Electronic Journal of Statistics
Vol. 12 (2018) 3443–3472
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1457

# On the prediction loss of the lasso in the partially labeled setting

**Pierre C. Bellec[1,3], Arnak S. Dalalyan[1], Edwin Grappin[1] and Quentin Paris[1,2]**

[1] *3 avenue Pierre Larousse, ENSAE ParisTech - CREST, 92245 Malakoff, France*

[2] *National Research University - Higher School of Economics, 3 Kochnovskiy drive, 125319 Moscow, Russian Federation*

[3] *Rutgers University, Dept. of Statistics and Biostatistics,501 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA*

**Abstract:** In this paper we revisit the risk bounds of the lasso estimator in the context of transductive and semi-supervised learning. In other terms, the setting under consideration is that of regression with random design under partial labeling. The main goal is to obtain user-friendly bounds on the off-sample prediction risk. To this end, the simple setting of bounded response variable and bounded (high-dimensional) covariates is considered. We propose some new adaptations of the lasso to these settings and establish oracle inequalities both in expectation and in deviation. These results provide non-asymptotic upper bounds on the risk that highlight the interplay between the bias due to the mis-specification of the linear model, the bias due to the approximate sparsity and the variance. They also demonstrate that the presence of a large number of unlabeled features may have significant positive impact in the situations where the restricted eigenvalue of the design matrix vanishes or is very small.

**MSC 2010 subject classifications:** Primary 62H30; secondary 62G08.
**Keywords and phrases:** Semi-supervised learning, sparsity, lasso, oracle inequality, transductive learning, high-dimensional regression.

Received January 2018.

## 1. Introduction

We consider the problem of prediction under the quadratic loss. That is, for a random feature-label pair $(\boldsymbol{X}, Y)$ drawn from a distribution $P$ on a product space $\mathcal{X} \times \mathcal{Y}$, we aim at predicting $Y$ as a function of $\boldsymbol{X}$. The goal is to find a measurable function $f : \mathcal{X} \to \mathcal{Y}$ such that the expected quadratic risk,

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} (y - f(\boldsymbol{x}))^2 \, P(d\boldsymbol{x}, dy) = \mathbb{E}\big[\big(Y - f(\boldsymbol{X})\big)^2\big]$$

is as small as possible. When $\mathcal{Y}$ is an interval of $\mathbb{R}$ and $\mathcal{X}$ is a measurable set in $\mathbb{R}^p$—which is the setting considered in the present work—the Bayes predictor, defined as the minimizer of $\mathcal{R}(f)$ over all measurable functions $f : \mathcal{X} \to \mathcal{Y}$, is the regression function (Vapnik, 1998)

$$f^\star(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}].$$

Using $f^\star$, the problem can be rewritten in a form which is more familiar in Statistics, namely

$$Y = f^\star(\boldsymbol{X}) + \xi,$$

where the noise variable $\xi$ satisfies $\mathbb{E}[\xi|\boldsymbol{X}] = 0$, $P_X$-almost surely[1]. In the present work, we tackle the prediction problem in the case where the available data $\mathcal{D}_{\text{all}}$ is of the form $\mathcal{D}_{\text{all}} = \mathcal{D}_{\text{labeled}} \cup \mathcal{D}_{\text{unlabeled}}$, where

$$\mathcal{D}_{\text{labeled}} = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\} \quad \text{and} \quad \mathcal{D}_{\text{unlabeled}} = \{\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_N\}.$$

The labeled sample $\mathcal{D}_{\text{labeled}}$ is composed of independent and identically distributed (i.i.d.) feature-label pairs with distribution $P$. The unlabeled sample $\mathcal{D}_{\text{unlabeled}}$ contains only i.i.d. features, with distribution $P_X$, and is independent of $\mathcal{D}_{\text{labeled}}$. This formal setting accounts for a number of realistic situations in which the labeling process is costly while the unlabeled data points are available in abundance (see, for instance, Balcan et al., 2005; Guillaumin et al., 2010; Brouard et al., 2011), that is $n$ may be quite small compared to $N$. Here, the baseline idea is to build upon the sample $\mathcal{D}_{\text{unlabeled}}$ to improve the supervised prediction process based on $\mathcal{D}_{\text{labeled}}$ alone. In this context, our study encompasses two closely related settings: semi-supervised learning and transductive learning.

In the *semi-supervised learning* setting, one aims at constructing a predictor $\widehat{f}$, based on the data $\mathcal{D}_{\text{all}}$, such that the excess risk

$$\mathcal{E}(\widehat{f}) = \mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star) = \int_{\mathbb{R}^p} \big(\widehat{f}(\boldsymbol{x}) - f^\star(\boldsymbol{x})\big)^2 P_X(d\boldsymbol{x}) = \|\widehat{f} - f^\star\|^2_{L_2(P_X)} \quad (1)$$

is as small as possible. This learning framework differs from the classical supervised learning only in that the data set is enriched by the unlabeled features.

In contrast with this, the goal of *transductive learning* is to predict solely the labels of the observed unlabeled features. This amounts to considering the same setting as above but to measure the quality of a prediction function $f$ by the excess risk

$$\mathcal{E}_{\text{TL}}(f) = \frac{1}{N-n} \sum_{i=n+1}^{N} \big(f(\boldsymbol{X}_i) - f^\star(\boldsymbol{X}_i)\big)^2. \quad (2)$$

We refer the reader to (Chapelle et al., 2006; Zhu, 2008) and the references therein for a comprehensive survey on the topic of semi-supervised and transductive learning. Theoretical analysis of the generalisation error and the excess risk in this context can be found in (Rigollet, 2007; Wang and Shen, 2007; Lafferty and Wasserman, 2007), whereas the closely related area of manifold learning is studied in (Belkin et al., 2006; Nadler et al., 2009; Niyogi, 2013). The purpose of the present work differs from these papers in that we put the emphasis on the high-dimensional setting and the sparsity assumption. The goal is to understand whether the unlabeled data can help in predicting the unknown labels using the $\ell_1$-penalized empirical risk minimizers. From another

[1] Notation $P_X$ is used for the marginal distribution of $\boldsymbol{X}$.

perspective—that of multi-view learning—the problem of sparse semi-supervised learning is investigated in (Sun and Shawe-Taylor, 2010). During the reviewing process of the present work, two papers were posted on arxiv (Azriel et al., 2016; Chakrabortty and Cai, 2017), studying linear regression in semi-supervized setting. Some of the ideas used in these works are close to those of the present paper. However, their theoretical results are of different nature since they are asymptotic with fixed dimension and increasing sample size.

When the feature vector is high dimensional, it is reasonable to consider prediction strategies based on "simple" functions $f$ in order to limit the computational cost. A widely used approach is then to look for a good linear predictor

$$f_{\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}, \qquad \boldsymbol{\beta} \in \mathbb{R}^p.$$

When the dimension $p$ is of the same order as (or larger than) the size $n$ of the labeled sample, the simple empirical risk minimizer (*i.e.*, the least squares estimator) is a poor predictor since it suffers from the curse of dimensionality. To circumvent this shortcoming, one popular approach is to use the $\ell_1$-penalised empirical risk minimizer, also known as the lasso estimator (Tibshirani, 1996): $\widehat{f}^{\text{lasso}} = f_{\widehat{\boldsymbol{\beta}}^{\text{lasso}}}$ where[2]

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \|\boldsymbol{Y} - \mathbf{X}_{\text{lab}} \boldsymbol{\beta}\|_2^2 + 2\lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{3}$$

where $\lambda > 0$ stands for a tuning parameter and

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X}_{\text{lab}} = \begin{bmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{bmatrix}.$$

Statistical properties of the lasso with regard to the prediction error were studied in many papers, the most relevant (to our purposes) of which will be discussed in the next section. We also refer the reader to (Bühlmann and van de Geer, 2011) for an overview of related topics. The rationale behind this approach is that (a) the term $\frac{1}{n}\|\boldsymbol{Y} - \mathbf{X}_{\text{lab}}\boldsymbol{\beta}\|_2^2 - \mathbb{E}[\xi^2]$ is an unbiased estimator of the excess risk $\mathcal{E}(f_{\boldsymbol{\beta}})$ and (b) the $\ell_1$-penalty term favors predictors $f_{\boldsymbol{\beta}}$ defined via a (nearly) sparse vector $\boldsymbol{\beta}$.

The prediction rules we are going to analyze in the present work are suitable adaptations of the (supervised) lasso to the semi-supervised and the transductive settings. More precisely, we consider the estimator

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\text{lab}} \boldsymbol{\beta} + 2\lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{4}$$

where $\lambda > 0$ and $\mathbf{A} \in \mathbb{R}^{p \times p}$ are parameters to be chosen by the statistician. This definition is based on the following observation. The unlabeled sample may be used to get an improved estimator of the excess risk $\mathcal{E}(f_{\boldsymbol{\beta}}) =$

[2]To ease notation, we assume that both labels and features are centered, that is $\mathbb{E}[Y] = 0$ and $\mathbb{E}[\boldsymbol{X}] = 0$, so that there is no need to include an intercept in the linear combination $f_{\boldsymbol{\beta}}$.

$\mathbb{E}[f^\star(\boldsymbol{X})^2] - 2\mathbb{E}[Y\boldsymbol{X}^\top]\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}$, where $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ is the $p \times p$ covariance matrix. Indeed, the population covariance matrix can be estimated using both labeled and unlabeled data. A similar observation holds for the transductive excess risk $\mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}})$.

