CREST
Centre de recherche
en économie
et statistique

Série des Documents de Travail

n° 2016-10
Variable selection with Hamming loss

C.Butucea[1]
N.A.Stepanova[2]
A.B.Tsybakov[3]

[1] Université Paris-Est Marne-la-Vallée, LAMA (UMR8050), UPEM, UPEC, CNRS, ENSAE. E-mail: cristina.butucea@u-pem.fr

[2] School of Mathematics and Statistics (Carleton University). E-mail: anstepanova@hse.ru

[3] CREST, ENSAE, CNRS. E-mail: alexandre.tsybakov@ensae.fr

Variable selection with Hamming loss

Butucea, C.[1,2], Stepanova, N.A.[3], and Tsybakov, A.B.[2]

[1] Université Paris-Est Marne-la-Vallée, LAMA(UMR 8050), UPEM, UPEC, CNRS,
F-77454, Marne-la-Vallée, France

[2] ENSAE, UMR CNRS 9194, 3, avenue P. Larousse 92245 Malakoff Cedex, France

[3] School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6 Canada

December 8, 2015

Abstract

We derive non-asymptotic bounds for the minimax risk of variable selection under expected Hamming loss in the Gaussian mean model in $\mathbb{R}^d$ for classes of s-sparse vectors separated from 0 by a constant $a > 0$. In some cases, we get exact expressions for the non-asymptotic minimax risk as a function of d, s, a and find explicitly the minimax selectors. Analogous results 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 adaptive to the parameters of the classes.

Keywords: adaptive variable selection, almost full recovery, exact recovery, Hamming loss, minimax selectors, nonasymptotic minimax selection bounds, phase transitions

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, [11, 18, 15, 10, 14, 16, 12, 5, 7] and the references cited therein. Several other recent papers deal with sparse variable selection in nonparametric regression; see, for example, [9, 2, 4, 6, 3].

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

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

where $\xi_1, \dots, \xi_d$ are i.i.d. standard Gaussian random variables, $\sigma > 0$ is the noise level, and $\theta = (\theta_1, \dots, \theta_d)$ 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:

$$\begin{aligned}\Theta_d(s,a) \quad = \quad & \Big\{\theta \in \mathbb{R}^d : \text{ there exists a set } S \subseteq \{1, \dots, d\} \text{ with } s \text{ elements} \\ & \text{such that } |\theta_j| \geq a \text{ for all } j \in S, \text{ and } \theta_j = 0 \text{ for all } j \notin S\Big\}\end{aligned}$$

or

$$\begin{aligned}\Theta_d^+(s,a) \quad = \quad & \Big\{\theta \in \mathbb{R}^d : \text{ there exists a set } S \subseteq \{1, \dots, d\} \text{ with } s \text{ elements} \\ & \text{such that } \theta_j \geq a \text{ for all } j \in S, \text{ and } \theta_j = 0 \text{ for all } j \notin S\Big\}.\end{aligned}$$

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

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

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

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

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

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

$$\mathbf{P}_\theta(\widehat{S} \neq S(\theta)) = \mathbf{P}_\theta(|\widehat{\eta} - \eta| \geq 1) \leq \mathbf{E}_\theta|\widehat{\eta} - \eta|. \tag{2}$$

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 $\mathbf{P}_\theta(\widehat{S} \neq S(\theta))$ is close to 0 in some asymptotic sense (see, for example, [11, 18, 15, 10, 14, 16, 12]). These papers consider high-dimensional linear regression settings with deterministic or random covariates. In particular, for the sequence model (1), one gets that if $a > C\sigma\sqrt{\log d}$ for some $C > 0$ large enough, then there exist selectors such that $\mathbf{P}_\theta(\widehat{S} \neq S(\theta))$ 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 [5, 7] 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 [5, 7] 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 [5, 7] 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 [8].

The present paper makes further steps in the analysis of variable selection with a Hamming loss initiated in [5, 7]. Unlike [5, 7], we study the sequence model (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 non-asymptotic results that are valid for any sample size. Remarkably, we get an exact expression for the non-asymptotic minimax risk and find explicitly the minimax selectors. Finally, we construct data-driven selectors that are simultaneously adaptive to the parameters a and s.

Specifically, we consider the minimax risk

$$\inf_{\tilde{\eta}} \sup_{\theta\in\Theta} \frac{1}{s} \mathbf{E}_\theta |\tilde{\eta} - \eta| \tag{3}$$

for $\Theta = \Theta_d(s, a)$ and $\Theta = \Theta_d^+(s, a)$, where $\inf_{\tilde{\eta}}$ denotes the infimum over all selectors $\widetilde{\eta}$. For the class $\Theta = \Theta_d^+(s, a)$ we find the exact value of the minimax risk and derive a minimax selector for any fixed $d, s, a > 0$ such that $s < d$, whereas for $\Theta = \Theta_d(s, a)$ we propose a selector attaining the minimax 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 $\mathbf{P}_\theta(\widehat{S} \neq S(\theta))$. In Section 3, as asymptotic corollaries of these results, we establish sharp conditions under which exact and almost full recovery are achievable. Section 4 is devoted to the construction of adaptive selectors that achieve almost full and exact recovery without the knowledge of the parameters a and s.

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) for classes Θ different from the sparsity classes that we consider here was analyzed in [3] and without the normalizing factor $1/s$ in [6]. Oracle inequalities for Hamming risks in the problem of multiple classification under sparsity constraints are established in [13]. The paper [17] introduces an asymptotically minimax approach based on the Hamming loss in the problem of community detection in networks.

2 Non-asymptotic 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

$$\hat{\eta}_j = I(|X_j| \geq t), \quad j = 1, \dots, d, \tag{4}$$

where the threshold is defined by

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

Set

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

where $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function, and $x_+ = \max(x, 0)$.

