{E} B_s B_t = \frac{1}{2} \left(t^{2H} + s^{2H} - |t-s|^{2H} \right).$$

Proof. For $H \in \left(0, \frac{1}{2}\right)$, we denote by $\gamma(t) := |t|^{2H}$. It is easy to check that γ satisfies $(\star)$. Using Proposition 8.2.1, the above series is well-defined since $\forall k \geq 1, \ c_k \leq 0$. Because of the independence between the Gaussian random variables Z, it follows immediately that

$$\begin{aligned}
\mathbf{E} B_s B_t &= \sum_{k=1}^{\infty} -\frac{c_k}{2} \left(\sin \frac{k\pi s}{T} \sin \frac{k\pi t}{T} + \left(1 - \cos \frac{k\pi s}{T} \right) \left(1 - \cos \frac{k\pi t}{T} \right) \right) \\
&= \sum_{k=1}^{\infty} -\frac{c_k}{2} \left(1 - \cos \frac{k\pi s}{T} - \cos \frac{k\pi t}{T} + \cos \frac{k\pi (t-s)}{T} \right) \qquad (8.5) \\
&= \sum_{k=1}^{\infty} \frac{c_k}{2} \left(\left(\cos \frac{k\pi s}{T} - 1 \right) + \left(\cos \frac{k\pi t}{T} - 1 \right) - \left(\cos \frac{k\pi (t-s)}{T} - 1 \right) \right).
\end{aligned}$$

We conclude using Lemma 8.2.1. □

The previous proof does not hold for the case $1/2 < H < 1$. In fact, the Fourier coefficients $c(\gamma)$ have alternating signs in this case. For $H > 1/2$, γ does not satisfy property $(\star)$. In particular, the change in the sign of $c(\gamma)$ is partially due to the smoothness of γ' around 0. One may still notice that γ'' satisfies $(\star)$. The next Lemma, gives a link between $c(\gamma)$ and $c(\gamma'')$.

Lemma 8.3.1. *Let γ be a twice differentiable function on $(0,T)$ such that $\gamma'(0) \neq \gamma'(T)$. Define f such that*

$$\forall t \in [0,T], \quad f(t) = \gamma(t) - \frac{\gamma'(T) - \gamma'(0)}{2T} \left(t + \frac{T\gamma'(0)}{\gamma'(T) - \gamma'(0)} \right)^2,$$

then $\forall k \in \mathbb{N}^*$ we have

$$c_k(f) = \left(\frac{T}{k\pi} \right)^2 c_k(-\gamma'').$$

Proof. The function f is constructed in such a way that $f'(0) = f'(T) = 0$. For all $k \in \mathbb{N}^*$, integration by parts yields

$$\begin{aligned}\int_0^T f(t)\cos\left(\frac{k\pi t}{T}\right)\mathrm{d}t &= \frac{T}{k\pi}\left[f(t)\sin\left(\frac{k\pi t}{T}\right)\right]_0^T - \frac{T}{k\pi}\int_0^T f'(t)\sin\left(\frac{k\pi t}{T}\right)\mathrm{d}t \\ &= \left(\frac{T}{k\pi}\right)^2\left[f'(t)\cos\left(\frac{k\pi t}{T}\right)\right]_0^T - \left(\frac{T}{k\pi}\right)^2\int_0^T f''(t)\cos\left(\frac{k\pi t}{T}\right)\mathrm{d}t \\ &= -\left(\frac{T}{k\pi}\right)^2\int_0^T \gamma''(t)\cos\left(\frac{k\pi t}{T}\right)\mathrm{d}t.\end{aligned}$$

The last equality is a consequence of orthogonality between constant functions and harmonics. □

The following theorem gives an explicit series expansion for fBm when $1/2 < H < 1$, assuming the series convergence.

Theorem 8.3.2. *Let $H \in \left(\frac{1}{2}, 1\right)$. Consider the function γ given by $\gamma(t) = -2H(2H-1)t^{2H-2}$, $\forall t \in [0,T]$. Denote by $c(\gamma)$ the sequence of its Fourier coefficients. Let B be a stochastic process given by the series expansion*

$$\forall t \in [0,T], \quad B_t = \sqrt{HT^{2H-2}}tZ_0 + \sum_{k=1}^{\infty}\frac{T}{k\pi}\sqrt{-\frac{c_k}{2}}\left(\sin\frac{k\pi t}{T}Z_k + \left(1-\cos\frac{k\pi t}{T}\right)Z_{-k}\right),$$

where $(Z_k)_{k\in\mathbb{Z}}$ denotes a sequence of independent standard Gaussian random variables. Then

$$\forall (s,t) \in [0,T]^2, \ \mathbf{E}B_sB_t = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right).$$

Proof. By considering $\gamma(t) = -2H(2H-1)t^{2H-2}$, we notice that γ satisfies $(\star)$ for $1/2 < H < 1$. Moreover

$$|\gamma'(t)| = \mathcal{O}\left(\frac{1}{t^{3-2H}}\right),$$

as $t \to 0^+$. Since $1 < 3-2H < 2$, we get using Proposition 8.2.1 that $c_k \leq 0$ for all $k = 1, 2, \ldots$, and $|c_k| = \mathcal{O}\left(\frac{1}{k^{2H-1}}\right)$. We also obtain, using Lemma 8.3.1 that

$$\frac{2}{T}\int_0^T \left(t^{2H} - HT^{2H-2}t^2\right)\cos\left(\frac{k\pi t}{T}\right)\mathrm{d}t = \left(\frac{T}{k\pi}\right)^2 c_k.$$

Since $\frac{|c_k|}{k^2} = \mathcal{O}\left(\frac{1}{k^{2H+1}}\right)$, the Fourier series converges uniformly and we can apply Lemma 8.2.1 to get

$$\forall t \in [-T,T], \quad |t|^{2H} = HT^{2H-2}t^2 + \sum_{k=1}^{\infty}\left(\frac{T}{k\pi}\right)^2 c_k\left(\cos\frac{k\pi t}{T} - 1\right).$$

Since $c_k \leq 0$ the series expansion is well defined and we have

$$\begin{aligned}
\mathbf{E}B_tB_s &= HT^{2H-2}ts - \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{T}{k\pi}\right)^2 c_k\left(\sin\frac{k\pi t}{T}\sin\frac{k\pi s}{T} + \left(1-\cos\frac{k\pi t}{T}\right)\left(1-\cos\frac{k\pi s}{T}\right)\right)\\
&= HT^{2H-2}ts - \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{T}{k\pi}\right)^2 c_k\left(1-\cos\frac{k\pi t}{T}-\cos\frac{k\pi s}{T}+\cos\frac{k\pi(t-s)}{T}\right)\\
&= \frac{1}{2}\left(HT^{2H-2}(t^2+s^2-(t-s)^2) + \sum_{k=1}^{\infty}\left(\frac{T}{k\pi}\right)^2 c_k\left(\cos\frac{k\pi t}{T}+\cos\frac{k\pi s}{T}-\cos\frac{k\pi(t-s)}{T}-1\right)\right)\\
&= \frac{|t|^{2H}+|s|^{2H}-|t-s|^{2H}}{2}.
\end{aligned}$$

□

Convergence and rate-optimality

After giving a new explicit representation of fBm, we prove its convergence in both mean square sense and almost surely. We also show its uniform rate-optimality. For the rest of the section, we denote more precisely by $(c_k)_{k\geq 0}$ the following sequence

$$\begin{cases} c_0 := 0, & 0 < H < 1/2,\\ c_0 := HT^{2H-2}, & 1/2 < H < 1, \end{cases} \tag{8.6}$$

and for $k = 1, 2, \ldots,$

$$\begin{cases} c_k := \frac{2}{T}\int_0^T t^{2H}\cos\frac{k\pi t}{T}\mathrm{d}t, & 0 < H < 1/2,\\ c_k := -\frac{4H(2H-1)T}{(k\pi)^2}\int_0^T t^{2H-2}\cos\frac{k\pi t}{T}\mathrm{d}t, & 1/2 < H < 1. \end{cases} \tag{8.7}$$

