Proceedings of Machine Learning Research vol 98:1–22, 2019 30th International Conference on Algorithmic Learning Theory

Interplay of minimax estimation and minimax support recovery under sparsity

Mohamed Ndaoud MOHAMED.NDAOUD@ENSAE.FR
Department of Statistics
CREST (UMR CNRS 9194), ENSAE
5, av. Henry Le Chatelier, 91764 Palaiseau, FRANCE

Editors: Aurélien Garivier and Satyen Kale

Abstract

In this paper, 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.

Keywords: High-dimensional estimation under sparsity, SLOPE estimator, Hamming loss, exact support recovery, non-asymptotic minimax risk, adaptive estimation.

1. Introduction

1.1. Statement of the problem

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

$$Y = \beta + \sigma\xi \tag{1}$$

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

$$Y = X\beta + \sigma\xi \tag{2}$$

where $X \in \mathbb{R}^{n\times p}$ is a given design or sensing matrix, and the noise is independent of X. Model (1) corresponds to the orthogonal design. In this paper, we mostly focus on model (1). We denote by $\mathbf{P}_\beta$ the distribution of Y in model (1) or of (Y, X) in model (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 $s \le p$, 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 $a \geq 0$, 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 (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 paper, 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 (1) if

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|^2 \right) = \sigma^2 s(1 + o(1)) \quad \text{as } \frac{s}{p} \to 0.$$

We also say that estimator β achieves *exact support recovery* in model (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 paper, 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 to Appendix.

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 $a_n \asymp b_n$ 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 $q \geq 1$, 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 $q \geq 1$ 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 $q \geq 1$.

1.2. 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 (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 (1). In model (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.

1.3. Main contribution

Inspired by the related literature, the present work is also motivated by the following questions.

- In model (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 paper.
- 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 paper is a sharp study of the minimax risk ϕ. What is more, we study a more general quantity given by

$$\inf_{\hat\beta}\sup_{\beta\in\Omega^p_{s,a}} \mathbf{E}_\beta\left(\|\hat\beta-\beta\|_q^q\right), \tag{3}$$

for any $q\geq 1$. We give lower bounds and corresponding upper bounds for (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.

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 1 *For any $s < p$, $q \geq 1$ and any probability measure π on $\mathbb{R}^p$, there exists $C_q > 0$ such that*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Theta_{s,a}} \mathbf{E}_\beta \|\hat{\beta} - \beta\|_q^q \geq \inf_{\hat{T} \in \mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |\hat{T}_j(Y) - \beta_j|^q - C_q \, \mathbb{E}_\pi \left[\left(\mathbf{E}(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta_{s,a}) \right], \tag{4}$$

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 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 (2), where X is a deterministic design. The following lemma is useful to get more precise lower bounds in model (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 2 *Assume that Y satisfies model (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 3 *Assume that Y follows model (2) with a deterministic design X. For any $a > 0$, $q \geq 1$ and $s < p$ we have*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq \sigma_q^q \max_{|S|=s} \sum_{i \in S} \frac{1}{\|X_i\|_2^q}.$$

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 4 *Assume that Y follows model (2) with deterministic design X. For any $a > 0$, $q \geq 1$ and $s < p$ we have*

$$\forall s' \in (0, s), \quad \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq a^q \frac{s'}{s} \left(\frac{1}{2^q} \Psi(p, s, a, \sigma, X) - 2se^{-\frac{(s-s')^2}{2s}} \right).$$

The proof is based on arguments similar to Butucea et al. (2018). Assume now that we are under model (1) and set

$$\psi(p, s, a, \sigma) := (p - s)\, \mathbf{P}(\sigma\varepsilon > t(a)) + s\mathbf{P}(\sigma\varepsilon > a - t(a)),$$

where ε is a standard Gaussian random variable and

$$t(a) := \frac{a}{2} + \frac{\sigma^2 \log\left(\frac{p}{s} - 1\right)}{a}. \tag{5}$$

The minimax Hamming loss for model (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 (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

$$t^* = \sigma\sqrt{2\log\frac{p}{s} - 1}. \tag{6}$$

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 *Assume that Y follows model (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 6 *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, it turns out that the minimax rate $\sigma^2 s \log(p/s)$, 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.

3. Optimal scaled minimax estimators

In this section, we consider upper bounds for the scaled minimax risk under model (1). For $a > 0$ define the following estimator:

$$\hat{\beta}^a_j := Y_j \mathbf{1}_{\{|Y_j| \geq t(a \vee t^*)\}}, \quad \forall j \in 1, \ldots, p, \tag{7}$$

where $t(.)$ and t^* are defined respectively in (5) and (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

$$\psi_+(p, s, a, \sigma) := (p - s)\, \mathbf{P}(\sigma\varepsilon > t(a)) + s\mathbf{P}(\sigma\varepsilon > (a - t(a)_+)),$$

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 (7).