Theorem 2.1. *For any $a > 0$ and $s < d$ the selector $\hat{\eta}$ in (4) with the threshold t defined in (5) satisfies*

$$\sup_{\theta \in \Theta_d(s,a)} \frac{1}{s} \mathbf{E}_\theta |\hat{\eta} - \eta| \leq 2\Psi(d,s,a). \tag{6}$$

Proof. We have, for any $t > 0$,

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

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

$$\mathbf{E}\left(I\left(|\sigma\xi_j + \theta_j| < t\right)\right) \leq \mathbf{P}(|\theta_j| - |\sigma\xi_j| < t) \leq \mathbf{P}(|\xi| > (a-t)/\sigma) = \mathbf{P}(|\xi| > ((a-t)_+)/\sigma),$$

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

$$\frac{1}{s}\mathbf{E}_\theta |\hat{\eta} - \eta| \leq \left(\frac{d}{s} - 1\right) \mathbf{P}(|\xi| \geq t/\sigma) + \mathbf{P}(|\xi| > ((a-t)_+)/\sigma) = 2\Psi(d,s,a). \tag{7}$$

Note that the inequality here is valid for any $t > 0$, not necessarily for t defined in (5). □

We now turn to the class $\Theta_d^+(s,a)$. Consider a selector $\hat{\eta}^+$ with the components

$$\hat{\eta}_j^+ = I(X_j \geq t), \quad j = 1, \dots, d. \tag{8}$$

Set

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

Note that

$$\Psi(d,s,a) \le \Psi_+(d,s,a). \tag{9}$$

Theorem 2.2. *For any $a > 0$ and $s < d$ the selector $\hat{\eta}^+$ in (8) with the threshold t defined in (5) satisfies*

$$\sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta|\hat{\eta}^+ - \eta| \le \Psi_+(d,s,a). \tag{10}$$

Proof. Arguing as in the proof of Theorem 2.1, we obtain

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

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

$$\frac{1}{s}\mathbf{E}_\theta|\hat{\eta}^+ - \eta| \le \left(\frac{d}{s}-1\right)\mathbf{P}(\xi \ge t/\sigma) + \mathbf{P}(\xi < (t-a)/\sigma) = \Psi_+(d,s,a).$$

□

We now establish the lower bound on the minimax risk showing that the upper bound in Theorem 2.2 is sharp.

Theorem 2.3. *For any $a > 0$ and $s < d$ we have*

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

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

Proof. An estimator $\bar{\eta} = (\bar{\eta}_1, \dots, \bar{\eta}_d)$ of η (not necessarily a selector) will be called *separable* if $\bar{\eta}_j$ depends only on X_j for all $j = 1, \dots, d$. First note that instead of considering all selectors, it suffices to prove the lower bound for the class of separable estimators $\bar{\eta}$ with components $\bar{\eta}_j \in [0,1]$. Indeed, for any selector $\widetilde{\eta}$, using Jensen's inequality, we obtain

$$\mathbf{E}_\theta|\widetilde{\eta} - \eta| = \sum_{j=1}^d \mathbf{E}_\theta|\widetilde{\eta}_j - \eta_j| = \sum_{j=1}^d \mathbf{E}_{j,\theta_j}\mathbf{E}_{\{\theta_i, i\ne j\}}|\widetilde{\eta}_j - \eta_j| \ge \sum_{j=1}^d \mathbf{E}_{j,\theta_j}|\bar{\eta}_j - \eta_j|$$

where $\bar{\eta}_j = \mathbf{E}_{\{\theta_i, i\ne j\}}(\widetilde{\eta}_j)$, and the symbols $\mathbf{E}_{j,\theta_j}$ and $\mathbf{E}_{\{\theta_i, i\ne j\}}$ stand for the expectations over the distributions of X_j and $(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_d)$, respectively. Clearly, $\bar{\eta}_j$ depends only on X_j and takes on values in $[0,1]$. Thus,

$$\inf_{\widetilde{\eta}} \sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s}\mathbf{E}_\theta|\widetilde{\eta} - \eta| \ge \inf_{\bar{\eta}\in\mathcal{T}_{[0,1]}} \sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s}\sum_{j=1}^d \mathbf{E}_{j,\theta_j}|\bar{\eta}_j - \eta_j| \tag{11}$$

where $\mathcal{T}_{[0,1]}$ is the class of all separable estimators $\bar{\eta}$ with components $\bar{\eta}_j \in [0,1]$.

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 $\bar{\eta} \in \mathcal{T}_{[0,1]}$ we have

$$\begin{aligned}
\sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s} \sum_{j=1}^{d} \mathbf{E}_{j,\theta_j} |\bar{\eta}_j - \eta_j| \ &\geq \ \frac{1}{s|\Theta'|} \sum_{\theta \in \Theta'} \sum_{j=1}^{d} \mathbf{E}_{j,\theta_j} |\bar{\eta}_j - \eta_j| \qquad (12) \\
&= \ \frac{1}{s|\Theta'|} \sum_{j=1}^{d} \Big(\sum_{\theta \in \Theta' : \theta_j = 0} \mathbf{E}_{j,0}(\bar{\eta}_j) + \sum_{\theta \in \Theta' : \theta_j = a} \mathbf{E}_{j,a}(1 - \bar{\eta}_j) \Big) \\
&= \ \frac{1}{s} \sum_{j=1}^{d} \Big(\Big(1 - \frac{s}{d}\Big) \mathbf{E}_{j,0}(\bar{\eta}_j) + \frac{s}{d} \, \mathbf{E}_{j,a}(1 - \bar{\eta}_j) \Big) \\
&\geq \ \frac{d}{s} \inf_{T \in [0,1]} \Big(\Big(1 - \frac{s}{d}\Big) \mathbb{E}_0(T) + \frac{s}{d} \, \mathbb{E}_a(1 - T) \Big),
\end{aligned}$$

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 (12), $\mathbb{E}_u$ 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

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

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

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

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

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

Combining this with (11) and (12), we get

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

□

As a straightforward corollary of Theorems 2.2 and 2.3, we obtain that the estimator $\hat{\eta}^+$ is minimax in the exact sense for the class $\Theta_d^+(s,a)$ and the minimax risk satisfies

$$\inf_{\widetilde{\eta}} \sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s} \mathbf{E}_\theta |\widetilde{\eta} - \eta| = \sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s} \mathbf{E}_\theta |\hat{\eta}^+ - \eta| = \Psi_+(d, s, a).$$

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 $s < d$. Analogous non-asymptotic minimax result is valid for the class

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

We omit the details here. Finally, the following corollary is an immediate consequence of Theorems 2.1, 2.3, and inequality (9).