Denoting by $\widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}$ the empirical covariance matrix based on the labeled sample, that is

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^\top,$$

one checks that the vector $\widehat{\boldsymbol{\beta}}$ coincides with the lasso estimator (3) when $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}^{1/2}$. If an unlabeled sample is available, the foregoing discussion suggests a different choice for the matrix $\mathbf{A}$. This choice depends on the setting under consideration. Namely, defining the matrices

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{all}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i\boldsymbol{X}_i^\top \qquad \text{and} \qquad \widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}} = \frac{1}{N-n}\sum_{i=n+1}^{N} \boldsymbol{X}_i\boldsymbol{X}_i^\top,$$

we use $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}^{1/2}$ and $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}^{1/2}$ in the semi-supervised and transductive settings, respectively.

The following two assumptions made on the probability distribution $P$ will be repeatedly used throughout this work.

**(A1)** The random variables $Y$ and $\boldsymbol{X}$ have zero mean and finite variance. Furthermore, all the coordinates $X^j$ of the random vector $\boldsymbol{X}$ satisfy $\mathbb{E}[(X^j)^2] = 1$.

**(A2)** The random variables $Y$ and $X^j$ are almost surely bounded. That is, there exist constants $B_Y$ and $B_X$ such that $\mathbf{P}\big(|Y| \le B_Y; \max_{j\in[p]} |X^j| \le B_X\big) = 1$.

Assumption (A1) is fairly mild, since one can get close to it by centering and scaling the observed labels and features. For features, the centering and the scaling may be performed using the sample mean and the sample variance computed over the whole data-set. It is however important to require this assumption, since its violation may seriously affect the quality of the $\ell_1$-penalized least-squares estimator $\widehat{\boldsymbol{\beta}}$, unless the terms $|\beta_j|$ of the $\ell_1$-norm are weighted according to the magnitude of the corresponding feature $X^j$. The second assumption is less crucial both for practical and theoretical purposes, given that its primary aim is to allow for user-friendly, easy-to-interpret theoretical guarantees. In most situations, even if assumption (A2) is violated, the predictor $f_{\widehat{\boldsymbol{\beta}}}$ does have a fairly small prediction error rate.

The main contributions of the present work are:

- Review of the relevant recent literature on the off-sample performance of the lasso in the prediction problem.
- Non-asymptotic bounds for the prediction error of the lasso in the semi-supervised and transductive settings that guarantee the fast rate under the restricted eigenvalue condition. We did an effort for keeping the results easy to understand and to obtain small constants. These results are simple enough to be taught to graduate students.

- Oracle inequalities in expectation for the prediction error of the lasso. To the best of our knowledge, such results were not available in the literature until the very recent paper (Bellec et al., 2018+).

To give a foretaste of the results detailed in the rest of this work, let us state and briefly discuss a risk bound in the semi-supervised setting (the complete form of the result is provided in Theorem 7). For a matrix $\mathbf{A}$, we denote by $\|\mathbf{A}\|$ its largest singular value and by $\kappa_{\mathbf{A}}$ the compatibility constant (see Section 2 for a precise definition).

**Theorem.** *Let assumption (A1) be fulfilled and let the random variables $Y$, $X^j$ be bounded in absolute value by 1. For a prescribed tolerance level $\delta \in (0,1)$, assume that the overall sample size $N$ and the tuning parameter $\lambda$ satisfy $N \geq 18p\|\mathbf{\Sigma}^{-1}\| \log(3p/\delta)$ and*

$$\lambda \geq 4\left(\frac{2\log(6p/\delta)}{n}\right)^{1/2} + \frac{8\log(6p/\delta)}{3n}.$$

*Then, for every $J \subseteq \{1,\dots,p\}$, with probability at least $1-\delta$, the estimator $\widehat{\boldsymbol{\beta}}$ defined in (4) above with $\mathbf{A} = \widehat{\mathbf{\Sigma}}_{\mathrm{all}}^{1/2}$ satisfies*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2|J|}{2\kappa_{\widehat{\mathbf{\Sigma}}_{\mathrm{all}}}(J,3)}\right\}. \tag{5}$$

This result follows in the footsteps of many recent papers such as (Koltchinskii et al., 2011; Sun and Zhang, 2012; Dalalyan et al., 2014) among others. The term oracle inequality refers to the fact that it allows us to compare the excess risk of the predictor $f_{\widehat{\boldsymbol{\beta}}}$ to that of the best possible nearly sparse prediction function. (By nearly sparse we understand here a vector $\boldsymbol{\beta}$ such that for a set $J \subseteq \{1,\dots,p\}$ of small cardinality the entries of $\boldsymbol{\beta}$ with indices in $J^c$ have small magnitude; that is $\|\boldsymbol{\beta}_{J^c}\|_1 = \sum_{j\notin J}|\beta_j|$ is small.) Indeed, if we denote by $\bar{\boldsymbol{\beta}}$ a nearly $s$-sparse vector in $\mathbb{R}^p$ such that the excess risk $\mathcal{E}(f_{\bar{\boldsymbol{\beta}}})$ is small, then the aforestated risk bound is the sum of three terms having clear interpretation. The first term, $\mathcal{E}(f_{\bar{\boldsymbol{\beta}}})$, is a bias term due to the $s$-sparse linear approximation. The second term, $\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|$, is the bias due to approximate $s$-sparsity. (Note that it vanishes if $\bar{\boldsymbol{\beta}}$ is exactly $s$-sparse and $J$ is taken as its support.) Finally, the third term measures the magnitude of the stochastic error. Assuming the compatibility constant to be bounded away from 0, this last term is of the order $s\log(p)/n$, which is known to be optimal[3] over all possible estimators (Ye and Zhang, 2010; Raskutti et al., 2011; Rigollet and Tsybakov, 2011, 2012).

Inequality (5) readily shows the advantage of using the unlabeled data: the compatibility constant involved in the last term of the right hand side is computed for the overall covariance matrix. When the size of the labeled sample is small in regard to the dimension $p$, the corresponding constant computed for $\widehat{\mathbf{\Sigma}}_{\mathrm{lab}}$ may be very close (and even equal) to zero. This may downgrade the fast

[3]More precisely, the optimal rate is $\frac{s\log(1+p/s)}{n}$, which is of the same order as $\frac{s\log(p)}{n}$ for most values of $s$.

rate of the original lasso to the slow rate $\|\bar{\boldsymbol{\beta}}\|_1/\sqrt{n}$. Instead, if a large number of unlabeled features are used, it becomes more plausible to assume that the compatibility constant is bounded away from zero. In relation with this, it is important to underline that the unlabeled sample cannot help to improve the fast rate of convergence of the lasso, $s\log(p)/n$, which is optimal in the minimax sense. The best we can hope to achieve using the unlabeled sample is the relaxation of the conditions guaranteeing the fast rate. Another worthwhile remark is that the theorem stated above is valid when the size of the unlabeled sample is significantly larger than the dimension $p$. Interestingly, this condition is not required for getting the analogous result in the transductive set-up.

The rest is as follows. In Section 2, we introduce the notations used throughout the paper. Section 3 contains a review of the relevant literature and discusses the relation of the previous work with our results. Section 4 presents risk bounds for the prediction error of the lasso in the transductive setting, whereas Section 5 is devoted to the analogous results in the semi-supervised setting. Conclusions are made in Section 6. The proofs are postponed to Section 7. We also provide outlines of the proofs of Theorems 5 and 7 directly after stating these results.

## 2. Notations

In the sequel, for any integer $k$ we denote by $[k]$ the set $\{1,\dots,k\}$. For any $q \in [1,+\infty]$ the notation $\|\boldsymbol{v}\|_q$ refers to the $\ell_q$-norm of a vector $\boldsymbol{v}$ belonging to an Euclidean space $\mathbb{R}^k$ with arbitrary dimension $k$. Since there is no risk of confusion, we omit the dependence on $k$ in the notation. For any square matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ we denote by $\mathbf{A}^+$ its Moore-Penrose pseudoinverse and by $\|\mathbf{A}\|$ its spectral norm defined by

$$\|\mathbf{A}\| = \max_{\|\boldsymbol{v}\|_2=1} \|\mathbf{A}\boldsymbol{v}\|_2$$

We use boldface italic letters for vectors and boldface letters for matrices. Throughout the manuscript, the index $j$ will be used for referring to $p$ features, whereas the index $i$ will refer to the observations ($i \in [n]$ or $i \in [N]$). For any set of indices $J \subseteq [p]$ and any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_p)^\top \in \mathbb{R}^p$, we define $\boldsymbol{\beta}_J$ as the $p$-dimensional vector whose $j$-th coordinate equals $\boldsymbol{\beta}_j$ if $j \in J$ and 0 otherwise. We denote the cardinality of any $J \subseteq [p]$ by $|J|$. Also, we set $\operatorname{supp}(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta}_j \neq 0\}$. In particular, whenever $f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta}^\star$, we set $J^\star = \operatorname{supp}(\boldsymbol{\beta}^\star)$ and $s^\star = |J^\star|$. For $J \subseteq [p]$ and $c > 0$, we introduce the compatibility constants

$$\kappa_{\mathbf{A}}(J,c) = \inf\left\{ \frac{c^2|J|\ \|\mathbf{A}^{1/2}\boldsymbol{v}\|_2^2}{(c\|\boldsymbol{v}_J\|_1 - \|\boldsymbol{v}_{J^c}\|_1)^2} : \boldsymbol{v} \in \mathbb{R}^p, \|\boldsymbol{v}_{J^c}\|_1 < c\|\boldsymbol{v}_J\|_1 \right\}$$

and

$$\bar{\kappa}_{\mathbf{A}}(J,c) = \inf\left\{ \frac{|J|\ \|\mathbf{A}^{1/2}\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}_J\|_1^2} : \boldsymbol{v} \in \mathbb{R}^p, \|\boldsymbol{v}_{J^c}\|_1 < c\|\boldsymbol{v}_J\|_1 \right\}.$$

One easily checks that these two constants are of the same order of magnitude in the sense that

$$\frac{\bar{c}^2}{(\bar{c}+c)^2}\kappa_{\mathbf{A}}(J,\bar{c}+c) \le \bar{\kappa}_{\mathbf{A}}(J,c) \le \kappa_{\mathbf{A}}(J,c)$$

for every $c, \bar{c} > 0$. These constants are slightly larger[4] than the restricted eigenvalues (Bickel et al., 2009) defined by

$$\kappa^{\mathrm{RE}}_{\mathbf{A}}(J,c) = \inf\left\{\|\mathbf{A}^{1/2}\boldsymbol{v}\|_2^2 : \|\boldsymbol{v}_{J^c}\|_1 \le c\|\boldsymbol{v}_J\|_1 \text{ and } \|\boldsymbol{v}_J\|_2 = 1\right\}.$$

For more details, we refer the reader to van de Geer and Bühlmann (2009).