One may first notice, based on previous results, that $|c_k| = \mathcal{O}\left(\frac{1}{k^{2H+1}}\right)$. We will now consider the series expansion constructed in this section and given by

$$\forall t \in [0,T],\ B_t = \sqrt{c_0}tZ_0 + \sum_{k=1}^{\infty}\sqrt{-\frac{c_k}{2}}\left(Z_k\sin\frac{k\pi t}{T} + Z_{-k}\left(1-\cos\frac{k\pi t}{T}\right)\right). \tag{8.8}$$

The following theorems show the convergence of the series in mean square and almost surely uniformly. Let $N \in \mathbb{N}^*$. In what follows, we denote by B^N the truncated series of B that is given by

$$B_t^N = \sqrt{c_0}tZ_0 + \sum_{k=1}^{N}\sqrt{-\frac{c_k}{2}}\left(Z_k\sin\frac{k\pi t}{T} + Z_{-k}\left(1-\cos\frac{k\pi t}{T}\right)\right). \tag{8.9}$$

Theorem 8.3.3. *Let B_t and B_t^N be defined by (8.6)-(8.9), and $H \in (0,1)\backslash\{1/2\}$. Then,*

$$\sup_{t\in[0,T]}\sqrt{\mathbf{E}\left(B_t - B_t^N\right)^2} = \mathcal{O}\left(N^{-H}\right).$$

Proof. As we did previously, we will use the independence of Gaussian random variables $(Z_k)_{k\in\mathbb{Z}}$. It is straightforward to see that, for all $t \in [0,T]$,

$$\mathbf{E}\big(B_t - B_t^N\big)^2 = \sum_{k>N} -\frac{c_k}{2}\left((\sin\frac{k\pi t}{T})^2 + (1-\cos\frac{k\pi t}{T})^2\right) = \sum_{k>N} -c_k\left(1-\cos\frac{k\pi t}{T}\right).$$

Hence

$$\sup_{t\in[0,T]}\sqrt{\mathbf{E}\big(B_t - B_t^N\big)^2} \leq \sqrt{\sum_{k>N}|c_k|}.$$

Since $|c_k| = \mathcal{O}\left(\frac{1}{k^{2H+1}}\right)$, the result follows. □

The previous theorem shows the mean square convergence of B^N. Since B^N is a centered Gaussian process, and using the fact that the Gaussian Hilbert space is complete, we deduce that B is a centered Gaussian process with the same covariance as fBm. It follows that B is a fractional Brownian motion on $[0,T]$. We turn now to the question of rate-optimality of the series expansion.

Theorem 8.3.4. *Let B be the series expansion defined in (8.8). Almost surely, B_t^N converges uniformly, and its rate of convergence is given by*

$$\mathbf{E}\sup_{t\in[0,T]}\left|B_t - B_t^N\right| \asymp N^{-H}\sqrt{\log(N)}.$$

Proof. We will only need to prove that the rate of convergence of the above series is faster than $N^{-H}\sqrt{\log(N)}$ since the latter is the optimal rate of convergence for fBm as shown in Kühn and Linde (2002). By truncating the series, we have

$$\forall t\in[0,T],\quad B_t - B_t^N = \sum_{k=N+1}^{\infty}\sqrt{-\frac{c_k}{2}}\left(Z_k\sin\frac{k\pi t}{T} + Z_{-k}\left(1-\cos\frac{k\pi t}{T}\right)\right).$$

Since $\sqrt{-\frac{c_k}{2}} = \underset{k\to\infty}{\mathcal{O}}\left(\frac{1}{k^{H+1/2}}\right)$, and using the fact that $t\to\sin(t)$ and $t\to 1-\cos(t)$ are 1-Lipschitz functions we can directly use Theorem 8.2.1 to conclude the proof. □

Finally, we have derived a new series expansion for fBm. Theorem 8.3.4 shows that this series is moreover rate-optimal.

8.4 Generalization to special Gaussian processes

In this section, we develop a general framework for series expansion of special classes of Gaussian processes. For all these classes, we prove almost sure uniform convergence and give the corresponding rate of convergence. The question of rate-optimality of the presented series expansions is beyond the scope of this paper. We refer the reader to Proposition 4 in Luschgy and Pagès (2009) that gives some hints on rate-optimality.

The next theorem generalizes the case of fBm. We derive a series expansion for a class of Gaussian processes with stationary increments. Let γ be a function satisfying $(\star)$ for some $\delta\in[0,1)$, and let X be a centered Gaussian process. We say that X is a Gaussian process of type (A), if it is characterized by the following covariance structure

$$\forall t,s\in[0,T],\quad \mathbf{E}X_tX_s = \frac{1}{2}\left(\gamma(t)+\gamma(s)-\gamma(|t-s|)\right).$$

Theorem 8.4.1. *Assume that γ satisfies $(\star)$ for some $\delta \in [0,1)$, and let $c(\gamma)$ be the sequence of its Fourier coefficients. Let $(Z_k)_{k\in\mathbb{Z}}$ be a sequence of independent standard Gaussian random variables. Then, the series expansion*

$$X_t = \sum_{k=1}^{\infty} \sqrt{\frac{-c_k}{2}} \left(Z_k \sin \frac{k\pi t}{T} + Z_{-k} \left(1 - \cos \frac{k\pi t}{T} \right) \right), \quad t \in [0,T],$$

converges uniformly in $[0,T]$ almost surely and X_t is a Gaussian process of type (A). Moreover its rate of convergence is given by

$$\mathbf{E} \sup_{t\in[0,T]} \left| X_t - X_t^N \right| = \mathcal{O}\left(N^{-\frac{1-\delta}{2}} \sqrt{\log(N)} \right),$$

where X_t^N is the truncated series of X_t.

Proof. Applying Proposition 8.2.1, we obtain that $c_k \leq 0$ for $k = 1, 2, \ldots$. Hence, the series is well defined. Moreover, we also have that $|c_k| = \mathcal{O}\left(\frac{1}{k^{2-\delta}}\right)$. Since $\delta \in [0,1)$ we can use Theorem 8.2.1 to get that

$$\mathbf{E} \sup_{t\in[0,T]} \left| \sum_{k=N+1}^{\infty} \sqrt{\frac{-c_k}{2}} \left(Z_k \sin \frac{k\pi t}{T} + Z_{-k} \left(1 - \cos \frac{k\pi t}{T} \right) \right) \right| = \mathcal{O}\left(\frac{\sqrt{\log(N)}}{N^{\frac{1-\delta}{2}}} \right),$$

and that the series converges almost surely and uniformly in $[0,T]$. It follows that X_t is a centered Gaussian process. Its covariance function is

$$\forall s,t \in [0,T], \quad \mathbf{E} X_s X_t = \sum_{k=1}^{\infty} \frac{-c_k}{2} \left(1 - \cos \frac{k\pi t}{T} - \cos \frac{k\pi s}{T} + \cos \frac{k\pi (t-s)}{T} \right).$$

We can conclude using Lemma 8.2.1 that

$$\forall s,t \in [0,T], \quad \mathbf{E} X_s X_t = \frac{1}{2} \left(\gamma(t) + \gamma(s) - \gamma(|t-s|) \right).$$

Hence X_t is a Gaussian process of type (A). □

The next class of interest is a subclass of stationary Gaussian processes. Let γ be a function such that $-\gamma$ satisfies $(\star)$ for some $\delta \in [0,1)$, and let X be a centered Gaussian process. We say that X is a Gaussian process of type (B), if it is characterized by the following covariance structure

$$\forall t,s \in [0,T], \quad \mathbf{E} X_t X_s = \gamma(|t-s|).$$

Theorem 8.4.2. *Assume that γ is such that $-\gamma$ satisfies $(\star)$ for some $\delta \in [0,1)$, and let $c(\gamma)$ be the sequence of its Fourier coefficients. Let $(Z_k)_{k\in\mathbb{Z}}$ be a sequence of independent standard Gaussian random variables. If $c_0 \geq 0$, then the series*

$$X_t = \sqrt{c_0} Z_0 + \sum_{k=1}^{\infty} \sqrt{c_k} \left(\sin \frac{k\pi t}{T} Z_k + \cos \frac{k\pi t}{T} Z_{-k} \right), \quad t \in [0,T],$$

converges uniformly in $[0,T]$ almost surely, and X_t is a Gaussian process of type (B). Moreover the convergence is rate-optimal, and its rate is given by

$$\mathbf{E} \sup_{t\in[0,T]} \left| X_t - X_t^N \right| = \underset{N\to\infty}{\mathcal{O}} \left(N^{-\frac{1-\delta}{2}} \sqrt{\log(N)} \right),$$

where X_t^N is the truncated series of X_t.