Corollary 2.1. *For any $a > 0$ and $s < d$ the selector $\hat{\eta}$ in (4) with the threshold t defined in (5) satisfies*

$$\sup_{\theta\in\Theta_d(s,a)} \frac{1}{s}\mathbf{E}_\theta|\hat{\eta}-\eta| \le 2 \inf_{\widetilde{\eta}} \sup_{\theta\in\Theta_d(s,a)} \frac{1}{s}\mathbf{E}_\theta|\widetilde{\eta}-\eta|. \tag{13}$$

Thus, the risk of the thresholding estimator (4) cannot be greater than the minimax risk over the class $\Theta_d(s,a)$ multiplied by 2.

Remark 2.1. *From the proof of Theorem 2.3 we see that, for each j, the minimax optimal selector $\hat{\eta}_j^+$ coincides with the Bayes test of the null hypothesis $H_0 : \theta_j = 0$ against the alternative $H_0 : \theta_j = a$ with prior probabilities $1 - s/d$ and s/d, respectively.*

We now show that the above non-asymptotic minimax results can be extended to the probability of wrong recovery. For any selector $\widetilde{\eta}$, we denote by $S_{\widetilde{\eta}}$ the selected set of indices: $S_{\widetilde{\eta}} = \{j : \widetilde{\eta}_j = 1\}$. Let $\mathcal{T}$ be the set of all separable selectors $\widetilde{\eta}$, that is, the set of selectors $\widetilde{\eta}$ such that the jth component $\widetilde{\eta}_j$ depends only on X_j for all $j = 1, \dots, d$.

Theorem 2.4. *For any $a > 0$ and $s < d$ the selectors $\hat{\eta}$ in (4) and $\hat{\eta}^+$ in (8) with the threshold t defined in (5) satisfy*

$$\sup_{\theta\in\Theta_d^+(s,a)} \mathbf{P}_\theta(S_{\hat{\eta}^+} \ne S(\theta)) \le s\Psi_+(d,s,a), \tag{14}$$

and

$$\sup_{\theta\in\Theta_d(s,a)} \mathbf{P}_\theta(S_{\hat{\eta}} \ne S(\theta)) \le 2s\Psi(d,s,a). \tag{15}$$

Furthermore,

$$\inf_{\widetilde{\eta}\in\mathcal{T}} \sup_{\theta\in\Theta_d^+(s,a)} \mathbf{P}_\theta(S_{\widetilde{\eta}} \ne S(\theta)) \ge \frac{s\Psi_+(d,s,a)}{1+s\Psi_+(d,s,a)}. \tag{16}$$

Proof. The upper bounds (14) and (15) follow immediately from (2) and Theorems 2.1 and 2.2. We now prove the lower bound (16). To this end, first note that for any $\theta \in \Theta_d^+(s,a)$ and any $\widetilde{\eta} \in \mathcal{T}$ we have

$$\mathbf{P}_\theta(S_{\widetilde{\eta}} \ne S(\theta)) = \mathbf{P}_\theta(\cup_{j=1}^d\{\widetilde{\eta}_j \ne \eta_j\}) = 1 - \prod_{j=1}^d p_j(\theta)$$

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

$$\sup_{\theta \in \Theta_d^+(s,a)} \mathbf{P}_\theta(S_{\widetilde{\eta}} \neq S(\theta)) \geq \max_{\theta \in \Theta'} \mathbf{P}_\theta(S_{\widetilde{\eta}} \neq S(\theta)) = 1 - p_* \tag{17}$$

where Θ' is the subset of $\Theta_d^+(s,a)$ defined in the proof of Theorem 2.3, and $p_* = \min_{\theta \in \Theta'} \prod_{j=1}^d p_j(\theta)$.

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

$$\sup_{\theta \in \Theta_d^+(s,a)} \mathbf{P}_\theta(S_{\widetilde{\eta}} \neq S(\theta)) \quad \geq \quad \frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \mathbf{P}_\theta(|\widetilde{\eta} - \eta| = 1). \tag{18}$$

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

$$\begin{aligned}
&\frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \mathbf{P}_\theta(|\widetilde{\eta} - \eta| = 1) = \frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \sum_{j=1}^d \mathbf{P}_\theta(B_j) \\
=\ & \frac{1}{|\Theta'|} \sum_{j=1}^d \Big(\sum_{\theta \in \Theta' : \theta_j = 0} \mathbf{P}_{j,0}(\widetilde{\eta}_j = 1) \prod_{i \neq j} p_i(\theta) + \sum_{\theta \in \Theta' : \theta_j = a} \mathbf{P}_{j,a}(\widetilde{\eta}_j = 0) \prod_{i \neq j} p_i(\theta) \Big) \\
\geq\ & \frac{p_*}{|\Theta'|} \sum_{j=1}^d \Big(\sum_{\theta \in \Theta' : \theta_j = 0} \mathbf{P}_{j,0}(\widetilde{\eta}_j = 1) + \sum_{\theta \in \Theta' : \theta_j = a} \mathbf{P}_{j,a}(\widetilde{\eta}_j = 0) \Big) \\
=\ & \frac{p_*}{|\Theta'|} \sum_{j=1}^d \Big(\sum_{\theta \in \Theta' : \theta_j = 0} \mathbf{E}_{j,0}(\widetilde{\eta}_j) + \sum_{\theta \in \Theta' : \theta_j = a} \mathbf{E}_{j,a}(1 - \widetilde{\eta}_j) \Big)
\end{aligned} \tag{19}$$

where $\mathbf{P}_{j,u}$ denotes the distribution of X_j when $\theta_j = u$. We now bound the right-hand side of (19) by following the argument from the last three lines of (12) to the end of the proof of Theorem 2.3. Applying this argument yields that, for any $\widetilde{\eta} \in \mathcal{T}$,

$$\frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \mathbf{P}_\theta(|\widetilde{\eta} - \eta| = 1) \quad \geq \quad p^* d L^* \geq p^* s \Psi_+(d,s,a). \tag{20}$$

Combining (17), (18), and (20), we find that, for any $\widetilde{\eta} \in \mathcal{T}$,

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

□

Although Theorem 2.4 does not provide the exact minimax solution, it implies sharp asymptotic minimaxity. Indeed, an interesting case is when the minimax risk in Theorem 2.4 goes to 0 as $d \to \infty$. Assuming that s and a depend on d in some way, this corresponds to $s\Psi_+(d,s,a) \to 0$. In this natural asymptotic setup, the upper and lower bounds of Theorem 2.4 for the class $\Theta_d^+(s,a)$ are sharp. We discuss this issue in more detail in the next section, cf. Theorem 3.5.