Theorem 7 *Assume that Y follows model (1). For all $a > 0$, let $\hat{\beta}^a$ be the estimator (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, we deduce the next corollary.

Corollary 8 *Assume that Y follows model (1). For all $a > 0$, let $\hat{\beta}^a$ be the estimator (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 9 *Assume that Y follows model (1). For all $a > 0$, let $\hat{\beta}^a$ be the estimator (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 10 *Assume that Y follows model (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

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \frac{\mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q\right)}{s\sigma_q^q} = 1 + o(1) \quad \textit{if } a \geq a_q(1).$$

Inspecting the proof of Corollary 10, 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 10 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.

4. Adaptative scaled minimax estimators

In Section 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 11

- *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 $\sigma\sqrt{2\log\left(p/s\right)}$. Thus, it is very difficult to be adaptive to a in the transition regime. For $s \leq p/4$, define the following estimator:

$$\hat{\beta}_j^s := Y_j \mathbf{1}_{\{|Y_j| \geq t_s^*\}}, \quad \forall j \in 1, \ldots, p, \tag{8}$$

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 12 *Assume that Y follows model (1). Let $\hat{\beta}^s$ be the estimator (8). For all $q \geq 1$, $a > 0$ and $s < p/4$ we have*

$$\sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta}^s - \beta\|_q^q \right) \leq C_q \Psi_{ad}(a),$$

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. Conclusion:

In this paper, 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.

Acknowledgments

We would like to thank Alexandre Tsybakov for valuable comments on early versions of this manuscript. We would like to thank the reviewers for their insightful comments on the paper.

Appendix A. Sharp Gaussian tail bounds

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

$$\frac{e^{-y^2/2}}{\sqrt{2\pi}y+4} \leq \frac{1}{\sqrt{2\pi}}\int_y^\infty e^{-u^2/2}du \leq \frac{e^{-y^2/2}}{\sqrt{2\pi}y \vee 2}.$$

for all $y \geq 0$. These bounds are an immediate consequence of formula 7.1.13 in Abramowitz and Stegun (1965) with $x = y/\sqrt{2}$.

Appendix B. Proof of Theorem 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

$$\begin{aligned}\inf_{\hat\beta}\sup_{\beta\in A}\mathbf{E}_\beta|\hat\beta-\beta|_q^q &\geq \inf_{\hat\beta}\mathbb{E}_{\pi_A}\mathbf{E}_\beta|\hat\beta-\beta|_q^q = \inf_{\hat\beta}\mathbb{E}_{\pi_A}\mathbf{E}_\beta|\hat\beta-\beta^A|_q^q \\ &\geq \sum_{j=1}^p \inf_{\hat T_j}\mathbb{E}_{\pi_A}\mathbf{E}_\beta|\hat T_j-\beta_j^A|^q\end{aligned}$$

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

$$\mathbb{E}^A\big(|B_j^A(Y)-\beta_j^A|^q\big|Y\big) \leq \mathbb{E}^A\big(|\hat T_j(Y)-\beta_j^A|^q\big|Y\big) \tag{9}$$

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,

$$\inf_{\hat\beta}\sup_{\beta\in A}\mathbf{E}_\beta|\hat\beta-\beta|_q^q \geq \mathbb{E}_{\pi_A}\mathbf{E}_\beta\sum_{j=1}^p|B_j^A-\beta_j^A|^q. \tag{10}$$

Using this, we obtain

$$\begin{aligned}
\inf_{\hat{T}\in\mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}-\beta|_q^q &\le \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |B_j^A - \beta_j|^q \\
&= \mathbb{E}_\pi \mathbf{E}_\beta \Big(\sum_{j=1}^p |B_j^A - \beta_j|^q \mathbf{1}(\beta \in A) \Big) + \mathbb{E}_\pi \mathbf{E}_\beta \Big(\sum_{j=1}^p |B_j^A - \beta_j|^q \mathbf{1}(\beta \notin A) \Big) \\
&\le \mathbb{E}_{\pi_A} \mathbf{E}_\beta \sum_{j=1}^p |B_j^A - \beta_j^A|^q + \mathbb{E}_\pi \mathbf{E}_\beta \Big(\sum_{j=1}^p |B_j^A - \beta_j|^q \mathbf{1}(\beta \notin A) \Big) \\
&\le \mathbb{E}_{\pi_A} \mathbf{E}_\beta \sum_{j=1}^p |B_j^A - \beta_j^A|^q + \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p 2^{q-1} (|B_j^A|^q + |\beta_j|^q) \mathbf{1}(\beta \notin A).
\end{aligned} \tag{11}$$

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 (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 (10) and (11) yields (4).