## 3. Brief overview of related work

The material of this paper builds on the shoulders of giants and this section aims at providing a unified overview of some of the most relevant results in our setting, without having the ambition of being exhaustive. For each of the selected papers, we will discuss its strengths and limitations in relation with the results presented further in this work.

Some recent results, obtained in the context of matrix regression, can be specialized to our problem and should be put in perspective with our contribution. For instance, a large part of Chapter 9 in (Koltchinskii, 2011) is devoted to the problem of assessing the off-sample excess risk of the trace-norm penalized empirical risk minimizer in the setting of trace regression with random design. One can arguably consider that setting as an extension of the random design regression problem by restricting attention to the set of diagonal matrices. Then the estimator studied in Koltchinskii (2011) coincides with the lasso estimator (3). With our notations, the main result of Chapter 9 in (Koltchinskii, 2011) reads as follows.

**Theorem 1** (Theorem 9.3 in Koltchinskii, 2011)**.** *Assume that Assumptions (A1) and (A2) hold. Then there exist universal positive constants $c_1$ and $c_2$ such that, if*

$$\lambda \ge c_1 B_X \max\left\{\frac{B_Y \log(2p/\delta)}{n}, \left(\frac{B_Y \log(2p/\delta)}{n}\right)^{1/2}\right\}$$

*for some $\delta \in (0,1)$, the estimator (3) satisfies,*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \le \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{2\mathcal{E}(f_{\boldsymbol{\beta}}) + c_2\left[\frac{\|\boldsymbol{\beta}\|_0\lambda^2}{\bar{\kappa}_{\boldsymbol{\Sigma}}(\mathrm{supp}(\boldsymbol{\beta}),5)} + \left(\|\boldsymbol{\beta}\|_1 \vee \frac{q(\lambda)}{\lambda}\right)^2 \frac{\log(k/\delta)\log(n)}{n} + \frac{1}{n}\right]\right\},$$

*with probability larger than $1-\delta$, where*

$$k = \log(n \vee p \vee B_Y) \vee |\log(2\lambda)| \vee 2 \quad \textit{and} \quad q(\lambda) = \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\mathcal{E}(f_{\boldsymbol{\beta}}) + 2\lambda\|\boldsymbol{\beta}\|_1\right).$$

[4] We recall here that a larger compatibility constant provides a better risk bound.

This result can be briefly compared to the risk bound in (5). The main advantages of this result is that (a) it is established under much weaker assumptions on the boundedness of the random variables $\boldsymbol{X}$ and $Y$ than those of Assumption (A2), (b) it holds not only for the vector regression but also for matrix regression, (c) it contains no restriction on the sample size and (d) it involves the compatibility constant of the population covariance matrix $\boldsymbol{\Sigma}$. On the negative side, the oracle inequality in Theorem 1 is not sharp since the factor in front of $\mathcal{E}(f_{\boldsymbol{\beta}})$ is not equal to one and, more importantly, the rate of convergence of the remainder term is sub-optimal in most situations. Indeed, if the best linear predictor corresponds to an $s$-sparse vector the nonzero entries of which are all of the same order, then the term $\|\boldsymbol{\beta}\|_1^2 \log(k/\delta)\log(n)/n$, present in the right hand side, is of order $s^2 \log(n)\log\log(n+p)/n$, whereas the remainder term in (5) is of smaller order $s\log(p)/n$.

On a related note, Koltchinskii et al. (2011) establish sharp oracle inequalities for the trace-norm penalized least-squares estimator in the problem of matrix estimation and completion under low rank assumption. Using our notation, Theorem 2 in (Koltchinskii et al., 2011) yields the following result.

**Theorem 2** (Koltchinskii et al., 2011)**.** *Assume that the matrix* $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ *is known and let* $\widehat{\boldsymbol{\beta}}$ *be as in* (4) *with* $\mathbf{A} = \boldsymbol{\Sigma}^{1/2}$. *Suppose in addition that Assumption (A2) holds and that, for* $\delta \in (0,1)$,

$$\lambda \geq 4B_Y\Big(\frac{\log(p/\delta)}{n}\Big)^{1/2}\Big[1 + \frac{B_X}{3}\Big(\frac{\log(p/\delta)}{n}\Big)^{1/2}\Big].$$

*Then, with probability larger than* $1-\delta$, *we have*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{J\subseteq[p]}\inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2|J|}{4\kappa_{\boldsymbol{\Sigma}}(J,3)}\Big\}.$$

The original result (Koltchinskii et al., 2011, Theorem 2) is slightly different from the aforestated one. In particular, it is expressed in terms of the restricted eigenvalue constant with respect to the population covariance matrix $\boldsymbol{\Sigma}$. However, all these differences imply only minor modifications in the proofs. Theorem 2 is very similar to the risk bounds that we establish in the present work, but has the obvious shortcoming of requiring the covariance matrix $\boldsymbol{\Sigma}$ to be known. In fact, this corresponds to the situation in which infinitely many unlabeled feature vectors $\boldsymbol{X}_{n+1}, \boldsymbol{X}_{n+2}, \ldots$ are available, that is $N = +\infty$. To some extent, one of the purposes of the present work is to provide risk bounds analogous to the result of Theorem 2 but valid for a broad range of values of $N$. Note that the choice of the tuning parameter $\lambda$ advocated by all the aforementioned results is of the same order of magnitude.

To the best of our knowledge, the only paper establishing risk bounds for a transductive version of the lasso is (Alquier and Hebiri, 2012). In that paper, the authors considered the problem of transductive learning in a linear model $Y = \boldsymbol{X}^\top\boldsymbol{\beta}^\star + \xi$ under the sparsity constraint. The estimator they studied is

slightly different from ours and is defined by

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}^{1/2}\boldsymbol{\beta}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}^{+}\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}. \tag{6}$$

For the predictor $f_{\widehat{\boldsymbol{\beta}}}$ based on this estimator, the authors established the following risk bound.

**Theorem 3** (Theorems 4.3 and 4.4 in Alquier and Hebiri, 2012)**.** *Assume that for some $\boldsymbol{\beta}^\star \in \mathbb{R}^p$, the conditional distribution of $\xi := Y - \boldsymbol{X}^\top\boldsymbol{\beta}^\star$ given $\boldsymbol{X}$ is Gaussian $\mathcal{N}(0,\sigma^2)$. Let $\mathscr{E}_1$ be the event "all the unlabeled features $\{\boldsymbol{X}_{n+i} : i \in [N-n]\}$, belong to the linear span of the labeled features $\{\boldsymbol{X}_i : i \in [n]\}$" and let $\delta \in (0,1)$. Denote by $a_{n,N,p}$ the harmonic mean of the diagonal entries of the matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}^{+}\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}$. Then the estimator (6) with $\lambda = \sigma\sqrt{(2/n)a_{n,N,p}\log(p/\delta)}$ satisfies*

$$\mathbf{P}\left( \mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \le \frac{72\sigma^2 a_{n,N,p}}{\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}}(J^\star,3)} \cdot \frac{s^\star \log(p/\delta)}{n} \;\middle|\; \mathbf{X}_{\mathrm{all}} \right) \ge 1-\delta \quad \textit{on} \quad \mathscr{E}_1.$$

This result is close in spirit to the result that we establish in this work in the setting of transductive learning. Note however that there are three main differences. First, we do not confine our study to the well-specified situation in which the Bayes predictor is linear, $f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta}^\star$ for every $\boldsymbol{x} \in \mathbb{R}^p$, with a sparse vector $\boldsymbol{\beta}^\star$. Second, we avoid the unpleasant restriction that the unlabeled features are linear combinations of labeled features. Third, we replace the factor $a_{n,N,p}$—which may be quite large—by a more tractable quantity. This being said, the result of Alquier and Hebiri (2012)—in contrast with our results—does not require the unlabeled features to be drawn from the same distribution as the labeled features.

We also review a recent result from (Lecué and Mendelson, 2016). In that paper, the authors consider the isotropic case $\boldsymbol{\Sigma} = \mathbf{I}_p$, where $\mathbf{I}_p$ stands for the $p \times p$ identity matrix, but impose only weak assumptions on the moments of the noise. Translated to our notations, their result can be formulated as follows.