Remark 2.2. *In papers [5, 7, 8], a different Hamming loss defined in terms of the vectors of signs is used. 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 non-asymptotic 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 the next two sections, 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 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 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 [5, 7] but we define these properties in a different way, as asymptotic minimax properties for classes of vectors θ. The papers [5, 7] considered a Bayesian setup with random θ and studied a linear regression model with i.i.d. Gaussian regressors rather than the sequence model (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 3.1. *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*

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta |\hat{\eta}-\eta| = 0. \tag{21}$$

  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*

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d} \mathbf{E}_\theta |\hat{\eta}-\eta| = 0. \tag{22}$$

  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 3.2. *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*

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\tilde{\eta} - \eta| > 0, \tag{23}$$

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

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d}\mathbf{E}_\theta|\tilde{\eta} - \eta| > 0, \tag{24}$$

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

The following general characterization theorem is a straightforward corollary of the results of Section 2.

Theorem 3.1. *(i) Almost full recovery is possible for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if and only if*

$$\Psi_+(d, s_d, a_d) \to 0 \quad \text{as } d \to \infty. \tag{25}$$

In this case, the selector $\hat{\eta}$ defined in (4) *with threshold* (5) *achieves almost full recovery.*

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

$$s_d\Psi_+(d, s_d, a_d) \to 0 \quad \text{as } d \to \infty. \tag{26}$$

In this case, the selector $\hat{\eta}$ defined in (4) *with threshold* (5) *achieves exact recovery.*

Although this theorem gives a complete solution to the problem, conditions (25) and (26) 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:

$$\sqrt{\frac{2}{\pi}}\,\frac{e^{-y^2/2}}{y+\sqrt{y^2+4}} \leq \frac{1}{\sqrt{2\pi}}\int_y^\infty e^{-u^2/2}du \leq \sqrt{\frac{2}{\pi}}\,\frac{e^{-y^2/2}}{y+\sqrt{y^2+8/\pi}}, \quad \forall y > 0. \tag{27}$$

These bounds are an immediate consequence of formula 7.1.13. in [1] with $x = y/\sqrt{2}$.

Furthermore, we will need some non-asymptotic 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 3.2. *Assume that* $s < d/2$.

(i) If

$$a^2 \geq \sigma^2\Big(2\log((d-s)/s) + W\Big) \text{ for some } W > 0, \tag{28}$$

then the selector $\hat\eta$ *defined in* (4) *with threshold* (5) *satisfies*

$$\sup_{\theta\in\Theta_d(s,a)} \mathbf{E}_\theta|\hat\eta - \eta| \leq (2+\sqrt{2\pi})s\,\Phi(-\Delta), \tag{29}$$

where Δ *is defined by*

$$\Delta = \frac{W}{2\sqrt{2\log((d-s)/s) + W}}\,. \tag{30}$$

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

$$a^2 \leq \sigma^2\Big(2\log((d-s)/s) + W\Big) \text{ for some } W > 0, \tag{31}$$

then

$$\inf_{\widetilde\eta}\sup_{\theta\in\Theta_d(s,a)} \mathbf{E}_\theta|\widetilde\eta - \eta| \geq s\,\Phi(-\Delta), \tag{32}$$

where the infimum is taken over all selectors $\widetilde\eta$ *and* $\Delta > 0$ *is defined in* (30).

Proof. (i) In the proof of Theorem 2.1, we have obtained that

$$\sup_{\theta\in\Theta_d(s,a)} \frac{1}{s}\mathbf{E}_\theta|\hat\eta - \eta| \leq 2\left(\frac{d}{s}-1\right)\Phi(-t/\sigma) + 2\Phi(-(a-t)_+/\sigma), \tag{33}$$

where $t = \frac{a}{2} + \frac{\sigma^2}{a}\log\left(\frac{d}{s}-1\right)$ is the threshold (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, $\left(\frac{d}{s}-1\right)e^{-t^2/(2\sigma^2)} = e^{-(a-t)^2/(2\sigma^2)}$. These remarks and (27) 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 (33) 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 (29) it remains to note that under assumption (28),

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

Indeed, assumption (28) states that $a \geq a_0 \triangleq \sigma\Big(2\log((d-s)/s)+W\Big)^{1/2}$, and the function $a \mapsto \big(a^2 - 2\sigma^2\log((d-s)/s)\big)/a$ is monotonically increasing in $a > 0$. On the other hand,

$$\big(a_0^2 - 2\sigma^2\log((d-s)/s)\big)/(2a_0\sigma) = \Delta. \qquad (34)$$

(ii) We now prove (32). By Theorem 2.3,

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

Here,

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

Observe that the function $a \mapsto \big(2\sigma^2\log((d-s)/s)-a^2\big)/a$ is monotonically decreasing in $a > 0$ and that assumption (31) states that $a \leq a_0$. In view of (34), the value of its minimum for $a \leq a_0$ is equal to $-\Delta$. The bound (32) now follows by the monotonicity of $\Phi(\cdot)$. □

The next theorem is an easy consequence of Theorem 3.2. It describes a "phase transition" for a_d in the problem of almost full recovery.

Theorem 3.3. *Assume that* $\limsup_{d\to\infty} s_d/d < 1/2$.

(i) If, for all d large enough,

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

for an arbitrary sequence $A_d \to \infty$, as $d\to\infty$, then the selector $\hat\eta$ defined by (4) *and* (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 d large enough the reverse inequality holds:

$$a_d^2 \leq \sigma^2\Big(2\log((d-s_d)/s_d)+A\sqrt{2\log((d-s_d)/s_d)}\Big)$$

then almost full recovery is impossible:

$$\liminf_{d\to\infty} \inf_{\tilde\eta} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d}\mathbf{E}_\theta|\tilde\eta-\eta| \geq \Phi\Big(-\frac{A}{2}\Big) > 0.$$

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

Proof. Assume without loss of generality that d is large enough to have $(d-s_d)/s_d > 1$. We apply Theorem 3.2 with $W = A\sqrt{2\log((d-s_d)/s_d)}$. Then,

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

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 (29) imply part (*i*) of the theorem. Part (*ii*) follows from (32) by noticing that $\Delta^2 \le \sup_{x>0} \frac{A^2 x}{4(x+A)} = A^2/4$ for any fixed $A > 0$. $\square$

Under the natural assumption that

$$d/s_d \to \infty \qquad \text{as} \qquad d \to \infty, \tag{35}$$

Theorem 3.3 shows that the "phase transition" for almost full recovery occurs at the value $a_d = a_d^*$, where

$$a_d^* = \sigma\sqrt{2\log((d-s_d)/s_d)}\Big(1+o(1)\Big). \tag{36}$$

Furthermore, Theorem 3.3 details the behavior of the $o(1)$ term here.