Appendix C. Proof of Lemma 2

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

$$\forall a \in \mathbb{R}^p, \forall j = 1,\dots,p \quad \inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y) - \beta_j|^q = \inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y-a) - \beta_j|^q. \tag{12}$$

Using conditioning, one may also notice that

$$\inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y) - \beta_j|^q \ge \mathbb{E}_{\pi_{-j}} \left(\inf_{\hat{T}_j\in\mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X,Y) - \beta_j|^q \Big| \beta_{-j} \right), \tag{13}$$

where β_{-j} represents the vector β deprived of β_j and π_{-j} the corresponding prior. Hence, we get from (12) and (13) that

$$\inf_{\hat{T}_j \in \mathbb{R}} \mathbb{E}_\pi \mathbf{E}_\beta |\hat{T}_j(X,Y) - \beta_j|^q \geq \mathbb{E}_{\pi_{-j}} \left(\inf_{\hat{T}_j \in \mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X, \tilde{Y}_j) - \beta_j|^q \Big| \beta_{-j} \right),$$

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.

Appendix D. Proof of Theorem 3

We apply Theorem 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

$$\begin{aligned} \inf_{\hat{\beta}} \sup_{\beta \in \beta_{s,a}} \mathbf{E}_\beta |\hat{\beta} - \beta|^q \geq & \inf_{\hat{T} \in \mathbb{R}^p} \mathbb{E}_\pi \mathbf{E}_\beta \sum_{j=1}^p |\hat{T}_j(X,Y) - \beta_j|^q \\ & - C_q \, \mathbb{E}_\pi \left[\left(\mathbf{E}(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Omega_{s,a}) \right]. \end{aligned}$$

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 2 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 $\nu \to \infty$.

Appendix E. Proof of Theorem 4

We are going to mimic the previous proof using a different prior. We apply Theorem 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

$$\mathbb{E}_\pi \left[\left(\mathbf{E}^A(\|\beta^A\|_q^q | Y) + \|\beta\|_q^q \right) \mathbf{1}(\beta \notin \Theta_{s,a}) \right] \leq 2 a^q s' e^{-\frac{(s-s')^2}{2s}}.$$

Going back to the first term, we get the following lower bound using Lemma 2

$$\inf_{\hat{T}_j \in \mathbb{R}} \mathbb{E}_{\pi_j} \mathbf{E}_\beta |\hat{T}_j(X, \tilde{Y}_j) - \beta_j|^q = a^q \inf_{\hat{T}_j \in \mathbb{R}} \left(\frac{s'}{p} \mathbf{E}_a |\hat{T}_j(X, \tilde{Y}_j) - 1|^q + (1 - \frac{s'}{p}) \mathbf{E}_0 |\hat{T}_j(X, \tilde{Y}_j)|^q \right)$$

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.

Appendix F. Proof of Proposition 5

Combining Theorem 3 and Theorem 4 with $s' = s/2$, we get

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq s\sigma_q^q \vee \left(\frac{a^q}{2^{q+1}} \psi(p, s, a, \sigma) - sa^q e^{-s/8} \right).$$

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

  $$a^q \psi(p, s, a, \sigma) \leq Csa^q e^{-a^2/32\sigma^2} \leq C_q s.$$

  This shows that the term $\sigma_q^q s$ is dominating. As a result

  $$s\sigma_q^q \vee \left(\frac{a^q}{2^{q+1}} \psi(p, s, a, \sigma) - sa^q e^{-s/8} \right) \asymp s,$$

  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

  $$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq s\sigma_q^q \vee \left(\frac{a^q}{2^{q+1}} \psi(p, s, a, \sigma) - sC_{q'} \right).$$

  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

  $$\begin{aligned} \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq & \inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,t^*}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \\ \geq & \frac{1}{2^{q+1}} st^{*q} \mathbf{P}\left(\sigma\varepsilon \geq 0\right) - C_{q'} s \geq C_{q''} st^{*q}. \end{aligned}$$

  Hence

  $$\inf_{\hat{\beta}} \sup_{\beta \in \Omega^p_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q \right) \geq C_{q''} \sigma^q s \log\left(\frac{p}{s} - 1 \right)^{q/2}.$$