**Theorem 4** (Theorem 1.3 in Lecué and Mendelson, 2016)**.** *Let Assumption (A2) be satisfied and let $\boldsymbol{\Sigma} = \mathbf{I}_p$. Let $f_{\bar{\boldsymbol{\beta}}}$ be the best linear approximation in $L^2(P_X)$ of the regression function $f^\star$, that is $\bar{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{E}(f_{\boldsymbol{\beta}})$. Let $\delta \in (0,1)$ be a prescribed tolerance level. There are three constants $c_1(\delta)$, $c_2(\delta, B_X)$ and $c_3(\delta, B_X)$ such that, if $\bar{\boldsymbol{\beta}}$ is nearly $s$-sparse in the sense that*[5]

$$\sum_{j=s+1}^{p} |\bar{\beta}|_{(j)} \le c_1(\delta) B_Y s\Big(\frac{\log(2p)}{n}\Big)^{1/2}$$

*and $\lambda$ is chosen by $\lambda = c_2(\delta, B_X)B_Y\big(\frac{\log(2p)}{n}\big)^{1/2}$, then with probability at least $1-\delta$ the lasso estimator satisfies*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \le \mathcal{E}(f_{\bar{\boldsymbol{\beta}}}) + c_3(\delta, B_X) B_Y^2 \, \frac{s\log(2p)}{n}.$$

[5]We denote by $|\bar{\beta}|_{(j)}$ the $j$-th largest value of the sequence $|\bar{\beta}_1|, \dots, |\bar{\beta}_p|$, so that $|\bar{\beta}|_{(1)} \ge \cdots \ge |\bar{\beta}|_{(p)}$.

The principal strength of this result is that it is valid under a very weak assumption on the tails of the noise, but it has the shortcoming of requiring the minimizer of the excess risk to be nearly $s$-sparse with a quite precise upper bound on the authorized non-sparsity bias. From this point of view, an upper bound of the form (5) provides more information on the robustness of the prediction rule with respect to the model mis-specification.

The proofs of the results above assess the off-sample prediction error rate of the lasso by using direct arguments. An alternative approach (adopted, for example, in Raskutti et al., 2010; Koltchinskii, 2011; Oliveira, 2013; Rudelson and Zhou, 2013) consists in taking advantage of the in-sample risk bounds in order to assess the off-sample excess risk. In short, by means of nowadays well-known techniques (developed in Bickel et al., 2009; Juditsky and Nemirovski, 2011; Bühlmann and van de Geer, 2011; Belloni et al., 2014; Dalalyan et al., 2014, for instance) for a well-specified model[6], an upper bound on the in-sample risk,

$$\frac{1}{n}\|\mathbf{X}_{\mathrm{lab}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star)\|_2^2 = \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star)\|_2^2,$$

is obtained along with proving that the vector $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star$ belongs to the dimension-reduction cone appearing in the definition of the compatibility constant. Then, using suitably chosen concentration arguments, it is shown that (with high probability) the compatibility constant $\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}}(J^\star, c)$ of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}$ is lower bounded by a (multiple of a) compatibility constant $\kappa_{\boldsymbol{\Sigma}}(J^\star, c')$ of the population covariance matrix, provided that the sparsity $s$ is of order $n/\log(p)$. The main conceptual differences between the aforementioned papers are in the conditions on the random vectors $\boldsymbol{X}_i$. In (Raskutti et al., 2010), it is assumed that the $\boldsymbol{X}_i$'s are Gaussian. In Rudelson and Zhou (2013) and Theorem 9.2 in Koltchinskii (2011), sub-Gaussian and bounded designs are considered, whereas only a bounded moment condition is required in Oliveira (2013). We will not reproduce their results here because (a) they do not allow to account for the robustness to the model mis-specification and, to a lesser extent, (b) the constants involved in the bounds are not explicit.

## 4. Risk bounds in transductive setting

We first consider the case of transductive learning. From an intuitive point of view, this case is simpler than the case of semi-supervised learning since a prediction needs to be carried out only for the features in $\mathcal{D}_{\mathrm{unlabeled}}$. Indeed, recall from (2) that in this context, the excess risk of the linear predictor $f_{\boldsymbol{\beta}}$ is defined by

$$\mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) = \frac{1}{N-n}\sum_{i=n+1}^{N}\left(\boldsymbol{X}_i^\top\boldsymbol{\beta} - f^\star(\boldsymbol{X}_i)\right)^2$$

[6]This means that for a sparse vector $\boldsymbol{\beta}^\star$, it holds that $f^\star = f_{\boldsymbol{\beta}^\star}$.

and the suitably adapted lasso estimator is given by choosing $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}$ in (4), that is

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \|\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}\boldsymbol{\beta}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top \mathbf{X}_{\text{lab}}\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1 \Big\}.$$

Note here that the role of the term $\frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\text{lab}}$ is to estimate the term $\frac{2}{N-n}\sum_{i=n+1}^N f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i^\top$, which appears after developing the square in the excess risk. Since the latter belongs to the image of the matrix $\mathbf{X}_{\text{unlab}}$, one can slightly improve the estimator by projecting onto the subspace of $\mathbb{R}^p$ spanned by the unlabeled vectors $\mathbf{X}_i$. This amounts to replacing the term $\boldsymbol{Y}^\top\mathbf{X}_{\text{lab}}\boldsymbol{\beta}$ by $\boldsymbol{Y}^\top\mathbf{X}_{\text{lab}}\Pi_{\text{unlab}}\boldsymbol{\beta}$, where $\Pi_{\text{unlab}}$ stands for the orthogonal projector in $\mathbb{R}^p$ onto $\text{Span}(\boldsymbol{X}_{n+1},\dots,\boldsymbol{X}_N)$. However, from a theoretical point of view, this modification has no impact on the risk bound stated below. That is why we confine our attention to the lasso estimator that does not use this modification.

**Theorem 5.** *Let Assumptions (A1) and (A2) be fulfilled. Define $n_\star = n\wedge(N-n)$ and assume that, for a given $\delta\in(0,1)$, the tuning parameter $\lambda$ satisfies*

$$\lambda \geq 4B_Y\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big]. \tag{7}$$

*Then, with probability at least $1-\delta$, the predictor $f_{\widehat{\boldsymbol{\beta}}}$ satisfies*

$$\mathcal{E}_{\text{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p\\ J\subseteq[p]}} \Big\{ \mathcal{E}_{\text{TL}}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2|J|}{4\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}}(J,3)} \Big\}. \tag{8}$$

A few comments are in order. First, Theorem 5 holds for any pair of integers $n$ and $N$ larger than 1. However, it is especially relevant when the number $N-n$ of unlabeled features is larger than the number $n$ of labeled ones. As already mentioned, this kind of situation is frequent in applications where the labeling procedure is expensive. In this case, $n_\star = n$ and Theorem 5 takes the same form as (5) with the notable advantage that the size of the unlabeled sample does not need to be of larger order than the dimension $p$. Let us provide a roadmap of the proof, before presenting a few implications of this result in the well-specified case.

*Roadmap of the proof of Theorem 5.* We start the proof by applying the basic inequality of Lemma 1, which captures the convexity of the optimization problem (4). This yields that for any $\boldsymbol{\beta}$,

$$\mathcal{E}_{\text{TL}}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}_{\text{TL}}(f_{\boldsymbol{\beta}}) \leq 2\boldsymbol{\zeta}^\top(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) - \|\mathbf{A}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\|_2^2,$$

where $\boldsymbol{\zeta}$ is a noise term defined as a sum of iid zero mean random vectors (see Proposition 1 for the precise definition). Thanks to the Bernstein inequality, condition (7) is sufficient to prove that the event $\{\|\boldsymbol{\zeta}\|_\infty \leq \lambda/2\}$ has a probability at least $1-\delta$. On this event, $2\boldsymbol{\zeta}^\top(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \leq \lambda\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_1$ and the simple algebra described in Lemma 2 lets us bound the right hand side of the previous display by $4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9}{4}\lambda^2|J|/\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}}(J,3)$ for any $J\subset[p]$. □

**Well-specified case.** Recall that the well-specified case refers to the situation where there exists $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ such that the Bayes predictor $f^\star$ satisfies $f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^\star$, $P_X$-almost surely. In this case, the excess risk of a predictor $f_{\boldsymbol{\beta}}$ can be written as $\mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) = \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2$. In this form, the technical tractability of the transductive learning problem appears clearly since the matrix $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}^{1/2}$ used in the definition of the estimator $\widehat{\boldsymbol{\beta}}$ coincides with the one appearing in the excess loss. As we shall see later, this is indeed not the case for semi-supervised learning. Now, the choice of $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ and $J = J^\star$ in the right hand side of inequality (8) yields

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \frac{9\lambda^2 s^\star}{4\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}}(J^\star, 3)}.$$

The choice of $\lambda$ provided by the right hand side of inequality (7), along with the condition $n_\star \geq B_X^2 \log(2p/\delta)$, leads to the bound

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \frac{64 B_Y^2}{\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}}(J^\star, 3)} \cdot \frac{s^\star \log(p/\delta)}{n_\star},$$

with probability at least $1 - \delta$. Comparing our result with that of Alquier and Hebiri (2012) (cf. Theorem 3 above), we can note that Theorem 5 holds without the assumption that the unlabeled features belong to the linear span of the labeled ones. On the other hand, Alquier and Hebiri (2012) do not require the labeled and the unlabeled features to be drawn from the same distribution.

## 5. Risk bounds in semi-supervised setting

We now turn to the more challenging problem of semi-supervised learning. In this subsection, we first consider the well-specified setting in which the Bayes predictor $f^\star$ is linear. We start with risk bounds that hold with a probability close to one. Such bounds are often termed *in deviation* as opposed to those holding *in expectation*.