We now state a corollary of Theorem 3.3 under simplified assumptions.

Corollary 3.1. *Assume that* (35) *holds and set* $a_d = \sigma\sqrt{2(1+\delta)\log(d/s_d)}$ *for some* $\delta > 0$. *Then the selector* $\hat\eta$ *defined by* (4) *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 (35) is satisfied. Then $\log(d/s_d) = \beta(1+o(1))\log d$ and it follows from Corollary 3.1 that for $a_d = \sigma\sqrt{2\beta(1+\delta)\log d}$ the selector with components $\hat\eta_j = I\big(|X_j| > \sigma\sqrt{2\beta(1+\varepsilon)\log d}\big)$ achieves almost full recovery. This is in agreement with the findings of [5, 7] 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, a "phase transition" for a_d in the problem of exact recovery is described in the next theorem.

Theorem 3.4. *Assume that* $s_d \to \infty$ *as* $d \to \infty$*, and* $\limsup_{d\to\infty} s_d/d < 1/2$.

(i) If

$$a_d^2 \ge \sigma^2\Big(2\log((d-s_d)/s_d) + W_d\Big)$$

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

$$\liminf_{d\to\infty} \frac{W_d}{4\Big(\log(s_d) + \sqrt{\log(s_d)\log(d-s_d)}\Big)} \ge 1, \tag{37}$$

then the selector $\hat{\eta}$ defined by (4) and (5) achieves exact recovery:

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat{\eta}-\eta| = 0. \tag{38}$$

(ii) If the complementary condition holds:

$$a_d^2 \le \sigma^2\Big(2\log((d-s_d)/s_d) + W_d\Big)$$

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

$$\limsup_{d\to\infty} \frac{W_d}{4\Big(\log(s_d) + \sqrt{\log(s_d)\log(d-s_d)}\Big)} < 1, \tag{39}$$

then exact recovery is impossible, and moreover we have

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

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

Proof. 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 $W_*(s) \triangleq 4\Big(\log s + \sqrt{\log s \log(d-s)}\Big)$, and notice that

$$\frac{W_*(s_d)}{2\sqrt{2\log((d-s_d)/s_d) + W_*(s_d)}} = \sqrt{2\log s_d}, \tag{40}$$

$$2\log((d-s_d)/s_d) + W_*(s_d) = 2\Big(\sqrt{\log(d-s_d)} + \sqrt{\log s_d}\Big)^2. \tag{41}$$

If (37) holds, we have $W_d \ge W_*(s_d)$ for all d large enough. By the monotonicity of the quantity Δ defined in (30) with respect to W, this implies

$$\Delta_d \triangleq \frac{W_d}{2\sqrt{2\log((d-s_d)/s_d) + W_d}} \ge \frac{W_*(s_d)}{2\sqrt{2\log((d-s_d)/s_d) + W_*(s_d)}} = \sqrt{2\log s_d}\,. \tag{42}$$

Now, by Theorem 3.2 and using (27) we may write

$$\begin{aligned} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat{\eta}-\eta| &\le (2+\sqrt{2\pi})s_d\,\Phi\left(-\Delta_d\right) \le 3s_d \min\left\{1, \frac{1}{\Delta_d}\right\} \exp\left(-\frac{\Delta_d^2}{2}\right) \\ &= 3\min\left\{1, \frac{1}{\Delta_d}\right\} \exp\left(-\frac{\Delta_d^2 - 2\log s_d}{2}\right). \end{aligned} \tag{43}$$

This and (42) 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 non-positive W_d almost full recovery is impossible and the result follows from part (ii) of Theorem 3.3. If (39) 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 (30) with respect to W and in view of equation (40), this implies

$$\begin{aligned}
\Delta_d^2 - 2\log s_d &\le \frac{A^2 W_*^2(s_d)}{4(2\log((d-s_d)/s_d) + AW_*(s_d))} - \frac{W_*^2(s_d)}{4(2\log((d-s_d)/s_d) + W_*(s_d))} \\
&= \frac{(A-1)W_*^2(s_d)(AW_*(s_d) + 2(A+1)\log((d-s_d)/s_d))}{4(2\log((d-s_d)/s_d) + AW_*(s_d))(2\log((d-s_d)/s_d) + W_*(s_d))} \\
&\le \frac{(A-1)AW_*^2(s_d)}{4(2\log((d-s_d)/s_d) + W_*(s_d))} \\
&= \frac{2(A-1)A\Big(\log s_d + \sqrt{\log s_d \log(d-s_d)}\Big)^2}{\Big(\sqrt{\log(d-s_d)} + \sqrt{\log s_d}\Big)^2} = 2(A-1)A\log s_d, \qquad (44)
\end{aligned}$$

where we have used the fact that $A < 1$ and equations (40), (41). Next, by Theorem 3.2 and using (27), we have

$$\begin{aligned}
\inf_{\widetilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\widetilde{\eta} - \eta| &\ge s_d\,\Phi\left(-\Delta_d\right) \ge \frac{s_d}{4}\min\left\{\frac{1}{2}, \frac{1}{\Delta_d}\right\}\exp\left(-\frac{\Delta_d^2}{2}\right) \\
&= \frac{1}{4}\min\left\{\frac{1}{2}, \frac{1}{\Delta_d}\right\}\exp\left(-\frac{\Delta_d^2 - 2\log s_d}{2}\right).
\end{aligned}$$

Combining this inequality with (44), we find that, for all d large enough,

$$\inf_{\widetilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\widetilde{\eta} - \eta| \ge \frac{1}{4}\min\left\{\frac{1}{2}, \frac{1}{\Delta_d}\right\}\exp\left((1-A)A\log s_d\right).$$

Since $A < 1$ and $\Delta_d \le A\sqrt{2\log s_d}$ by (44), the last expression tends to ∞ as $s_d \to \infty$. This proves part (ii) of the theorem. □

Some remarks are in order here. First of all, Theorem 3.4 and (41) show that the "phase transition" for exact recovery occurs at the value $a_d = a_d^*$, where

$$a_d^* = \sigma\Big(\sqrt{2\log(d-s_d)} + \sqrt{2\log s_d}\Big). \qquad (45)$$

This is larger than the value a_d^* for almost full recovery, cf. (36), which is intuitively quite clear. The optimal threshold (5) corresponding to (45) 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 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 [5, 7] 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 4.1 below.