Appendix G. Proof of Theorem 7

Let β be a vector in $\Omega^p_{s,a}$, we have

$$\|\beta^a - \beta\|_q^q = \sum_{i\in S} \mathbf{E}\left|\hat{\beta}^a_i - \beta_i\right|^q + \sum_{i\in S^c} \mathbf{E}\left|\hat{\beta}^a_i - \beta_i\right|^q.$$

On S^c, we have

$$\hat{\beta}^a_i - \beta_i = \xi_i \mathbf{1}_{\{|\xi_i|>t(a\vee t^*)\}}.$$

Hence we get that

$$\mathbf{E}\left|\hat{\beta}^a_i - \beta_i\right|^q = \mathbf{E}\left(|\xi_i|^q \mathbf{1}_{\{|\xi_i|>t(a\vee t^*)\}}\right).$$

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

$$\begin{aligned}\mathbf{E}\left|\hat{\beta}^a_i - \beta_i\right|^q &= \mathbf{E}\left(|\xi_i|^q \mathbf{1}_{\{|\xi_i|>t(a\vee t^*)\}}\right)\\ &\leq C_q(t(a\vee t^*)^q + \sigma^q)\mathbf{P}\left(|\xi_i| \geq t(a\vee t^*)\right).\end{aligned}$$

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

$$\hat{\beta}^a_i - \beta_i = Y_i \mathbf{1}_{\{|Y_i|>t(a)\}} - \beta_i = -\xi_i - Y_i \mathbf{1}_{\{|Y_i|\leq t(a)\}}.$$

Hence and since $|x+y|^q \leq 2^{q-1}(|x|^q + |y|^q)$ we get

$$\begin{aligned}\mathbf{E}\left|\hat{\beta}^a_i - \beta_i\right|^q &\leq 2^{q-1}\sigma_q^q + 2^{q-1}\mathbf{E}\left(|Y_i|^q \mathbf{1}_{\{|Y_i|\leq t(a\vee t^*)\}}\right)\\ &\leq 2^{q-1}\sigma_q^q + 2^{q-1}t(a\vee t^*)^q\mathbf{P}\left(|Y_i| \leq t(a\vee t^*)\right)\\ &\leq 2^{q-1}\sigma_q^q + 2^{q-1}t(a\vee t^*)^q\mathbf{P}\left(|\xi_i| \geq (a - t(a\vee t^*))_+\right).\end{aligned}$$

We get that on S we have

$$\mathbf{E}\sum_{i\in S}|\hat{\beta}^a_i - \beta_i|^q \leq C_q\left(s\sigma_q^q + t(a)^q|S|\mathbf{P}\left(\sigma\varepsilon > (a - t(a\vee t^*)_+)\right)\right). \tag{14}$$

Since $(a - t(a\vee t^*)_+) \leq (a\vee t^* - t(a\vee t^*)_+)$, we get

$$\mathbf{E}\sum_{i\in S}|\hat{\beta}^a_i - \beta_i|^q \leq C_q s\sigma_q^q + C_q t(a\vee t^*)^q|S|\mathbf{P}\left(\sigma\varepsilon > (a\vee t^* - t(a\vee t^*)_+)\right).$$

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 $\log(p/s - 1) \geq 1$.

Appendix H. Proof of Theorem 9

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

$$\begin{aligned}|\xi_i - Y_i\mathbf{1}_{\{|Y_i|\le t(a)\}}|^q &\le |Y_i\mathbf{1}_{\{|Y_i|\le t(a)\}}|^q + q|\xi_i||\xi_i - Y_i\mathbf{1}_{\{|Y_i|\le t(a)\}}|^{q-1}\\ &\le |Y_i 1_{\{|Y_i|\le t(a)\}}|^q + q|\xi_i|2^q\left(|\xi_i|^{q-1} + |Y_i\mathbf{1}_{\{|Y_i|\le t(a)\}}|^{q-1}\right)\\ &\le t(a)^q + q2^q\left(|\xi i|^q + |\xi_i|t(a)^{q-1}\right).\end{aligned}$$

As a consequence

$$\mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \le t(a)^q + q2^q(\sigma_q^q + \sigma_1 t(a)^{q-1}) \le t(a)^q(1+o(1)).$$

The last inequality holds since $t(a) \to \infty$ as $s/p \to 0$. We conclude that

$$\sum_{i\in S}\mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \le \Phi_o(1+o(1)).$$

- 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

$$\sum_{i\in S}\mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \le \frac{st^{*q(1-\epsilon)}}{1+\sqrt{\frac{\pi}{2}}q\epsilon\log\log(p/s-1)}(1+o(1)).$$