Proof. Using the same steps as for Theorem 8.4.1, we get that the series is well defined, and that it converges uniformly almost surely. Hence X_t is a Gaussian process. Moreover we have

$$\forall s,t \in [0,T], \quad \mathbf{E}X_sX_t = \sum_{k=0}^{\infty} c_k \cos\frac{k\pi(t-s)}{T} = \gamma(|t-s|).$$

It follows that X_t is a Gaussian process of type (B). Since the basis $(e_k)_{k\in\mathbb{Z}}$ used in this expansion is orthogonal, we may apply the same argument as in Proposition 4 in Luschgy and Pagès (2009), and deduce that the convergence is rate-optimal. □

One immediate consequence is a series expansion for a stationary fractional Ornstein-Uhlenbeck process X_t with $0 < H < 1/2$, where a stationary fOU is a centered Gaussian process such that

$$\forall s,t \in [0,T], \quad \mathbf{E}X_sX_t = e^{-|t-s|^{2H}}.$$

The last series expansion was already derived in Luschgy and Pagès (2009).

The framework we are proposing here can also be applied to Gaussian processes that are neither stationary nor with stationary increments. As an example we apply it to another class of Gaussian processes. Let γ be a function defined on $(0,2T)$, such that $-\gamma$ satisfies $(\star)$ for some $\delta \in [0,1)$, and let X_t be a centered Gaussian process. We say that X_t is a Gaussian process of type (C), if it is characterized by the following covariance structure

$$\forall t,s \in [0,T],\ \mathbf{E}X_sX_t = \frac{1}{2}\Big(\gamma(|t-s|) - \gamma(t+s)\Big). \tag{8.10}$$

Theorem 8.4.3. *Assume that γ is such that $-\gamma$ satisfies $(\star)$ on $(0,2T)$ for some $\delta \in [0,1)$, and let $c(\gamma)$ be the sequence of its Fourier coefficients on the interval $(0,2T)$. Let $(Z_k)_{k\in\mathbb{N}^*}$ be a sequence of independent standard Gaussian random variables. Then, the series*

$$X_t = \sum_{k=1}^{\infty} \sqrt{c_k}\sin\frac{k\pi t}{2T} Z_k, \quad t \in [0,T],$$

converges uniformly in $[0,T]$ almost surely, and X_t is a Gaussian process of type (C). Moreover its rate of convergence is given by

$$\mathbf{E}\sup_{t\in[0,T]} \left|X_t - X_t^N\right| = \underset{N\to\infty}{\mathcal{O}}\left(N^{-\frac{1-\delta}{2}}\sqrt{\log(N)}\right),$$

where X_t^N is the truncated series of X_t.

Proof. Using the same steps as for Theorem 8.4.1, we get that the series is well defined, and that it converges uniformly almost surely. Hence X_t is a Gaussian process. Moreover we have

$$\forall t,s \in [0,T],\ \mathbf{E}X_sX_t = \sum_{k=1}^{\infty} c_k\left(\sin\frac{k\pi t}{2T}\sin\frac{k\pi s}{2T}\right) = \frac{1}{2}\sum_{k=1}^{\infty} c_k\left(\cos\frac{k\pi(t-s)}{2T} - \cos\frac{k\pi(t+s)}{2T}\right).$$

As a consequence we get that

$$\forall t,s \in [0,T],\ \mathbf{E}X_sX_t = \frac{1}{2}\Big(\gamma(|t-s|) - \gamma(t+s)\Big).$$

It follows that X_t is a Gaussian process of type (C). □

Remark 8.4.1. *One may notice that, in Theorem 8.4.3, we have considered a $4T$-periodic basis instead of $2T$-periodic because $\forall 0 < t, s < T, \quad 0 \leq s + t \leq 2T$.*

In order to illustrate Theorem 8.4.3, we give below two examples of corresponding series expansions.

Example 8.4.1. *Karhunen-Loève expansion of the Brownian motion.*
In this example, we consider the function $\gamma(t) = -|t|$ on $[0, T]$. It is easy to check that γ satisfies the conditions of Theorem 8.4.3. One may also notice that this process is a Brownian motion on $[0, T]$ since

$$\forall t, s \in [0, T], \quad \frac{1}{2}\left(-|t - s| + |t + s|\right) = \min(t, s).$$

An explicit evaluation of the sequence $(c_k)_{k\in\mathbb{N}^*}$ gives

$$\begin{aligned} \forall k \in \mathbb{N}^*, \quad c_k &= \frac{1}{T}\int_0^{2T} -t\cos\frac{k\pi t}{2T}\mathrm{d}t \\ &= \frac{2}{k\pi}\int_0^{2T} \sin\frac{k\pi t}{2T}\mathrm{d}t \\ &= \left(1 - (-1)^k\right)\left(\frac{2}{k\pi}\right)^2 T. \end{aligned} \tag{8.11}$$

Applying Theorem 8.4.3, it follows that

$$\forall t \in [0, T], \quad X_t = \sqrt{2}\sum_{k=1}^{\infty} \frac{\sqrt{T}}{(k - \frac{1}{2})\pi} Z_k \sin\frac{(k - \frac{1}{2})\pi t}{T},$$

is a series expansion for Brownian motion on $[0, T]$, where $(Z_k)_{k\in\mathbb{N}^*}$ is a sequence of independent standard Gaussian random variables.

Example 8.4.2. *A new series expansion for the generalized Ornstein-Uhlenbeck process.*
In this example we consider the non-stationary Ornstein-Uhlenbeck process $(Y_t)_{t\geq 0}$ where Y_0 is a Gaussian random variable with the following distribution $\mathcal{N}(\mu, \sigma_0^2)$. This process is a Gaussian process characterized by

$$\forall t \geq 0, \quad \mathbf{E}Y_t = \mu e^{-\theta t} + \alpha(1 - e^{-\theta t}),$$

and

$$\forall s, t \geq 0, \quad \mathbf{E}\left(Y_s - \mathbf{E}Y_s, Y_t - \mathbf{E}Y_t\right) = \sigma_0^2 e^{-\theta(t+s)} + \frac{\sigma^2}{2\theta}\left(e^{-\theta(|t-s|)} - e^{-\theta(t+s)}\right),$$

for some $\theta > 0$ and $\alpha, \sigma \in \mathbb{R}$. By setting $\gamma(t) = \frac{\sigma^2}{\theta}e^{-\theta t}$, we have that $-\gamma$ satisfies conditions of Theorem 8.4.3. Hence, and applying Theorem 8.4.3, the following expansion

$$\forall t \in [0, T], \quad X_t = Y_0 e^{-\theta t} + \alpha(1 - e^{-\theta t}) + \sum_{k=1}^{\infty} \sqrt{c_k} Z_k \sin\frac{k\pi t}{2T}, \tag{8.12}$$

is a series expansion of the generalized Ornstein-Uhlenbeck process on $[0, T]$, where $(Z_k)_{k\geq 1}$ is a sequence of independent standard Gaussian random variables, that are also independent from Y_0.