Finally, we state an asymptotic corollary of Theorem 2.4 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 3.5. *Assume that exact recovery is possible for the classes* $(\Theta_d(s_d, a_d))_{d\ge 1}$ *and* $(\Theta_d^+(s_d, a_d))_{d\ge 1}$*, that is, condition* (26) *holds. Then, for the selectors* $\hat\eta$ *and* $\hat\eta^+$ *defined by* (4)*,* (8)*, and* (5) *we have*

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

and

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

$$\liminf_{d\to\infty} \inf_{\tilde\eta\in\mathcal{T}} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{\mathbf{P}_\theta(S_{\tilde\eta} \neq S(\theta))}{s_d\Psi_+(d, s_d, a_d)} \ge 1.$$

Note that the threshold (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.

4 Adaptive selectors

In this section, we consider the asymptotic setup as in Section 3 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, 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. (45)), 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 4.1. *Assume that $s_d \to \infty$ as $d \to \infty$ and that $\limsup_{d\to\infty} s_d/d < 1/2$. Let the sequence $(a_d)_{d\geq 1}$ be above the phase transition level for exact recovery, that is, $a_d \geq a_d^*$ for all d, where a_d^* is defined in* (45)*. Then the selector $\hat\eta$ defined by* (4) *with threshold $t = \sigma\sqrt{2\log d}$ achieves exact recovery.*

Proof. By (7), for any $\theta \in \Theta_d(s, a)$, and any $t > 0$ we have

$$\mathbf{E}_\theta|\hat\eta - \eta| \leq (d - s)\,\mathbf{P}(|\xi| \geq t/\sigma) + s\mathbf{P}(|\xi| > (a - t)_+/\sigma),$$

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 $t > 0$,

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

It suffices to consider $d \geq 9$, and $2 \leq s_d \leq d/2$. Then, using (45) we get

$$a_d^* \geq \sigma \min_{2\leq x\leq d/2}\left(\sqrt{2\log(d-x)} + \sqrt{2\log x}\right) = 2\sigma\sqrt{2\log(d/2)} \geq 2\sigma\sqrt{\log d}.$$

Thus, $(a_d^* - t)_+/\sigma \geq (2 - \sqrt{2})\sqrt{\log d}$ for our choice of t. Using this inequality, (27) and (45), we find

$$\begin{aligned}\sup_{\theta\in\Theta_d(s_d,a_d)} \mathbf{E}_\theta|\hat\eta - \eta| \;&\leq\; \frac{1}{\sqrt{2\log d}} + \frac{s_d \exp\left(-\left[\sqrt{\log(d-s_d)} + \sqrt{\log(s_d)} - \sqrt{\log d}\right]^2\right)}{(2-\sqrt{2})\sqrt{\log d}} \\ &\leq\; \frac{1}{\sqrt{2\log d}} + \frac{\exp\left(2(\sqrt{\log d} - \sqrt{\log(d-s_d)})\sqrt{\log(s_d)}\right)}{(2-\sqrt{2})\sqrt{\log d}}\end{aligned}$$

and the theorem follows since, under our assumptions, $(\sqrt{\log d} - \sqrt{\log(d-s_d)})\sqrt{\log(s_d)} \leq (\sqrt{\log d} - \sqrt{\log d - \log 2})\sqrt{\log(s_d)} = O(1)$ as $d \to \infty$. □

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 that this includes a much broader range of values than in case of exact recovery. Thus, using the adaptive selector of Theorem 4.1 for almost full recovery does not give a satisfactory result, and we have to take a different approach.

Following Section 3, we will use the notation

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

As shown in Section 3, 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 we have used the assumption that $d/s_d \to \infty$ (the sparsity assumption), which is now transformed into the condition

$$s_d \in \mathcal{S}_d \triangleq \{1, 2, \ldots, s_d^*\} \text{ where } s_d^* \text{ is an integer such that } \frac{d}{s_d^*} \to \infty \text{ as } d \to \infty. \tag{46}$$

Assuming s_d to be known, we have shown in Section 3 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 3.3. However, for adaptive estimation considered in this section we will need the following mild assumption on the growth of A_d:

$$A_d \geq c_0 \left(\log\log\left(\frac{d}{s_d^*} - 1\right)\right)^{1/2}, \tag{47}$$

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 (47) 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, $m = 1, ..., M$, we define a selector

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

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:

$$\widehat{m} = \min\Big\{ m \in \{2, \ldots, M\} : \sum_{j=1}^{d} I\big(w(g_k) \leq |X_j| < w(g_{k-1})\big) \leq \tau g_k \text{ for all } k \geq m \Big\}, \tag{48}$$

where

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

Finally, we define an adaptive selector as

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

Theorem 4.2. *Let $c_0 \geq 4$. Then the selector $\hat{\eta}^{\mathrm{ad}}$ adaptively achieves almost full recovery in the following sense:*

$$\lim_{d\to\infty} \sup_{\theta \in \Theta_d(s_d, a_d)} \frac{1}{s_d} \mathbf{E}_\theta |\hat{\eta}^{\mathrm{ad}} - \eta| = 0 \tag{49}$$

for all sequences $(s_d, a_d)_{d \geq 1}$ such that (46) holds and $a_d \geq a_0(s_d, A_d)$, where A_d satisfies (47).

Proof. 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

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s,a_0(s,A))} \frac{1}{s}\mathbf{E}_\theta|\hat{\eta}^{\mathrm{ad}} - \eta| = 0. \tag{50}$$

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. We first decompose the risk as follows:

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

where

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

with m_0 being the index of the minimal element of the grid g_{m_0} that is greater than the true underlying s: $g_{m_0-1} \leq s < g_{m_0}$.