- case $a \geq a(1)$:
  In this case it suffices to prove the result for $a = a(1)$ since the minimax risk in increasing with respect to a.

$$\begin{aligned}|\xi_i - Y_i \mathbf{1}_{\{|Y_i| \leq t(a)\}}|^q &\leq |\xi_i|^q + q|Y_i \mathbf{1}_{\{|Y_i| \leq t(a)\}}||\xi_i - Y_i \mathbf{1}_{\{|Y_i| \leq t(a)\}}|^{q-1} \\ &\leq |\xi_i|^q + C_q \left(t^* |\xi_i|^{q-1} \mathbf{1}_{\{|Y_i| \leq t(a)\}} + t^{*q} \mathbf{1}_{\{|Y_i| \leq t(a)\}}\right).\end{aligned}$$

  In the previous case we proved that

$$\mathbf{E}(|\xi_i|^{q-1} \mathbf{1}_{\{|Y_i| \leq t(a)\}}) \leq \log(t^*)^{q-1} \mathbb{P}(|Y_i| \leq t(a)),$$

  and that

$$\mathbf{P}(|Y_i| \leq t(a)) \leq C \frac{t^{*-q}}{\log\log(p/s) + 1}.$$

  Hence

$$\mathbf{E}\left(t^* |\xi_i|^{q-1} \mathbf{1}_{\{|Y_i| \leq t(a)\}} + t^{*q} \mathbf{1}_{\{|Y_i| \leq t(a)\}}\right) \leq \frac{C}{\log\log(p/s)} = o(\sigma_q^q).$$

  It follows that

$$\sum_{i \in S} \mathbf{E}|\hat{\beta}_i^a - \beta_i|^q \leq s\sigma_q^q (1 + o(1)).$$

  This concludes the proof of this theorem.

Appendix I. Proof of Corollary 10

Based on the fact that

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q\right) \geq s\sigma_q^q,$$

we observe that the second result is a direct consequence of Theorem 9. In order to conclude, we need to show that for $a < a_q(1)$ we have

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega_{s,a}^p} \frac{\mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q\right)}{\Phi_o(a)} \geq \frac{1}{4} + o(1).$$

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

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|_q^q\right) \geq \frac{1}{2} s a^q \mathbf{E}\left(T_1^q\right) - C_{q''} s a(1)^q a(0)^{-2q},$$

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

$$sa(1)^q a(0)^{-2q} = o(\Phi_o(a)).$$

It is sufficient to prove that

$$sa^q \mathbf{E}\left(T_1^q\right) \geq \frac{1}{2}\Phi_o(a)(1+o(1)).$$

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 Appendix are sharp combined with the proof of the previous upper bound we can verify that for $q=1$

$$sa\mathbf{E}\left(T_1\right) \geq \frac{1}{2}\Phi_o(a)(1+o(1)).$$

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.

Appendix J. Proof of Theorem 12

First notice that $t_s^* = t(a(1))$ and that $\hat{\beta}^s = \hat{\beta}^{a(1)}$. Hence and using Theorem 9 we get for all $a \geq a(1)$

$$\sup_{\beta \in \Omega_{s,a}^p} \mathbf{E}_\beta\left(\|\hat{\beta}^s - \beta\|_q^q\right) \leq \sup_{\beta \in \Omega_{s,a(1)}^p} \mathbf{E}_\beta\left(\|\hat{\beta}^s - \beta\|_q^q\right) \leq C_q s \sigma_q^q.$$

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

$$\mathbf{E}_\beta \left(\sum_{i \in S} |\hat{\beta}_i^s - \beta_i|^q \right) \leq C_q (s\sigma_q^q + s t_s^{*q}).$$

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.

References

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

Pierre C Bellec. The noise barrier and the large signal bias of the lasso and other convex estimators. *arXiv preprint arXiv:1804.01230*, 2018.

Pierre C Bellec, Guillaume Lecué, Alexandre B Tsybakov, et al. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slopeadaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3): 1103, 2015.

Cristina Butucea, Mohamed Ndaoud, Natalia A Stepanova, and Alexandre B Tsybakov. Variable selection with hamming loss. *The Annals of Statistics*, 46(5):1837–1875, 2018.

David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 41–81, 1992.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Mohamed Ndaoud and Alexandre B Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *arXiv preprint arXiv:1809.03145*, 2018.

Weijie Su and Emmanuel Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

Zheyang Wu and Harrison H Zhou. Model selection and sharp asymptotic minimaxity. *Probability Theory and Related Fields*, 156(1-2):165–191, 2013.