**Well-specified case.** We assume here that

$$f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^\star, \quad P_X\text{-almost surely.} \tag{9}$$

In this context, the excess risk of the linear predictor $f_{\boldsymbol{\beta}}$, defined in (1), becomes $\mathcal{E}(f_{\boldsymbol{\beta}}) = \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2$. This setting is more restrictive than the mis-specified setting considered below, but it has the advantage of allowing us to obtain risk bounds that are small even if the sample size $N$ is not necessarily larger than the dimension $p$. The next result assesses the performance of the predictor $f_{\widehat{\boldsymbol{\beta}}}$ where

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}^{1/2} \boldsymbol{\beta}\|_2^2 - \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} \boldsymbol{\beta} + 2\lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{10}$$

corresponding to the choice $\mathbf{A} = \widehat{\mathbf{\Sigma}}_{\text{all}}^{1/2}$ in (4). In the next result, we set

$$\kappa_{\mathbf{A}}^{\text{RE}}(s, c) = \min_{J \subseteq [p] : |J| \leq s} \kappa_{\mathbf{A}}^{\text{RE}}(J, c),$$

where the restricted eigenvalue $\kappa_{\mathbf{A}}^{\text{RE}}(J, c)$ is defined in Section 2.

**Theorem 6.** *Let Assumptions (A1), (A2) and (9) be fulfilled. Let $\delta \in (0, 1)$ be a tolerance level and let the tuning parameter $\lambda$ satisfy*

$$\lambda \geq 4B_Y \Big( \frac{\log(4p/\delta)}{n} \Big)^{1/2} \Big[ 1 + \frac{B_X}{2} \Big( \frac{\log(4p/\delta)}{n} \Big)^{1/2} \Big].$$

*With probability at least $1 - \delta$, it holds*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \left( \frac{6\lambda s^\star}{\bar{\kappa}_{\widehat{\mathbf{\Sigma}}_N}(J^\star, 3)} \right)^2 \bigwedge \frac{9\|\mathbf{\Sigma}\|\lambda^2 s^\star}{\kappa^{\text{RE}}_{\widehat{\mathbf{\Sigma}}_N}(s^\star, 3)^2}. \tag{11}$$

*In addition, if the overall sample size $N$ is such that $16 s^\star B_X^2 \sqrt{2\log(4p^2/\delta)} \leq \bar{\kappa}_{\mathbf{\Sigma}}(J^\star, 3)\sqrt{N}$ then, with probability at least $1 - \delta$, the predictor $f_{\widehat{\boldsymbol{\beta}}}$ satisfies the inequality*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \frac{9\lambda^2 s^\star}{\bar{\kappa}_{\mathbf{\Sigma}}(J^\star, 3)}. \tag{12}$$

This theorem provides three different risk bounds, all of them being valid for the same choice of the tuning parameter $\lambda$, that clearly show the benefits of using unlabeled data. The first two bounds are stated in eq. (11). They share the common feature of depending on a characteristic (compatibility constant or restricted eigenvalue) of the sample covariance matrix. The latter is computed using both labeled and unlabeled data. For large values of $N$, it is more likely that these characteristics are bounded away from zero than those of the sample covariance matrix based on the labeled data only. In the asymptotic setting where $s^\star$ goes to infinity with the sample size and the dimension, the second term in the right hand side of eq. (11) is of smaller order than the first one and is rate optimal, provided that the restricted eigenvalue is lower bounded by a fixed positive constant. However, for finite and small values of $s^\star$ the first term in the right hand side of eq. (11) might be smaller than the second term.

This being said, it might be more insightful to look at the non random upper bounds on the excess risk as the one stated in eq. (12). It basically tells us that if the overall sample size $N$ is larger than a multiple of $(s^\star)^2 \log p$, then the off-sample prediction risk of the semi-supervised lasso estimator achieves the fast rate $\frac{s^\star \log p}{n}$. Note that if we use only the labeled data points, the best known results—as recalled in Section 2 above—provide the fast rate when $n$ is larger than a multiple of $s^\star \log p$. Thus, if $N$ is of the same order as $n$, our result above is not the sharpest possible, but it has the advantage of being easy to prove and, nevertheless, demonstrating the gain of using the unlabeled data. In particular, the proof of results providing the fast rate under the condition

$n \geq Cs^\star \log(p)$, for some $C > 0$, involve the important step of lower bounding the compatibility constant of the sample covariance matrix by its population counterpart. This step uses concentration arguments which are often tedious and come with implicit (or unreasonably large) constants. Instead, our proof makes use of much simpler tools essentially boiling down to the classical Bernstein inequality and leads to explicit and small constants.

**Mis-specified case.** Mathematical analysis of the semi-supervised lasso under mis-specification is more involved, since it requires careful control of the bias terms corresponding to the nonlinearity and the non-sparsity of the model. We first state results providing risk bounds in deviation, then state their counterpart in expectation.

**Theorem 7.** *Let Assumptions (A1) and (A2) be fulfilled. Fix $J \subseteq [p]$ and $\delta \in (0,1)$. Suppose in addition that*

$$N \geq 18 B_X^2 p \|\boldsymbol{\Sigma}^{-1}\| \log(3p/\delta) \tag{13}$$

*and*

$$\lambda \geq 8 B_X B_Y \Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2} \Big[1 + \frac{B_X}{3}\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\Big]. \tag{14}$$

*Then the semi-supervised lasso estimator $\widehat{\boldsymbol{\beta}}$ defined in (10) above satisfies*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda \|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{2\kappa_{\widehat{\boldsymbol{\Sigma}}_{\rm all}}(J,3)} \Big\}, \tag{15}$$

*with probability larger than $1-\delta$.*

*Roadmap of the proof of Theorem 7.* We start the proof by applying the basic inequality of Lemma 1, which leverages the convexity of the optimization problem (10) and implies that for any $\boldsymbol{\beta}$,

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) \leq \underbrace{2\boldsymbol{u}^\top(\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1)}_{\mathbf{T}_1} + \underbrace{\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} - \|\mathbf{A}\boldsymbol{u}\|_2^2}_{\mathbf{T}_2}.$$

In this inequality, $\boldsymbol{u} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, and for each $i = 1, 2$, vector $\boldsymbol{\zeta}^{(i)}$ is a sum of iid random vectors with zero mean (see Proposition 1 for a precise definition). Similarly to the proof of Theorem 5, condition (14) is sufficient to prove that, thanks to the Bernstein inequality, the event $\{\|\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\|_\infty \leq \lambda/2\}$ has a probability at least $1 - \delta/2$. However, in the semi-supervised setting, the term $\mathbf{T}_2$ appears because the matrix $\mathbf{A}^\top\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\rm all}$ is not equal to the true covariance $\boldsymbol{\Sigma}$ from the excess risk. Using matrix concentration inequalities, condition (13) is sufficient to prove that the event $\{\mathbf{T}_2 \leq -\frac{1}{2}\|\mathbf{A}\boldsymbol{u}\|^2\}$ has a probability at least $1 - \delta/2$. On the intersection of these two events, the simple algebra described in Lemma 2 lets us bound the right hand side of the previous display to obtain (15). □

The novelty of Theorem 7 lies in the semi-supervised nature of the estimator (10), which involves all the unlabeled features through the matrix $\mathbf{A} = \widehat{\mathbf{\Sigma}}_{\text{all}}^{1/2}$ in eq. (4). In particular, Theorem 7 quantifies the natural intuition according to which, if $N$ is large enough, the matrix $\mathbf{A} = \widehat{\mathbf{\Sigma}}_{\text{all}}^{1/2}$ is a good estimator of $\mathbf{\Sigma}$ and a result similar to Theorem 2 should hold. As mentioned in the introduction, an attractive feature of the upper bound in eq. (15) is that it is of the same form as the recent oracle inequalities established in the case of fixed design regression (see, for instance, Dalalyan et al., 2014; Pensky, 2014, and the references therein) and quantifies in an easy-to-understand manner the error terms accounting for the non-linearity and the non-sparsity of the true regression function $f^\star$.

Let us mention that there is a difference between the formulations of Theorem 7 and Theorem 5. Indeed, the latter contains an infimum over $J$ in the upper bound, whereas the former does not have this infimum but is valid for every $J$. It turns out that in the semi-supervised setting it is harder to get a good upper bound than in the transductive setting. The difference between the two statements reflects this difficulty. In fact, the oracle inequality of Theorem 5 (transductive setting) holds true on an event which has a probability at least $1-\delta$ and the definition of which does not rely on a specific choice of $J$. This is why we can put the inf over $J$ inside the probability. On the other side, in Theorem 7, we are not able to exhibit a unique event of probability at least $1-\delta$ for all the sets $J$ for which the oracle inequality is true. What we succeed to prove is that for every $J$, the stated inequality is true on an event $\mathscr{E}_J$ of probability at least $1-\delta$. Unfortunately, the dependence of the event on $J$ prevents us from moving the inf within the probability.

The minimal number $N$ of features satisfying (13) depends on $\|\mathbf{\Sigma}^{-1}\| = \lambda_{\min}^{-1}(\mathbf{\Sigma})$, reflecting the fact that the quality of approximation of the identity matrix $\mathbf{I}_p$ by $\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Sigma}}_{\text{all}}\mathbf{\Sigma}^{-1/2}$ depends on $\|\mathbf{\Sigma}^{-1}\|$. One can remark that under constraint (13), the lowest eigenvalue of the sample covariance matrix is close to its population counterpart (Vershynin, 2010) and provides a simple lower bound on the compatibility constant $\kappa_{\widehat{\mathbf{\Sigma}}_{\text{all}}}(J,3)$ appearing in eq. (15). These considerations lead to the following corollary.