We now evaluate I_1. First note that $I_1 = 0$ if $m_0 = 1$ since by definition $\widehat{m} \geq 2$. Thus, consider the case $m_0 \geq 2$. Using the fact that $\hat{\eta}_j(g_m)$ is monotonically increasing in m and the definition of $\widehat{m}$, we obtain that, on the event $\{\widehat{m} \leq m_0\}$,

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

where we have used that $g_m = 2^m$ and the fact that, by the choice of m_0, we have $s \geq g_{m_0-1} = g_{m_0}/2$. Thus,

$$\begin{aligned} I_1 &\leq \frac{1}{s}\mathbf{E}_\theta\left(|\hat{\eta}(g_{\widehat{m}}) - \hat{\eta}(g_{m_0})| I(\widehat{m} \leq m_0)\right) + \frac{1}{s}\mathbf{E}_\theta|\hat{\eta}(g_{m_0}) - \eta| \\ &\leq 4\tau + \frac{1}{s}\mathbf{E}_\theta|\hat{\eta}(g_{m_0}) - \eta|. \end{aligned} \tag{51}$$

By (7), for any $\theta \in \Theta_d(s, a_0(s,A))$ we have

$$\frac{1}{s}\mathbf{E}_\theta|\hat{\eta}(g_{m_0}) - \eta| \leq \left(\frac{d}{s} - 1\right)\mathbf{P}(|\xi| \geq w(g_{m_0})) + \mathbf{P}(|\xi| > (a_0(s,A) - w(g_{m_0}))_+) \tag{52}$$

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

$$\begin{aligned}
\left(\frac{d}{s}-1\right)\mathbf{P}(|\xi| \geq w(g_{m_0})) &\leq \frac{d/s-1}{d/g_{m_0}-1}\ \frac{\pi^{-1/2}}{\sqrt{\log(d/g_{m_0}-1)}} \qquad (53)\\
&\leq \frac{d-s}{d-2s}\ \frac{2\pi^{-1/2}}{\sqrt{\log(d/s-1)}} \leq \frac{3\pi^{-1/2}}{\sqrt{\log(d/s_d^*-1)}}\ .
\end{aligned}$$

To bound the second probability on the right-hand side of (52), we use the following lemma.

Lemma 4.1. *Under the assumptions of Theorem 4.2, for any $m \geq m_0$ we have*

$$\mathbf{P}(|\xi| > (a_0(s,A) - w(g_m))_+) \leq \big(\log\left(d/s_d^* - 1\right)\big)^{-\frac{1}{2}}. \qquad (54)$$

Combining (52), (53) and (54) with $m = m_0$, we find

$$\frac{1}{s}\mathbf{E}_\theta|\hat{\eta}(g_{m_0}) - \eta| \ \ \leq\ \ \frac{3\pi^{-1/2}+1}{\sqrt{\log(d/s_d^*-1)}}\,, \qquad (55)$$

which together with (51) leads to the bound

$$I_1 \ \ \leq\ \ 4\tau + \frac{3\pi^{-1/2}+1}{\sqrt{\log(d/s_d^*-1)}}\,. \qquad (56)$$

We now turn to the evaluation of I_2. We have

$$\begin{aligned}
I_2 \ &= \ \frac{1}{s}\sum_{m=m_0+1}^{M}\mathbf{E}_\theta\left(|\hat{\eta}(g_{\widehat{m}}) - \eta| I(\widehat{m} = m)\right) \qquad (57)\\
&\leq \ \frac{1}{s}\sum_{m=m_0+1}^{M}\big(\mathbf{E}_\theta\left(|\hat{\eta}(g_m) - \eta|^2\right)\big)^{1/2}\big(\mathbf{P}_\theta(\widehat{m} = m)\big)^{1/2}\\
&= \ \frac{1}{s}\sum_{m=m_0+1}^{M}\big(\mathbf{E}_\theta|\hat{\eta}(g_m) - \eta|\big)^{1/2}\big(\mathbf{P}_\theta(\widehat{m} = m)\big)^{1/2}.
\end{aligned}$$

By definition, the event $\{\widehat{m} = m\}$ occurs if and only if there exists some $\ell \geq m$ such that $\sum_{j=1}^d I(w_\ell \leq |X_j| < w_{\ell-1}) > \tau g_\ell \triangleq v_\ell$, where we set for brevity $w_\ell = w(g_\ell)$. Thus,

$$\mathbf{P}_\theta(\widehat{m} = m) \ \ \leq\ \ \sum_{\ell=m}^{M}\mathbf{P}_\theta\left(\sum_{j=1}^{d} I(w_\ell \leq |X_j| < w_{\ell-1}) > v_\ell\right)\ . \qquad (58)$$

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

$$\begin{aligned}
&\mathbf{P}_\theta\left(\sum_{j=1}^{d} I(w_\ell \leq |X_j| < w_{\ell-1}) - \mathbf{E}_\theta\left(\sum_{j=1}^{d} I(w_\ell \leq |X_j| < w_{\ell-1})\right) > t\right)\\
&\qquad \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^d \mathbf{E}_\theta\left(I(w_\ell \leq |X_j| < w_{\ell-1})\right) + 2t/3}\right), \qquad (59)
\end{aligned}$$

where we have used that, for random variables with values in $\{0,1\}$, the variance is smaller than the expectation. Now, similar to (7), for any $\theta \in \Theta_d(s, a_0(s,A))$,

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

where ξ is a standard Gaussian random variable. Since $\ell \ge m_0+1$, from Lemma 4.1 we get

$$\mathbf{P}(|\xi| > (a_0(s,A)-w_{\ell-1}))_+) \le \big(\log\left(d/s_d^*-1\right)\big)^{-\frac{1}{2}}. \tag{60}$$

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

$$(d-s)\mathbf{P}\left(|\xi| \ge w_\ell\right) \le \frac{d-s}{d/g_\ell - 1}\frac{\pi^{-1/2}}{\sqrt{\log(d/g_\ell-1)}} \le \frac{(4/3)\pi^{-1/2}g_\ell}{\sqrt{\log(d/s_d^*-1)}}. \tag{61}$$

We now deduce from (60) and (61), and the inequality $s \le g_\ell$ for $\ell \ge m_0+1$, that

$$\mathbf{E}_\theta\Big(\sum_{j=1}^d I(w_\ell \le |X_j| < w_{\ell-1})\Big) \le \frac{\big((4/3)\pi^{-1/2}+1\big)g_\ell}{\sqrt{\log(d/s_d^*-1)}} \le 2\tau g_\ell. \tag{62}$$

Taking in (59) $t = 3\tau g_\ell = 3v_\ell$ and using (62), we find

$$\mathbf{P}_\theta\left(\sum_{j=1}^d I(w_\ell \le |X_j| < w_{\ell-1}) > v_\ell\right) \le \exp(-C_1 v_\ell) = \exp(-C_1 2^\ell \tau),$$

for all $\ell \ge m_0+1$ and some absolute constant $C_1 > 0$. This implies

$$\mathbf{P}_\theta(\widehat{m} = m) \le \sum_{\ell=m}^M \exp(-C_1 2^\ell\tau) \le C_2 2^{-m}\tau^{-1}\exp(-C_1 2^m \tau) \tag{63}$$

for some absolute constant $C_2 > 0$.