An explicit evaluation of the sequence $(c_k)_{k\in\mathbb{N}^*}$ *gives*

$$\begin{aligned}\forall k \geq 1, \quad \frac{\theta}{\sigma^2}c_k &= \frac{1}{T}\int_0^{2T} e^{-\theta t}\cos\frac{k\pi t}{2T} = Re\left(\frac{1}{T}\int_0^{2T} e^{(-\theta+i\frac{k\pi}{2T})t}\mathrm{d}t\right)\\ &= Re\left(\frac{1-(-1)^k e^{-2\theta T}}{\theta T - \frac{ik\pi}{2}}\right) = \frac{1}{1+\left(\frac{k\pi}{2\theta T}\right)^2}\frac{1-(-1)^k e^{-2\theta T}}{\theta T}.\end{aligned} \tag{8.13}$$

The expansion (8.12) *is easier to use compared to the one known so far that includes the zeros of Bessel functions.*

8.5 Application: *Functional quantization*

Quantization consists in approximating a random variable taking a continuum of values in $\mathbb{R}$ by a discrete random variable. While vector quantization deals with finite dimensional random variables, functional quantization extends the concept to the infinite dimensional setting, as it is the case for stochastic processes. Quantization of random vectors can be considered as a discretization of the probability space, providing in some sense the best approximation to the original distribution. The quantization of a random variable X taking values in $\mathbb{R}$ consists in approximating it by the best discrete random variable Y taking finite values in $\mathbb{R}$. If we set N to be the maximum number of values taken by Y, the problem is equivalent to minimizing the following error

$$\xi_N(X) = \left\{\mathbf{E}\left(X - \mathrm{Proj}_\Gamma(X)\right)^2, \quad \Gamma \subset \mathbb{R} \text{ such that } |\Gamma| \leq N\right\}. \tag{8.14}$$

A solution of (8.14) is an L^2-optimal quantizer of X.

For a multidimensional Gaussian random variable X optimal quantization is expensive. One way to mitigate this cost, is to consider product-quantization, that is to use a cartesian product of one-dimensional optimal-quantizers of each marginale as in Printems (2005). The resulting quantizer is stationnary when marginals of X are independent. In Luschgy and Pagès (2007), it is shown that Karhunen-Loève product-quantization, while it is sub-optimal, remains rate-optimal in the case of Gaussian processes.

We consider now a continuous Gaussian process $(X_t)_{t\in[0,T]}$ such that $\int_0^T \mathbf{E}|X_t^2|\mathrm{d}t < \infty$, and its expansion

$$\forall t \in [0,T], \quad X_t = \sum_{i=0}^{\infty} \lambda_i e_i(t) Z_i,$$

where $(\lambda_i)_{i\in\mathbb{N}}$ is a sequence of real numbers such that $\sum_{i=0}^{\infty}\lambda_i^2 < \infty$, $(e_i)_{i\in\mathbb{N}}$ is an orthonormal sequence of continuous functions, and $(Z_i)_{i\in\mathbb{N}}$ a sequence of independent standard Gaussian random variables. Notice that the Karhunen-Loève expansion is a special case of what we are introducing. In this case the error induced by replacing the process by a rate-optimal quantizer of its truncation up to order m is given by

$$\xi_N(X)^2 = \int_0^T \mathbf{E}\left(X_t - \sum_{i=0}^{m}\lambda_i e_i(t) Y_i\right)^2 \mathrm{d}t,$$

where $\forall\ 0 \leq i \leq m$, Y_i is an optimal quantizer of Z_i taking N_i values and $\prod_{i=0}^m N_i \leq N$. More precisely we get that

$$\xi_N(X)^2 = \sum_{i=m+1}^{\infty}\lambda_i^2 + \sum_{i=0}^{m}\xi_{N_i}(\mathcal{N}(0,\lambda_i^2)).$$

If moreover $\lambda_N^2 \asymp \frac{1}{N^\delta}$, with $1 < \delta < 3$, it is shown in Luschgy and Pagès (2002) that, the optimal product-quantization of level N is achieved when the dimension of the quantizer m is of order $\log N$ and that it satisfies

$$\xi_N(X) \asymp (\log N)^{\frac{1-\delta}{2}}.$$

When the basis $(e_k)_{k\in\mathbb{N}}$ chosen in the series expansion is not orthonormal, an alternative rate-optimal quantization method is presented in Junglen and Luschgy (2010). The idea consists in truncating the series up to the optimal order $m \asymp \log N$ and to consider the finite-dimensional covariance operator K^m of the truncation given by

$$\forall t, s \in [0,T], \quad K^m(t,s) = \sum_{i=0}^{m} \lambda_i^2 e_i(t) e_i(s).$$

More specifically, consider H a linear subspace of L^2 defined by $H = \mathrm{Span}((e_i)_{0\le i\le m})$. The operator T_{K^m} is given by

$$T_{K^m} : \begin{cases} H \to H \\ f \to \int_0^T K^m(s,.) f(s) \mathrm{d}s. \end{cases}$$

T_{K^m} is clearly an endomorphism. Unlike the Karhunen-Loève theorem, in this case we deal with a linear and symmetric operator in finite dimension. Hence there exists $(\mu_i^m)_{0\le i\le m}$ a sequence of positive real numbers and $(f_i^m)_{0\le i\le m}$ an orthonormal basis of H such that

$$\forall t, s \in [0,T], \quad K^m(t,s) = \sum_{i=0}^{m} \mu_i^m f_i^m(t) f_i^m(s).$$

We can then assert that there exists $(Y_i^m)_{0\le i\le m}$ a sequence of independent standard Gaussian random variables such that

$$\forall t \in [0,T], \quad \sum_{i=0}^{m} \lambda_i e_i(t) Z_i = \sum_{i=0}^{m} \sqrt{\mu_i^m} f_i^m(t) Y_i^m \quad a.s.$$

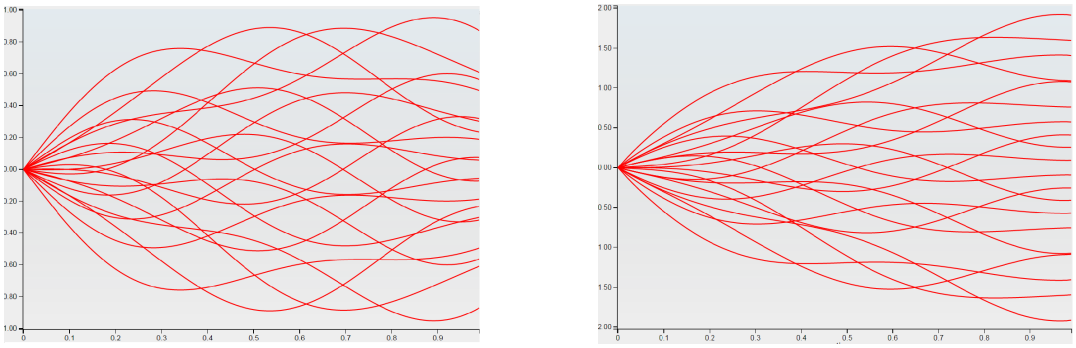
Following the same argument as in Junglen and Luschgy (2010), if we set $m \asymp \log N$ and replace the process by a rate-optimal quantizer of $\sum_{i=0}^{m} \sqrt{\mu_i^m} f_i^m(t) Y_i^m$, we will get the quadratic quantization error

$$\xi_N(X)^2 = \underset{N\to\infty}{\mathcal{O}} \left(\sum_{i=m+1}^{\infty} \lambda_i^2 + \sum_{i=0}^{m} \xi_{N_i}(\mathcal{N}(0,\mu_i^2)) \right).$$

Moreover, this error is rate-optimal. As an illustration, we give a rate-optimal quantization for both fBm and generalized Ornstein-Uhlenbeck with $T = 1$ and $N = 20$.