**Corollary 1.** *Under the conditions of Theorem 7, with probability at least* $1-\delta$,

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \le \inf_{J\subseteq[p]} \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \tfrac{27\|\mathbf{\Sigma}^{-1}\|}{4}\,\lambda^2|J| \Big\}.$$

Let us also mention that the factor $B_X^2 p\|\mathbf{\Sigma}^{-1}\|$ present in the right hand side of eq. (13) is an upper bound on the norm $\|\mathbf{\Sigma}^{-1/2}\boldsymbol{X}_i\|_2^2$ under assumption (A2). Under additional assumptions on the support of the features $\boldsymbol{X}_i$, this expression may be replaced by a smaller one leading thus to a relaxation of condition (13).

**Sharp oracle inequality in expectation.** All the previously stated results assert that the lasso estimator has a small prediction error on an event of overwhelming probability. However, in these results, the choice of the tuning parameter $\lambda$ and, therefore, the final predictor $f_{\widehat{\boldsymbol{\beta}}}$, depends on the prescribed level of tolerance. A consequence of this dependence is that one can not integrate out

the bounds in deviation in order to get a bound in expectation. This is probably one of the reasons why the bounds in expectation for the lasso are scarce in the literature. To fill this caveat, we state below a risk bound in expectation that can be easily deduced from the bounds in deviation.

**Theorem 8.** *Let Assumptions (A1) and (A2) be fulfilled. Suppose that the overall sample size is such that* $N \geq 18B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|\log(3pN^2)$. *Then, for the tuning parameter*

$$\lambda = 8B_X B_Y\Big(\frac{\log(6pN^2)}{n}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{\log(6pN^2)}{n}\Big)^{1/2}\Big]$$

*the semi-supervised lasso estimator* $\widehat{\boldsymbol{\beta}}$ *defined in* (10) *above satisfies*

$$\mathbb{E}[\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}})] \leq \inf_{J\subseteq[p]}\inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{27\|\boldsymbol{\Sigma}^{-1}\|}{4}\lambda^2|J|\Big\} + \frac{2B_Y^2}{N^2} + \frac{B_Y^2}{2^7 n\log^2(6pN^2)}.$$

The proof of this theorem is postponed to section 7.2.3. The bound above is not optimal in terms of its dependence on $N$. In particular, it blows up when $N$ goes to infinity and all the other parameters are fixed. However, this divergence is only logarithmic in $N$. The dominating term in the risk bound above is (at least in the well specified setting) of the order $\lambda^2|J| \asymp \frac{s\log(pN)}{n}$.

## 6. Conclusion

We have reviewed some recent results on the prediction accuracy of the lasso in the problem of regression with random design and have proposed their extensions to the setting where the labels of some data points are not available. Theoretical guarantees stated in previous sections are formulated as oracle inequalities that allow us to compare the excess risk of a suitable adaptation of the lasso to the best possible (nearly) sparse prediction function. We have opted for considering only those risk bounds that provide the fast rate and are valid under some conditions on the design such as the restricted eigenvalue condition or the compatibility condition. Some of the established upper bounds involve the compatibility constant of the sample covariance matrix. Using results on random matrices (Rudelson and Zhou, 2013; Oliveira, 2013; Bah and Tanner, 2014) they can be further worked out to get deterministic upper bounds. However, the evaluation of the restricted eigenvalues and related quantities of the random covariance-type matrices is a dynamically evolving research area and we expect that new advances will be made in near future.

The main high level message of the contributions of this paper is that one can take advantage of the unlabeled sample for improving the prediction accuracy of the lasso. Roughly speaking, if the size of the unlabeled sample is larger than the ambient dimension, then the modified lasso predictor has a prediction risk

that converges to zero at the optimal rate even if the sample covariance matrix based only on the labeled sample does not satisfy the compatibility or the restricted eigenvalue condition. However, it should be acknowledged that when the model is well specified (that is there exists a sparse linear combination of the features with an extremely low approximation error) and the population covariance matrix is well-conditioned, then the original lasso might perform as well as, or even better than, the modified lasso proposed in this work. Therefore, one can conclude that the use of the unlabeled sample improves on the robustness of the lasso to the model mis-specification.

We would like also to emphasize that, pursuing pedagogical goals, we have restricted our attention to the simple case of bounded feature vectors and bounded labels. All the proofs presented in this paper are based on elementary arguments and are fairly simple. Using more involved arguments, they can be carried over the case of sub-Gaussian design and labels. It would be interesting to explore their extensions to other settings such as regression with structured sparsity, low rank matrix regression or matrix completion, *etc.*

## 7. Proofs

We start with a general result that holds for the penalized least squares predictor with arbitrary convex penalty. This result is of independent interest. It generalizes the corresponding result of (Koltchinskii et al., 2011) established for the matrix trace-norm penalties. The proof that we present here is different from the one in (Koltchinskii et al., 2011) in that it does not rely on the precise form of the sub-differential of the penalty function.

**Lemma 1.** *Let $n, p \geq 1$. Let* $\mathrm{pen} : \mathbb{R}^p \to \mathbb{R}$ *be any convex function and* $\widehat{\boldsymbol{\beta}}$ *be defined by*

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} \boldsymbol{\beta} + \mathrm{pen}(\boldsymbol{\beta}) \right\},$$

*where* $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$ *and* $\mathbf{X}_{\mathrm{lab}} \in \mathbb{R}^{n \times p}$. *Then, for all* $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 \leq \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\widehat{\boldsymbol{\beta}}) - \|\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2. \tag{16}$$

*Proof.* Let us introduce the function $\Phi(\boldsymbol{\beta}) = \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} \boldsymbol{\beta} + \mathrm{pen}(\boldsymbol{\beta})$ for every $\boldsymbol{\beta} \in \mathbb{R}^p$, so that $\widehat{\boldsymbol{\beta}}$ is a minimum point of $\Phi$. Since the latter is a convex function, we know that the zero vector $\mathbf{0}_p$ of $\mathbb{R}^p$ belongs to the sub-differential $\partial\Phi(\widehat{\boldsymbol{\beta}})$ of $\Phi$ at $\widehat{\boldsymbol{\beta}}$. For all $\boldsymbol{\beta} \in \mathbb{R}^p$, let

$$\psi(\boldsymbol{\beta}) = \|\mathbf{A}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\|_2^2, \qquad \bar{\Phi}(\boldsymbol{\beta}) = \Phi(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta}).$$

The function $\psi$ is proper and convex. It is also differentiable on $\mathbb{R}^p$ and the sub-differential of $\psi$ at $\widehat{\boldsymbol{\beta}}$ is reduced to its gradient at $\widehat{\boldsymbol{\beta}}$, so that $\partial\psi(\widehat{\boldsymbol{\beta}}) = \{\nabla\psi(\widehat{\boldsymbol{\beta}})\} = \{\mathbf{0}_p\}$. The function $\bar{\Phi}$ defined on $\mathbb{R}^p$ is the sum of an affine function

and the convex function pen, thus it is also convex. The functions $\psi, \bar{\Phi}$ are proper and convex, the function $\psi$ is continuous on $\mathbb{R}^p$ so by the Moreau-Rochafellar Theorem,

$$\partial\Phi(\widehat{\boldsymbol{\beta}}) = \partial\psi(\widehat{\boldsymbol{\beta}}) + \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}}) = \{\mathbf{0}_p\} + \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}}) = \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}}).$$

Thus $\mathbf{0}_p \in \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}})$, which can be rewritten as

$$\bar{\Phi}(\boldsymbol{\beta}) \geq \bar{\Phi}(\widehat{\boldsymbol{\beta}}), \qquad \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

By adding $\psi(\boldsymbol{\beta})$ on both sides of the previous display, we obtain

$$\Phi(\boldsymbol{\beta}) \geq \Phi(\widehat{\boldsymbol{\beta}}) + \|\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2, \qquad \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

Rearranging the terms of this inequality, we get the claim of the lemma. □

We will also repeatedly use the following result.

**Lemma 2.** *For any pair of vectors $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p$, for any pair of scalars $\mu > 0$ and $\gamma > 1$, for any $p \times p$ symmetric matrix $\mathbf{A}$ and for any set $J \subseteq [p]$, the following inequality is true*

$$\begin{aligned} &2\mu\gamma^{-1}\Big(\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\boldsymbol{\beta}'\|_1\Big) - \|\mathbf{A}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_2^2 \\ &\quad \leq 4\mu\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{(\gamma+1)^2\mu^2|J|}{\gamma^2\kappa_{\mathbf{A}^2}(J, c_\gamma)}, \end{aligned}$$

*where $c_\gamma = (\gamma+1)/(\gamma-1)$.*

*Proof.* To ease notation, we set $\boldsymbol{u} = \boldsymbol{\beta} - \boldsymbol{\beta}'$. Using that $\|\boldsymbol{\beta}_J\|_1 - \|\boldsymbol{\beta}'_J\|_1 \leq \|\boldsymbol{u}_J\|_1$ and $\|\boldsymbol{\beta}_{J^c}\|_1 + \|\boldsymbol{\beta}'_{J^c}\|_1 \geq \|\boldsymbol{u}_{J^c}\|_1$, we obtain

$$\begin{aligned} \|\boldsymbol{u}\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\boldsymbol{\beta}'\|_1 &= \|\boldsymbol{u}\|_1 + \gamma\big(\|\boldsymbol{\beta}_J\|_1 + \|\boldsymbol{\beta}_{J^c}\|_1 - \|\boldsymbol{\beta}'_J\|_1 - \|\boldsymbol{\beta}'_{J^c}\|_1\big) \\ &= \|\boldsymbol{u}\|_1 + 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + \gamma\big(\|\boldsymbol{\beta}_J\|_1 - \|\boldsymbol{\beta}'_J\|_1\big) \\ &\quad - \gamma\big(\|\boldsymbol{\beta}'_{J^c}\|_1 + \|\boldsymbol{\beta}_{J^c}\|_1\big) \\ &\leq \|\boldsymbol{u}\|_1 + 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + \gamma\|\boldsymbol{u}_J\|_1 - \gamma\|\boldsymbol{u}_{J^c}\|_1 \\ &= 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + (\gamma+1)\|\boldsymbol{u}_J\|_1 - (\gamma-1)\|\boldsymbol{u}_{J^c}\|_1 \\ &= 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + (\gamma+1)\big(\|\boldsymbol{u}_J\|_1 - c_\gamma^{-1}\|\boldsymbol{u}_{J^c}\|_1\big). \end{aligned}$$