On the other hand, notice that the bounds (52), and (53) are valid not only for g_{m_0} but also for any g_m with $m \ge m_0+1$. Using this observation and Lemma 4.1 we get that, for any $\theta \in \Theta_d(s, a_0(s,A))$ and any $m \ge m_0+1$,

$$\begin{aligned}\mathbf{E}_\theta|\hat\eta(g_m) - \eta| &\le s\left[\frac{d/s-1}{d/g_m-1}\ \frac{\pi^{-1/2}}{\sqrt{\log(d/g_m-1)}} + \big(\log\left(d/s_d^*-1\right)\big)^{-\frac{1}{2}}\right] \\ &\le \frac{\big((4/3)\pi^{-1/2}+1\big)g_m}{\sqrt{\log(d/s_d^*-1)}} \triangleq \tau' g_m = \tau' 2^m,\end{aligned} \tag{64}$$

where the last inequality follows from the same argument as in (61).

Now, we plug (63) and (64) in (57) to obtain

$$\begin{aligned}I_2 &\le \frac{C_2^{1/2}(\tau'/\tau)^{1/2}}{s}\sum_{m=m_0+1}^M \exp(-C_1 2^{m-1}\tau) \\ &\le C_3(\tau')^{1/2}\tau^{-3/2}\exp(-C_1 2^{m_0}\tau) \le C_3(\tau')^{1/2}\tau^{-3/2}\end{aligned} \tag{65}$$

for some absolute constant $C_3 > 0$. Notice that $(\tau')^{1/2} = O\big(\big(\log(d/s_d^* - 1)\big)^{-\frac{1}{4}}\big)$ as $d/s_d^* \to \infty$ while $\tau^{-3/2} = O\big(\big(\log(d/s_d^* - 1)\big)^{\frac{3}{14}}\big)$. Thus, $I_2 = o(1)$ as $d \to \infty$. Since from (56) we also get that $I_1 = o(1)$ as $d \to \infty$, the proof is complete. □

Proof of Lemma 4.1. Since $g_m > s$ for $m \geq m_0$, we have $w(g_m) < w(s)$. It follows that

$$a_0(s, A) - w(g_m) \geq a_0(s, A) - w(s) \geq \frac{\sqrt{A}}{2\sqrt{2}} \min\left(\frac{\sqrt{A}}{\sqrt{2}},\ \log^{1/4}(d/s - 1)\right),$$

where we have used the elementary inequalities

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

with $x = 2\log(d/s - 1)$ and $y = A\sqrt{\log(d/s-1)}$. By assumption, $A \geq 4\sqrt{\log\log\big(d/s_d^* - 1\big)}$, so that we get

$$a_0(s, A) - w(g_m) \geq a_0(s, A) - w(s) \geq \left(\log\log\left(\frac{d}{s_d^*} - 1\right)\right)^{1/2}.$$

This and the bound on the Gaussian tail probability imply

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

□

Acknowledgements. We would like to thank Felix Abramovich for helpful discussion of the results. The work of N.A. Stepanova was supported by an NSERC grant. The work of A.B. Tsybakov was supported by GENES and by the French National Research Agency (ANR) under the grants IPANEMA (ANR-13-BSH1-0004-02), and Labex ECODEC (ANR - 11-LABEX-0047). It was also supported by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

References

[1] M. Abramowitz, I. A. Stegun (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series.* Washington, D.C.

[2] K. Bertin and G.Lecué (2008). Selection of variables and dimension reduction in high-dimensional nonparametric regression. *Electronic J. Statist.*, **2**, 1224–1241.

[3] C. Butucea and N. Stepanova (2015). Adaptive variable selection in nonparametric sparse additive models. *http://arxiv.org/abs/1508.06660*

[4] L. Comminges and A. S. Dalalyan (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40** (5), 2667–2696.

[5] C. R. Genovese, J. Jin, L. Wasserman, and Z. Yao. (2012). A comparison of the Lasso and Marginal Regression. *J. Mach. Learn. Res.*, **13**, 2107–2143.

[6] Yu. I. Ingster and N. A. Stepanova (2014). Adaptive variable selection in nonparametric sparse regression. *Journal of Mathematical Sciences*, **199**, 184–201.

[7] P. Ji, and J. Jin (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.*, **40** (1), 73–103.

[8] J.Jin, C.-H.Zhang, and Q.Zhang (2014) Optimality of graphlet screening in high dimensional variable selection. *J. of Machie Learning Research*, **15**, 2723–2772.

[9] J. Lafferty, and L. Wasserman (2008). Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.*, **36**, 28–63.

[10] K. Lounici (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic J. Statist.*, **2**, 90–102.

[11] N. Meinshausen and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34** (3), 1436–1462.

[12] N. Meinshausen and P. Bühlmann (2010). Stability selection. *J. Roy. Stat. Soc. Ser. B*, **72** (4), 417–473.

[13] P. Neuvial and E. Roquain (2012). On false discovery rate thresholding for classification under sparsity. *Ann. Statist.*, **40** (5), 2572–2600.

[14] L. Wasserman and K. Roeder (2009). High-dimensional variable selection. *Ann. Statist.*, **37** (5A), 2178–2201.

[15] M. Wainwright (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory*, **55** (5), 2183–2202.

[16] C.-H. Zhang (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38** (2), 894–942.

[17] A. Y. Zhang, H. H. Zhou (2015). Minimax rates of community detection in stochastic block models. *http://arxiv.org/abs/1507.05313*

[18] P. Zhao and B. Yu (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.