8.6 Conclusion

This paper presents a new framework to derive series expansions for a specific class of Gaussian processes based on harmonic analysis. One of the main results is a new, simple



(a) generalized OU $\theta = 2, \alpha = 0$ (b) fBm $H = 0.4$

Figure 8.1: Product quantization of a centered Ornstein-Uhlenbeck process, starting from $Y_0 = 0$ (left), and a fBm (right).

and rate-optimal series expansion for fractional Brownian motion. One of the advantages of this expansion is that its coefficients are easily computed which can reduce the complexity of simulation, especially for the case $H < 1/2$ where no other trigonometric series expansion is known. Our approach is general and gives series expansions for a large class of Gaussian processes, in particular to the generalized Ornstein-Uhlenbeck process. The application to quantization is interesting in particular for fBm, where the series basis is not orthonormal. In this case, we show how to deal with non-orthonormality and construct a rate-optimal quantizer.

8.7 Appendix: Proofs

Proof of Proposition 8.2.1. We first prove that γ is integrable. The function γ' is continuous on $(0,T]$, positive and for $\delta \in (0,2)$ we have that there exists $M>0$ and $\epsilon>0$ such that

$$\forall x \in (0,\epsilon), \quad 0 \leq \gamma'(x) \leq \frac{M}{x^\delta}. \tag{8.15}$$

By integrating (8.15), we get

$$\gamma(x) = \mathcal{O}\left(1 + x^{1-\delta}\right),$$

as $x \to 0^+$. The last result holds also for $\delta = 0$. Since $0 \leq \delta < 2$ and γ is continuous, it comes out that γ is integrable on (0,T]. It follows that $c(\gamma)$ is well defined.

Before showing the second part, one may first notice that γ' is positive and decreasing since γ is concave and increasing. By a change of variable in (8.3), we get

$$\forall k \in \mathbb{N}^*, \quad c_k = \frac{2}{T}\frac{T}{k\pi}\int_0^{k\pi} \gamma\left(\frac{Tu}{k\pi}\right)\cos(u)\mathrm{d}u.$$

Since

$$\gamma(u)\sin(u) = \mathcal{O}\left(\sin(u) + u^{1-\delta}\sin(u)\right),$$

as $u \to 0^+$, then

$$\lim_{u\to 0^+} \gamma(u)\sin(u) = 0.$$

Using integration by parts we obtain, for all $k \in \mathbb{N}^*$,

$$\begin{aligned} c_k &= -\frac{2}{T}\left(\frac{T}{k\pi}\right)^2 \int_0^{k\pi} \gamma'\left(\frac{Tu}{k\pi}\right)\sin(u)\mathrm{d}u \\ &= -\frac{2}{T}\left(\frac{T}{k\pi}\right)^2 \sum_{n=0}^{k-1}(-1)^n \int_0^{\pi} \gamma'\left(\frac{T(u+n\pi)}{k\pi}\right)\sin(u)\mathrm{d}u. \end{aligned} \tag{8.16}$$

For $0 \leq n < k$, we define

$$v_{k,n} := \int_0^{\pi} \gamma'\left(\frac{T(u+n\pi)}{k\pi}\right)\sin(u)\mathrm{d}u.$$

It is immediate that, $\forall k \in \mathbb{N}^*$, $(v_{k,n})_{n<k}$ is nonnegative and decreasing with respect to n. Regrouping each pair of elements in (8.16), we get

$$c_k = -\frac{2}{T}\left(\frac{T}{k\pi}\right)^2 \left(\sum_{n=0}^{\lfloor \frac{k}{2} \rfloor - 1}\left(v_{k,2n} - v_{k,2n+1}\right) + \frac{1-(-1)^k}{2}v_{k,k-1}\right).$$

It follows that $c_k \leq 0, \forall k \in \mathbb{N}^*$. For the last point, we use again the second part of (8.16) and get

$$c_k = -\frac{2}{T}\left(\frac{T}{k\pi}\right)^2 \sum_{n=0}^{k-1}(-1)^n v_{k,n}.$$

Since the sequence $\left((-1)^n v_{k,n}\right)_{n<k}$ has alternating signs and a decreasing modulus, it turns out that

$$|c_k| \leq \frac{2T^2}{k^2\pi^2} v_{k,0} \leq \frac{2T^2}{k^2\pi^2} \int_0^\pi \gamma'\left(\frac{Tu}{k\pi}\right) \sin(u) \mathrm{d}u.$$

In order to conclude, it is enough to prove that

$$\int_0^\pi \gamma'\left(\frac{Tu}{k\pi}\right) \sin(u) \mathrm{d}u = \mathcal{O}(k^\delta).$$

Since $\gamma \in \Gamma$, we can check that $x \to x^\delta \gamma'(x)$ is uniformly bounded on $[0, T]$. Let M be this uniform bound, then

$$0 \leq \int_0^\pi \gamma'\left(\frac{Tu}{k\pi}\right) \sin(u) \mathrm{d}u \leq M \left(\frac{k\pi}{T}\right)^\delta \int_0^\pi \frac{\sin(u)}{u^\delta} \mathrm{d}u.$$

As δ belongs to $[0, 2)$, it follows that

$$|c_k| = \mathcal{O}(k^{\delta-2}).$$

This concludes the proof.

□

Proof of Theorem 8.2.1. We will use the following standard bound on the maximum of centered Gaussian random variables. If $X_1, \ldots, X_M$ are centered Gaussian random variables, then there exists $c > 0$, such that

$$\mathbf{E} \max_{1 \leq i \leq M} |X_i| \leq c\sqrt{\log M} \max_{1 \leq i \leq M} \sqrt{\mathbf{E} X_i^2}. \tag{8.17}$$

We denote by $v_k(t) := \lambda_k e_k(\frac{k\pi t}{T}) Z_k$ for $k \in \mathbb{N}$ and $t \in [0, T]$. The proof is divided in two parts. We first show that, for some $A > 0$, we have

$$\forall n \in \mathbb{N}^*, \quad \mathbf{E} \sup_{t \in [0,T]} \left| \sum_{k=2^n}^{2^{n+1}-1} v_k(t) \right| \leq A\sqrt{n} 2^{-nH}. \tag{8.18}$$

Let $N \in \mathbb{N}$, for all $0 \leq j \leq N-1$, we denote by $I_j = \left[j\frac{T}{N}, (j+1)\frac{T}{N}\right]$ and t_j the corresponding center i.e. $t_j = (j + 1/2)\frac{T}{N}$. Let $n \in \mathbb{N}^*$, we have

$$\begin{aligned}
\mathbf{E} \sup_{t \in [0,T]} \left| \sum_{k=2^n}^{2^{n+1}-1} v_k(t) \right| &= \mathbf{E} \sup_{0 \leq j < N} \sup_{t \in I_j} \left| \sum_{k=2^n}^{2^{n+1}-1} v_k(t) \right| \\
&\leq \mathbf{E} \sup_{0 \leq j < N} \left| \sum_{k=2^n}^{2^{n+1}-1} v_k(t_j) \right| + \mathbf{E} \sup_{0 \leq j < N} \sup_{t \in I_j} \left| \sum_{k=2^n}^{2^{n+1}-1} (v_k(t) - v_k(t_j)) \right|.
\end{aligned} \tag{8.19}$$

Using (8.17) we get

$$\begin{aligned}
\mathbf{E} \sup_{0 \leq j < N} \left| \sum_{k=2^n}^{2^{n+1}-1} v_k(t_j) \right| &\leq c\sqrt{\log N} \sup_{0 \leq j < N} \sqrt{\mathbf{E} \left| \sum_{k=2^n}^{2^{n+1}-1} v_k(t_j) \right|^2} \\
&\leq c\sqrt{\log N} \sup_{0 \leq j < N} \sqrt{\sum_{k=2^n}^{2^{n+1}-1} \mathbf{E} v_k(t_j)^2} \\
&\leq C'\sqrt{\log N} 2^{-nH},
\end{aligned} \tag{8.20}$$

for some $C' > 0$. The last inequality comes from the fact that $\mathbf{E}v_k(t_j)^2 \leq \lambda_k^2 ||e_k||_\infty^2 \leq \frac{C}{k^{1+2H}}$, for some $C > 0$.
For the second part of (8.19), we have

$$\mathbf{E} \sup_{0\leq j<N} \sup_{t\in I_j} \left| \sum_{k=2^n}^{2^{n+1}-1} (v_k(t) - v_k(t_j)) \right| \leq \mathbf{E} \sup_{0\leq j<N} \sum_{k=2^n}^{2^{n+1}-1} \sup_{t\in I_j} |v_k(t) - v_k(t_j)| \,. \tag{8.21}$$