If $c_\gamma\|\boldsymbol{u}_J\|_1 < \|\boldsymbol{u}_{J^c}\|_1$, the claim of the lemma is straightforward. Otherwise, $\|\boldsymbol{u}_{J^c}\|_1 \leq c_\gamma\|\boldsymbol{u}_J\|_1$ and using the definition of the compatibility constant we get

$$\begin{aligned} &\frac{2\lambda(\gamma+1)}{\gamma}\big(\|\boldsymbol{u}_J\|_1 - c_\gamma^{-1}\|\boldsymbol{u}_{J^c}\|_1\big) - \|\mathbf{A}\boldsymbol{u}\|_2^2 \\ &\quad \leq \frac{2\lambda(\gamma+1)}{\gamma}\left(\frac{|J|\cdot\|\mathbf{A}\boldsymbol{u}\|_2^2}{\kappa_{\mathbf{A}^2}(J, c_\gamma)}\right)^{1/2} - \|\mathbf{A}\boldsymbol{u}\|_2^2 \\ &\quad \leq \frac{(\gamma+1)^2\lambda^2|J|}{\gamma^2\kappa_{\mathbf{A}^2}(J, c_\gamma)}, \qquad \text{[by Cauchy-Schwarz]} \end{aligned}$$

which completes the proof. □

To close this subsection of auxiliary results, we provide simple upper bounds on the quantiles of some random noise variables.

**Proposition 1.** *Let $m = N - n$ and $n_\star = n \wedge m$. Introduce the random vectors $\boldsymbol{\zeta}^{(1)} = \frac{1}{n}\sum_{i=1}^n Y_i \boldsymbol{X}_i - \mathbb{E}[Y\boldsymbol{X}]$,*

$$\boldsymbol{\zeta} = \frac{1}{n}\sum_{i=1}^{n} Y_i \boldsymbol{X}_i - \frac{1}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i \quad \textit{and}$$
$$\bar{\boldsymbol{\zeta}} = \frac{1}{n}\sum_{i=1}^{n} Y_i \boldsymbol{X}_i - \frac{1}{N}\sum_{i=1}^{N} f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i.$$

*Under Assumptions (A1) and (A2), and for any $\delta \in (0,1)$, each of the following inequalities*

$$\|\boldsymbol{\zeta}^{(1)}\|_\infty \le 2B_Y\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big[1 + \frac{B_X}{3}\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big]$$
$$\|\boldsymbol{\zeta}\|_\infty \le 2B_Y\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big[1 + \frac{B_X}{3}\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big]$$
$$\|\bar{\boldsymbol{\zeta}}\|_\infty \le 2B_Y\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big[1 + \frac{B_X}{2}\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big]$$

*holds with probability at least $1 - \delta$.*

*Proof.* We will only prove the inequality corresponding to $\boldsymbol{\zeta}$. The others being very similar are left to the reader. Denote $\boldsymbol{\mu} = \mathbb{E}[Y\boldsymbol{X}] = \mathbb{E}[f^\star(\boldsymbol{X})\boldsymbol{X}] \in \mathbb{R}^p$, and introduce the random vectors

$$\boldsymbol{Z}_i = \begin{cases} N(Y_i\boldsymbol{X}_i - \boldsymbol{\mu})/n, & i \in [n], \\ N(\boldsymbol{\mu} - f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i)/m, & i \in [N]\setminus[n]. \end{cases}$$

The vectors $\boldsymbol{Z}_i$ are independent, centered, bounded and satisfy

$$\boldsymbol{\zeta} = \frac{\boldsymbol{Z}_1 + \cdots + \boldsymbol{Z}_N}{N}.$$

Furthermore, Assumption (A2) implies that $\|\boldsymbol{Z}_i\|_\infty \le 2NB_YB_X/n$ if $i \le n$ and that $\|\boldsymbol{Z}_i\|_\infty \le 2NB_YB_X/m$ if $i > n$. One can also bound from above the variance of the $j$-th component $Z_{ij}$ of $\boldsymbol{Z}_i$ as follows. If $i \le n$ then, in view of Assumptions (A1) and (A2), $\mathbb{E}[Z_{ij}^2] \le (N/n)^2\mathbb{E}[Y_i^2X_{ij}^2] \le (NB_Y/n)^2$. Similarly, if $i > n$ then $\mathbb{E}[Z_{ij}^2] \le (NB_Y/m)^2$. Hence, we may easily deduce that, for all $j \in [p]$,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Z_{ij}^2] \le \frac{2NB_Y^2}{n_\star}.$$

Therefore, using the Bernstein inequality recalled in Proposition 4 of Appendix A, for every $j \in [p]$ and every $\delta > 0$, we get that inequality

$$|\zeta_j| > 2B_Y\left(\frac{\log(2p/\delta)}{n_\star}\right)^{1/2} + \frac{2B_YB_X\log(2p/\delta)}{3n_\star}$$

holds with probability at most $\delta/p$. The claim of Proposition 1 follows from the union bound. □

**Remark 7.1.** *One can easily check that the inequality* $\mathbb{E}[Z_{ij}^2] \le (NB_Y/n)^2$, *for* $i = 1, \dots, n$, *used in the previous proof can be replaced by* $\mathbb{E}[Z_{ij}^2] \le (NL_Y B_X/n)^2$, *where* $L_Y = (\mathbb{E}[Y_i^2])^{1/2}$. *This may lead to a better risk bound in the cases where the random variable* $Y_i$ *is not well concentrated around its average value.*

We are now in a position to prove the main theorems of this paper.

### 7.1. *Proof of Theorem 5*

The proof of Theorem 5 follows directly from Proposition 1 and Proposition 2 below. For simplicity, the parameter $\gamma > 1$ introduced in Proposition 2 is fixed at the value $\gamma = 2$ in Theorem 5.

**Proposition 2.** *Let* $\boldsymbol{\zeta}$ *be as in Proposition 1. For any* $\gamma > 1$, *we set* $c_\gamma = (\gamma+1)/(\gamma-1)$. *On the event* $\mathscr{E} = \{\|\boldsymbol{\zeta}\|_\infty \le \lambda/\gamma\}$, *for every* $\boldsymbol{\beta} \in \mathbb{R}^p$ *and every* $J \subseteq [p]$, *we have*

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \le \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{(\gamma+1)^2\lambda^2|J|}{\gamma^2\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}}(J, c_\gamma)}.$$

*Proof.* Along the proof, we will use for convenience the shorthand notations $m = N - n$ and $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}^{1/2}$. First, notice that developing the square in the expression $\mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) = \frac{1}{m}\sum_{i=n+1}^{N}\left(\boldsymbol{X}_i^\top\boldsymbol{\beta} - f^\star(\boldsymbol{X}_i)\right)^2$, we get

$$\begin{aligned}\mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) &= \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \left(\frac{2}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i^\top\right)\boldsymbol{\beta} + \frac{1}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)^2 \\ &= \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + 2\boldsymbol{\zeta}^\top\boldsymbol{\beta} - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\mathrm{lab}}\boldsymbol{\beta} + \frac{1}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)^2.\end{aligned}$$

This implies that for every $\boldsymbol{\beta} \in \mathbb{R}^p$, we have

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) = \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + 2\boldsymbol{\zeta}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\mathrm{lab}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Using Lemma 1 with the convex penalty term $\mathrm{pen}(\boldsymbol{\beta}) = 2\lambda\|\boldsymbol{\beta}\|_1$, we deduce that, for every $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) \le 2\boldsymbol{\zeta}^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) - \|\mathbf{A}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\|_2^2. \quad (17)$$

On the event $\mathscr{E}$, note that $2\boldsymbol{\zeta}^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \le 2\|\boldsymbol{\zeta}\|_\infty\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 \le \frac{2\lambda}{\gamma}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1$, which leads to

$$2\boldsymbol{\zeta}^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \le \frac{2\lambda}{\gamma}\left(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1\right). \quad (18)$$

Combining equations (17) and (18), we get that on the event $\mathscr{E}$, for every $\boldsymbol{\beta} \in \mathbb{R}^p$ and every $J \subseteq [p]$,

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) \leq 2\lambda\gamma^{-1}\big(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1\big) - \|\mathbf{A}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\|_2^2. \tag{19}$$

The claim of the proposition follows from eq. (19) by applying Lemma 2 with $\mu = \lambda$. □

To conclude the proof of Theorem 5, it suffices to note that in view of Proposition 1, the probability of the event $\mathscr{E} = \{\|\boldsymbol{\zeta}\|_\infty \leq \lambda/\gamma\}$ is larger than $1 - \delta$ provided that

$$\lambda \geq 2\gamma B_Y \Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2} \Big[1 + \frac{B_X}{3}\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big].$$