Observing that $\forall t \in I_j$ we have $|t - t_j| \leq \frac{T}{N}$, we get that

$$\begin{aligned} \sup_{t\in I_j} |v_k(t) - v_k(t_j)| &\leq |\lambda_k||Z_k| \left| e_k\left(\frac{k\pi t}{T}\right) - e_k\left(\frac{k\pi t_j}{T}\right) \right| \\ &\leq C' k^{\frac{1}{2}-H} |Z_k| \frac{\pi}{N}. \end{aligned} \tag{8.22}$$

Replacing in (8.21), it follows that

$$\begin{aligned} \mathbf{E} \sup_{0\leq j<N} \sup_{t\in I_j} \left| \sum_{k=2^n}^{2^{n+1}-1} (v_k(t) - v_k(t_j)) \right| &\leq \frac{C''}{N} \sum_{k=2^n}^{2^{n+1}-1} k^{\frac{1}{2}-H} \\ &\leq \frac{C^*}{N} 2^{n(\frac{3}{2}-H)}. \end{aligned} \tag{8.23}$$

Combining (8.20) and (8.23) we deduce that

$$\mathbf{E} \sup_{t\in[0,T]} \left| \sum_{k=2^n}^{2^{n+1}-1} v_k(t) \right| \leq C\sqrt{\log N} 2^{-nH} + \frac{C^*}{N} 2^{n(\frac{3}{2}-H)}. \tag{8.24}$$

Replacing $N = 2^{2n}$ in (8.24), we prove (8.18).

The previous result holds even if we replace $\left|\sum_{k=2^n}^{2^{n+1}-1} v_k(t)\right|$ by $\left|\sum_{k=M}^{2^{n+1}-1} v_k(t)\right|$ for some $M \in [2^n, 2^{n+1} - 1]$. Let N be a positive integer. By taking $m = \lfloor \log N / \log 2 \rfloor$, we get

$$\mathbf{E} \sup_{t\in[0,T]} \left| \sum_{k=N+1}^{\infty} v_k(t) \right| \leq \mathbf{E} \sup_{t\in[0,T]} \left| \sum_{k=N+1}^{2^{m+1}-1} v_k(t) \right| + \sum_{i=m+1}^{\infty} \mathbf{E} \sup_{t\in[0,T]} \left| \sum_{k=2^i}^{2^{i+1}-1} v_k(t) \right|. \tag{8.25}$$

We can conclude, using (8.18), that

$$\mathbf{E} \sup_{t\in[0,T]} \left| \sum_{k=N+1}^{\infty} v_k(t) \right| \leq A \sum_{i=m}^{\infty} \sqrt{i} 2^{-iH} \leq A' \sqrt{m} 2^{-mH}. \tag{8.26}$$

It suffices to observe that $2^m \leq N < 2^{m+1}$, to obtain the rate of the uniform convergence. The uniform tightness implies that $\sum_{k=0}^{N} v_k$ has a weak limit in C[0,T] the space of continuous functions on [0,T]. We remind the reader that we endow this space with the supremum metric. By the Itô-Nisio theorem, we get, as in Itô et al. (1968), that the process $\sum_{k=0}^{N} v_k$ converges in $C[0,T]$ almost surely. □

Conclusion

In this thesis, we focused on some problems specific to high-dimensional statistics and their applications to machine learning.

Our main contribution is with respect to the problem of variable selection in high-dimensional linear regression. We derive non-asymptotic bounds for the minimax risk of support recovery under expected Hamming loss in the Gaussian mean model in $\mathbf{R}^d$ for classes of s-sparse vectors separated from 0 by a constant $a > 0$. In some cases, we also find explicitly the corresponding minimax selectors and their adaptive variants. As corollaries, we characterize precisely an asymptotic sharp phase transition for both almost full and exact recovery.

As for the problem of exact support recovery in Compressed Sensing, we propose an algorithm for exact support recovery in the setting of noisy compressed sensing where all entries of the design matrix are i.i.d standard Gaussian. This algorithm is the first polynomial time procedure to achieve the same conditions of exact recovery as the exhaustive search (maximal likelihood) decoder that were studied in Rad (2011) and Wainwright (2009a). Our procedure has an advantage over the exhaustive search of being adaptive to all parameters of the problem, robust and computable in polynomial time.

Motivated by the interplay between estimation and support recovery, we introduce a new notion of scaled minimaxity for sparse estimation in high-dimensional linear regression model. We present more optimistic lower bounds than the one given by the classical minimax theory and hence improve on existing results. We recover the sharp result of Donoho et al. (1992) for global minimaxity as a consequence of our study. Fixing the scale of the signal-to-noise ratio, we prove that estimation error can be much smaller than the global minimax error. Among other findings, we show that exact support recovery is not necessary to achieve the optimal scaled minimax rate.

As for the problem of clustering in the two components Gaussian Mixture Model, we provide a precise non-asymptotic characterization of the minimax Hamming risk. As a consequence we recover the sharp phase transition for exact recovery in this model. Namely, the phase transition occurs around the threshold $\Delta = \bar{\Delta}_n$ such that

$$\bar{\Delta}_n^2 = \sigma^2 \left(1 + \sqrt{1 + \frac{2p}{n \log n}}\right) \log n.$$

Our procedure achieves the previous threshold. It is a variant of Lloyd's iterations initialized by a spectral method. This procedure is fully adaptive, rate optimal and computationally simple. It turns out that our procedure is, to the best of our knowledge, the first fast method to achieve optimal exact recovery.

Another main contribution is devoted to some effects of adaptivity under sparsity, where adaptivity is either with respect to the noise level or its nominal law. We derive

the minimax optimal rates and exhibit minimax estimators for estimation of the noise variance σ^2 for different classes of noise. For instance, when the noise distribution is exactly known, then σ^2 can be estimated faster if the noise have known polynomial tails rather than belongs to the class of sub-Gaussian noise. Similar results are derived for the problem of minimax estimation of $\|\boldsymbol{\theta}\|_2$. Finally, we study the minimax optimality of estimation of $\boldsymbol{\theta}$ when the noise belongs to a class of distributions with polynomial tails or exponential tails. We derive the minimax rates for these settings. An unexpected conclusion is that in the sparse mean model, the optimal rates are much slower and depend on the polynomial index of the noise as opposed to the rates in regression with "well spread" regressors.

In our last contribution, we propose a new approach to derive series expansions for some Gaussian processes based on harmonic analysis of their covariance function. In particular, a new simple rate-optimal series expansion is derived for fractional Brownian motion. The convergence of the latter series holds in mean square and uniformly almost surely, with a rate-optimal decay of the remainder term of the series. We also develop a general framework of convergent series expansion for certain classes of Gaussian processes.

Bibliography

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*.

Abbe, E., Bandeira, A. S., and Hall, G. (2014). Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*.

Abramovich, F. and Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In *Wavelets and statistics*, pages 5–14. Springer.

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653.

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation.

Aeron, S., Saligrama, V., and Zhao, M. (2010). Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130.

Aksoylar, C., Atia, G. K., and Saligrama, V. (2017). Sparse signal processing with linear and nonlinear observations: A unified Shannon-theoretic approach. *IEEE Transactions on Information Theory*, 63(2):749–776.

Alon, N., Matias, Y., and Szegedy, M. (1999). The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147.

Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1):1983–2001.

Awasthi, P. and Sheffet, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer.

Ayache, A. and Taqqu, M. S. (2003). Rate optimality of wavelet series approximations of fractional brownian motion. *Journal of Fourier Analysis and Applications*, 9(5):451–471.

Banks, J., Moore, C., Vershynin, R., Verzelen, N., and Xu, J. (2018). Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Transactions on Information Theory*.

Bellec, P. C. (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. *arXiv preprint arXiv:1804.01230*.

Bellec, P. C., Lecué, G., and Tsybakov, A. B. (2018). Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642.

Belloni, A., Chernozhukov, V., and Wang, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788.

Belomestny, D., Trabs, M., and Tsybakov, A. B. (2017). Sparse covariance matrix estimation in high-dimensional deconvolution. *arXiv preprint arXiv:1710.10870*.

Benaych-Georges, F. and Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521.

Benaych-Georges, F. and Nadakuditi, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135.

Berthet, Q. and Rigollet, P. (2013). Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope - adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103.

Broadbent, S. R. and Hammersley, J. M. (1957). Percolation processes: I. crystals and mazes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 53, pages 629–641. Cambridge University Press.

Butucea, C., Ingster, Y. I., and Suslina, I. A. (2015). Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *ESAIM: Probability and Statistics*, 19:115–134.

Butucea, C., Matias, C., et al. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, 11(2):309–340.

Butucea, C., Ndaoud, M., Stepanova, N. A., and Tsybakov, A. B. (2018). Variable selection with hamming loss. *The Annals of Statistics*, 46(5):1837–1875.

Butucea, C. and Stepanova, N. (2017). Adaptive variable selection in nonparametric sparse additive models. *Electronic Journal of Statistics*, 11(1):2321–2357.

Cai, T. T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145.

Cai, T. T. and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Carpentier, A., Collier, O., Comminges, L., Tsybakov, A. B., and Wang, Y. (2018). Minimax rate of testing in sparse linear regression. *arXiv preprint arXiv:1804.06494*.

Carpentier, A. and Verzelen, N. (2019). Adaptive estimation of the sparsity in the gaussian vector model. *The Annals of Statistics*, 47(1):93–126.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185.

Cavalier, L., Golubev, G., Picard, D., and Tsybakov, A. (2002). Oracle inequalities for inverse problems. *Annals of Statistics*, 30(3):843–874.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019). Fast mean estimation with sub-gaussian rates. *arXiv preprint arXiv:1902.01998*.

Collier, O., Comminges, L., and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958.

Collier, O., Comminges, L., Tsybakov, A. B., and Verzelen, N. (2018). Optimal adaptive estimation of linear functionals under sparsity. *The Annals of Statistics*, 46(6A):3130–3150.

Collier, O. and Dalalyan, A. S. (2017). Minimax estimation of a p-dimensional linear functional in sparse gaussian models and robust estimation of the mean. *arXiv preprint arXiv:1712.05495*.

Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. (2018). Adaptive robust estimation in sparse vector model. *arXiv preprint arXiv:1802.04230v3*.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696.

Corlay, S. (2010). Functional quantization-based stratified sampling methods. *arXiv preprint arXiv:1008.4441*.

Dasgupta, S. and Schulman, L. (2007). A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. (2005). A direct formulation for sparse pca using semidefinite programming. In *Advances in neural information processing systems*, pages 41–48.

Deheuvels, P. (2007). A karhunen–loève expansion for a mean-centered brownian bridge. *Statistics & probability letters*, 77(12):1190–1200.

Derumigny, A. (2018). Improved bounds for square-root lasso and square-root slope. *Electronic Journal of Statistics*, 12(1):741–766.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664. IEEE.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017). Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org.

Dietrich, C. and Newsam, G. N. (1997). Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107.

Ding, L., Yurtsever, A., Cevher, V., Tropp, J. A., and Udell, M. (2019). An optimal-storage approach to semidefinite programming using approximate complementarity. *arXiv preprint arXiv:1902.03373*.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67.

Dzhaparidze, K. and Van Zanten, H. (2004). A series expansion of fractional brownian motion. *Probability theory and related fields*, 130(1):39–55.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fei, Y. and Chen, Y. (2018). Hidden integrality of sdp relaxation for sub-gaussian mixture models. *arXiv preprint arXiv:1803.06510*.

Feldman, V., Perkins, W., and Vempala, S. (2015). Subsampled power iteration: a unified algorithm for block models and planted csp's. In *Advances in Neural Information Processing Systems*, pages 2836–2844.

Feng, L. and Zhang, C.-H. (2017). Sorted concave penalized regression. *arXiv preprint arXiv:1712.09941*.

Féral, D. and Péché, S. (2007). The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in mathematical physics*, 272(1):185–228.

Fletcher, A. K., Rangan, S., and Goyal, V. K. (2009). Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772.

Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959.

Foucart, S. and Lecué, G. (2017). An iht algorithm for sparse recovery from subexponential measurements. *IEEE Signal Processing Letters*, 24(9):1280–1283.

Gamarnik, D. and Zadik, I. (2017). Sparse high-dimensional linear regression. Algorithmic barriers and a local search algorithm. *arXiv preprint arXiv:1711.04952*.

Gao, C., Lu, Y., and Zhou, D. (2016). Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning*, pages 603–611.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.

Gao, Z. and Stoev, S. (2018). Fundamental limits of exact support recovery in high dimensions. *arXiv preprint arXiv:1811.05124*.

Gautier, E. and Tsybakov, A. B. (2013). Pivotal estimation in high-dimensional regression via linear programming. In *Empirical inference*, pages 195–204. Springer.

Genovese, C. R., Jin, J., Wasserman, L., and Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, 13(Jun):2107–2143.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Giraud, C. and Verzelen, N. (2018). Partial recovery bounds for clustering with the relaxed k means. *arXiv preprint arXiv:1807.07547*.

Golubev, Y. and Krymova, E. (2017). On estimation of the noise variance in high-dimensional linear models. *arXiv preprint arXiv:1711.09208*.

Guo, Z., Wang, W., Cai, T. T., and Li, H. (2018). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association*, pages 1–12.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732.

Harris, T. E. (1960). A lower bound for the critical probability in a certain percolation process. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 56, pages 13–20. Cambridge University Press.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hopkins, S. B. (2018). Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Huber, P. J. (2011). *Robust statistics*. Springer.

Ibragimov, I. A. and Has'Minskii, R. Z. (2013). *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media.

Iglói, E. (2005). A rate-optimal trigonometric series expansion of the fractional brownian motion. *Electron. J. Probab*, 10:1381–1397.

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351.

Ingster, Y. and Stepanova, N. (2014). Adaptive variable selection in nonparametric sparse regression. *Journal of Mathematical Sciences*, 199(2):184–201.

Itô, K., Nisio, M., et al. (1968). On the convergence of sums of independent banach space valued random variables. *Osaka Journal of Mathematics*, 5(1):35–48.

Janson, L., Barber, R. F., and Candes, E. (2017). Eigenprism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1037–1065.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622.

Jerrum, M. R., Valiant, L. G., and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188.

Ji, P. and Jin, J. (2012). Ups delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, 40(1):73–103.

Jin, J., Zhang, C.-H., and Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *Journal of Machine Learning Research*, 15(1):2723–2772.

Joseph, A. (2013). Variable selection in high-dimension with random designs and orthogonal matching pursuit. *Journal of Machine Learning Research*, 14(1):1771–1800.

Junglen, S. and Luschgy, H. (2010). A constructive sharp approach to functional quantization of stochastic processes. *Journal of Applied Mathematics*, 2010.

Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer.

Ke, T., Jin, J., and Fan, J. (2014). Covariance assisted screening and estimation. *Annals of statistics*, 42(6):2202.

Kesten, H. (1980). The critical probability of bond percolation on the square lattice equals 1/2. *Communications in mathematical physics*, 74(1):41–59.

Klusowski, J. M. and Brinda, W. (2016). Statistical guarantees for estimating the centers of a two-component gaussian mixture by em. *arXiv preprint arXiv:1608.02280*.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Kühn, T. and Linde, W. (2002). Optimal series representation of fractional brownian sheets. *Bernoulli*, pages 669–696.

Kumar, A. and Kannan, R. (2010). Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63.

Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Liu, H. and Barber, R. F. (2018). Between hard and soft thresholding: optimal iterative thresholding algorithms. *arXiv preprint arXiv:1804.08841*.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2:90–102.

Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. (2011). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39(5):2164–2204.

Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

Luschgy, H. and Pagès, G. (2002). Functional quantization of gaussian processes. *Journal of Functional Analysis*, 196(2):486–531.

Luschgy, H. and Pagès, G. (2007). High-resolution product quantization for gaussian processes under sup-norm distortion. *Bernoulli*, 13(3):653–671.

Luschgy, H. and Pagès, G. (2009). Expansions for gaussian processes and parseval frames. *Electron. J. Probab*, 14(42):1198–1221.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Mixon, D. G., Villar, S., and Ward, R. (2016). Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*.

Moitra, A. and Valiant, G. (2010). Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE.

Mossel, E., Neeman, J., and Sly, A. (2015). Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 69–75. ACM.

Ndaoud, M. (2018a). Harmonic analysis meets stationarity: A general framework for series expansions of special gaussian processes. *arXiv preprint arXiv:1810.11850*.

Ndaoud, M. (2018b). Sharp optimal recovery in the two component gaussian mixture model. *arXiv preprint arXiv:1812.08078*.

Ndaoud, M. (2019). Interplay of minimax estimation and minimax support recovery under sparsity. *ALT 2019*.

Ndaoud, M. and Tsybakov, A. B. (2018). Optimal variable selection and adaptive noisy compressed sensing. *arXiv preprint arXiv:1809.03145*.

Nemirovskii, A. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York.

Neuvial, P. and Roquain, E. (2012). On false discovery rate thresholding for classification under sparsity. *The Annals of Statistics*, 40(5):2572–2600.

Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. (2018). Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory*. Clarendon Press, Oxford.

Printems, J. (2005). Functional quantization for numerics with an application to option pricing. *Monte Carlo Methods and Applications mcma*, 11(4):407–446.

Rad, K. R. (2011). Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Transactions on Information Theory*, 57(7):4672–4679.

Royer, M. (2017). Adaptive clustering through semidefinite programming. In *Advances in Neural Information Processing Systems*, pages 1795–1803.

Saligrama, V. and Zhao, M. (2011). Thresholded basis pursuit: LP algorithm for order-wise optimal support recovery for sparse and approximately sparse signals from noisy random measurements. *IEEE Transactions on Information Theory*, 57(3):1567–1586.

Sanjeev, A. and Kannan, R. (2001). Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM.

Shen, J. and Li, P. (2017). A tight bound of hard thresholding. *The Journal of Machine Learning Research*, 18(1):7650–7691.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*, volume 59. Siam.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Su, W. and Candes, E. (2016). Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Szabo, B. and van Zanten, H. (2017). An asymptotic analysis of distributed nonparametric methods. *arXiv preprint arXiv:1711.03149*.

Tropp, J. A. and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematician*, volume 2, pages 523–531.

Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *In: CompressedSensing. Cambridge Univ. Press, Cambridge*, pages 210–268.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90.

Verzelen, N. and Gassiat, E. (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710.

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the r th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Wainwright, M. J. (2009a). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741.

Wainwright, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wang, S., Weng, H., and Maleki, A. (2017). Which bridge estimator is optimal for variable selection? *arXiv preprint arXiv:1705.08617*.

Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Transactions on Information Theory*, 56(6):2967–2979.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Wegkamp, M. (2003). Model selection in nonparametric regression. *Annals of Statistics*, 31(1):252–273.

Wei, X. and Minsker, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, pages 2859–2868.

Wigner, E. (1958). On the distribution of the roots of certain symmetric matrices. *The Annals of Mathematics*, 67:325–328.

Wu, Z. and Zhou, H. H. (2013). Model selection and sharp asymptotic minimaxity. *Probability Theory and Related Fields*, 156(1-2):165–191.

Xia, D. and Zhou, F. (2017). The sup-norm perturbation of hosvd and low rank tensor denoising. *arXiv preprint arXiv:1707.01207*.

Zhang, A., Cai, T. T., and Wu, Y. (2018). Heteroskedastic pca: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*.

Zhang, A. Y. and Zhou, H. H. (2016). Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144.

Zhang, T. (2011a). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE transactions on information theory*, 57(7):4689–4708.

Zhang, T. (2011b). Sparse recovery with orthogonal matching pursuit under rip. *IEEE Transactions on Information Theory*, 57(9):6215–6221.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

List of Figures

List of Tables

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Quelques contributions à la sélection de variables, au clustering et à l'estimation statistique en grande dimension

Mots clés : sélection de variable, grande dimension, clustering, regression linéaire, transition de phase.

Résumé : Cette thèse traite les problèmes statistiques suivants : la sélection de variables dans le modèle de régression linéaire en grande dimension, le clustering dans le modèle de mélange Gaussien, l'estimation statistique adaptative sous l'hypothèse de parcimonie ainsi que la simulation des processus Gaussiens.
Sous l'hypothèse de parcimonie, la sélection de variables correspond au recouvrement du petit ensemble de variables significatives. Nous étudions les propriétés non-asymptotiques de la sélection de variables dans la régression linéaire en grande dimension, et nous caractérisons les conditions nécessaires et suffisantes pour la sélection de variables dans ce modèle. Nous étudions également le problème de l'adaptation dans le modèle de bruit additif parcimonieux et nous obtenons les taux d'estimation optimaux du vecteur objectif, sa norme euclidienne lorsque le niveau de bruit et/ou sa loi nominale sont inconnus, ainsi que de la variance du bruit.
Le clustering est une tâche d'apprentissage statistique non supervisée visant à regrouper des observations proches les unes des autres dans un certain sens. Nous étudions le problème de clustering dans le modèle de mélange Gaussien à deux composantes, et caractérisons (à la constante précise) la séparation minimale entre les groupes afin de les recouvrir de façon exacte. Nous fournissons également une procédure computationnelle rapide permettant un clustering optimal.
Les processus Gaussiens sont extrêmement utiles dans la pratique, par exemple lorsqu'il s'agit de modéliser des phénomènes du monde réel. Néanmoins, leur simulation n'est pas facile, en général. Nous proposons et étudions un nouveau développement en série à taux optimal pour simuler une grande classe de processus Gaussiens.

Title : Contributions to variable selection, clustering and statistical estimation in high dimension

Keywords : variable selection, high dimension, clustering, linear regression, phase transition.

Abstract : This PhD thesis deals with the following statistical problems : Variable selection in high-dimensional linear regression, clustering in Gaussian Mixture Model, adaptive estimation under sparsity, simulation of Gaussian processes.
Under the sparsity assumption, variable selection corresponds to recovering a small set of significant variables. We study non-asymptotic properties of variable selection in sparse high-dimensional linear regression, and we establish necessary and sufficient conditions of recovery in this model. We also explore adaptation issue in the sparse mean model and we find optimal rates of estimation of the target vector and of its ℓ_2 norm when the variance and/or the distribution of noise are unknown, as well as optimal rates of variance estimation.
Clustering is a non-supervised machine learning task aiming to group observations that are close to each other in some sense. We study the problem of clustering in Gaussian Mixture Model with two components, and we find (with sharp rate and constant) the minimal separation distance between clusters allowing for exact recovery. We also propose a computationally fast procedure achieving this optimal clustering.
Gaussian processes are very useful to model real world phenomena. Nevertheless, their simulation is, in general, not easy. We propose and study a new rate-optimal series expansion to simulate a large class of Gaussian processes.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France