### *7.2. Proofs for the semi-supervised version of the lasso*

We start this section by some arguments that are shared by the proofs of both theorems stated in Section 5. Let $J \subseteq [p]$ and let $\boldsymbol{\beta}$ be a minimizer of the right hand side of (15). Note in particular that $\boldsymbol{\beta}$ is a deterministic vector depending on the unknown distribution $P$ of the data. In addition, if the model is well-specified and $J = J^\star$ then $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$. We will also use the notation $\boldsymbol{u} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and

$$\boldsymbol{\zeta}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} Y_i \boldsymbol{X}_i - \mathbb{E}[Y\boldsymbol{X}] \quad \text{and} \quad \boldsymbol{\zeta}^{(2)} = \big(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}\big)\boldsymbol{\beta}. \tag{20}$$

Furthermore, to ease notation, we set $\widehat{\boldsymbol{\Sigma}}_N = \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}$, $\widehat{\boldsymbol{\Sigma}}_n = \widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}$, $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_N^{1/2}$. First, observe that the excess risk $\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \int_{\mathcal{X}} \big(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}} - f^\star(\boldsymbol{x})\big)^2 P_X(\mathrm{d}\boldsymbol{x})$ of the predictor $f_{\widehat{\boldsymbol{\beta}}}$ satisfies

$$\begin{aligned}\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) &= \int_{\mathcal{X}} \big\{(\boldsymbol{x}^\top\boldsymbol{u})^2 + 2\boldsymbol{u}^\top\boldsymbol{x}\big(\boldsymbol{x}^\top\boldsymbol{\beta} - f^\star(\boldsymbol{x})\big) + \big(\boldsymbol{x}^\top\boldsymbol{\beta} - f^\star(\boldsymbol{x})\big)^2\big\} P_X(\mathrm{d}\boldsymbol{x}) \\ &= \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 + 2\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\boldsymbol{u}^\top\mathbb{E}\left[\boldsymbol{X} f^\star(\boldsymbol{X})\right] + \mathcal{E}(f_{\boldsymbol{\beta}}).\end{aligned} \tag{21}$$

Next, notice that

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 = \boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} + \|\mathbf{A}\boldsymbol{u}\|_{\mathbf{2}}^{\mathbf{2}}, \tag{22}$$

and that

$$\begin{aligned}2\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} &= 2\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{\beta} + 2\boldsymbol{u}^\top\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\beta} \\ &= 2\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{\beta} + \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_{\mathbf{2}}^{\mathbf{2}} - \|\mathbf{A}\boldsymbol{u}\|_{\mathbf{2}}^{\mathbf{2}} - \|\mathbf{A}\boldsymbol{\beta}\|_{\mathbf{2}}^{\mathbf{2}},\end{aligned} \tag{23}$$

where in the last line we have used the identity $2a^\top b = \|a + b\|_2^2 - \|a\|_2^2 - \|b\|_2^2$ with $a = \mathbf{A}\boldsymbol{u}$ and $b = \mathbf{A}\boldsymbol{\beta}$. Transforming eq. (21) thanks to (22) and (23)

we obtain

$$\begin{aligned}\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) &= \boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} + 2\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{\beta} + \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - 2\boldsymbol{u}^\top\mathbb{E}[Y\boldsymbol{X}] \\ &= \boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} + 2\boldsymbol{u}^\top\boldsymbol{\zeta}^{(2)} + \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + 2\boldsymbol{u}^\top\boldsymbol{\zeta}^{(1)} - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_n\boldsymbol{u},\end{aligned} \tag{24}$$

where we have used the identity $\mathbb{E}[Y\boldsymbol{X}] = \mathbb{E}[\boldsymbol{X}f^\star(\boldsymbol{X})]$ and the definitions of $\boldsymbol{\zeta}^{(1)}$ and $\boldsymbol{\zeta}^{(2)}$. Applying Lemma 1 with $\mathrm{pen}(\boldsymbol{\beta}) = 2\lambda\|\boldsymbol{\beta}\|_1$ and combining its result with (24), we arrive at

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) \le \underbrace{2\boldsymbol{u}^\top(\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1)}_{\mathbf{T}_1} + \underbrace{\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} - \|\mathbf{A}\boldsymbol{u}\|_2^2}_{\mathbf{T}_2}. \tag{25}$$

*7.2.1. Proof of Theorem 6*

As mentioned earlier, in the well-specified setting we have $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ and, therefore, $\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2$ and $\mathcal{E}(f_{\boldsymbol{\beta}^\star}) = 0$. Hence, (25) yields

$$2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \le 2\boldsymbol{u}^\top\big(\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\big) + 2\lambda(\|\boldsymbol{\beta}^\star\|_1 - \|\boldsymbol{\beta}^\star + \boldsymbol{u}\|_1). \tag{26}$$

Combining the duality inequality $|\boldsymbol{u}^\top\big(\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\big)| \le \|\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\|_\infty\|\boldsymbol{u}\|_1$ with the following one $\|\boldsymbol{\beta}^\star\|_1 - \|\boldsymbol{\beta}^\star + \boldsymbol{u}\|_1 = \|\boldsymbol{\beta}^\star_{J^\star}\|_1 - \|\boldsymbol{\beta}^\star_{J^\star} + \boldsymbol{u}_{J^\star}\|_1 - \|\boldsymbol{u}_{(J^\star)^c}\|_1 \le \|\boldsymbol{u}_{J^\star}\|_1 - \|\boldsymbol{u}_{(J^\star)^c}\|_1$, we infer from inequality (26) that on the event $\mathscr{E} = \big\{2\|\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\|_\infty \le \lambda\big\}$, we have

$$2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \le \lambda(3\|\boldsymbol{u}_{J^\star}\|_1 - \|\boldsymbol{u}_{(J^\star)^c}\|_1). \tag{27}$$

This implies that $\|\boldsymbol{u}_{(J^\star)^c}\|_1 \le 3\|\boldsymbol{u}_{J^\star}\|_1$ and, therefore,

$$2\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_N}(J^\star, 3)\|\boldsymbol{u}_{J^\star}\|_1^2 \le 2s^\star\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \le 3\lambda s^\star\|\boldsymbol{u}_{J^\star}\|_1.$$

This yields $\|\boldsymbol{u}_{J^\star}\|_1 \le 3\lambda s^\star/(2\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_N}(J^\star, 3))$ and, since $\max_{j,j'}|\boldsymbol{\Sigma}_{j,j'}| \le 1$, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2 \le \|\boldsymbol{u}\|_1 \le 4\|\boldsymbol{u}_{J^\star}\|_1$, which implies that

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \le \left(\frac{6\lambda s^\star}{\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_N}(J^\star, 3)}\right)^2. \tag{28}$$

On the other hand, if we denote by $I$ the set of the $s^\star$ largest entries of the vector $|\boldsymbol{u}|$, inequality (27) implies that $2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \le \lambda(3\|\boldsymbol{u}_I\|_1 - \|\boldsymbol{u}_{I^c}\|_1)$.

Therefore, using the definition of the restricted eigenvalue and similar arguments as above, we deduce that $\|\boldsymbol{u}_I\|_2 \le 3\lambda\sqrt{s^\star}/(2\kappa^{\mathrm{RE}}_{\widehat{\boldsymbol{\Sigma}}_N}(I, 3))$. Furthermore,

$\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{u}_I\|_2^2 + \|\boldsymbol{u}_{I^c}\|_2^2 \le \|\boldsymbol{u}_I\|_2^2 + \|\boldsymbol{u}_{I^c}\|_\infty \|\boldsymbol{u}_{I^c}\|_1 \le \|\boldsymbol{u}_I\|_2^2 + (s^\star)^{-1}\|\boldsymbol{u}_I\|_1 \|\boldsymbol{u}_{I^c}\|_1 \le \|\boldsymbol{u}_I\|_2^2 + 3(s^\star)^{-1}\|\boldsymbol{u}_I\|_1^2 \le 4\|\boldsymbol{u}_I\|_2^2$. This yields

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \le \|\boldsymbol{\Sigma}\| \cdot \|\boldsymbol{u}\|_2^2 \le 4\|\boldsymbol{\Sigma}\| \cdot \|\boldsymbol{u}_I\|_2^2 \le \frac{9\|\boldsymbol{\Sigma}\|\lambda^2 s^\star}{\kappa^{\mathrm{RE}}_{\widehat{\boldsymbol{\Sigma}}_N}(I,3)^2}\,. \tag{29}$$

Combining (28) and (29), we get the first claim of the theorem.

To get the second claim of the theorem, we go back to (27) and use the following inequalities:

$$\begin{aligned} 2\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 &= 2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 + 2\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} \\ &\le 3\lambda\|\boldsymbol{u}_{J^\star}\|_1 + 2\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N\|_\infty \|\boldsymbol{u}\|_1^2 \\ &\le 3\lambda\|\boldsymbol{u}_{J^\star}\|_1 + 32\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N\|_\infty \|\boldsymbol{u}_{J^\star}\|_1^2. \end{aligned} \tag{30}$$

In the sequel, let us denote $\kappa = \bar{\kappa}_{\boldsymbol{\Sigma}}(J^\star, 3)$ for brevity. Then, upper bounding the two instances of $\|\boldsymbol{u}_{J^\star}\|_1$ in (30) by $(s^\star\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2/\kappa)^{1/2}$, we infer that on $\mathscr{E}$,

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \le \frac{3\lambda\sqrt{s^\star}}{2\sqrt{\kappa}}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2 + \frac{16 s^\star}{\kappa}\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N\|_\infty \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2.$$

Dividing both sides by $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2$ (if this quantity vanishes then the claim of the theorem is obviously true) and after some algebra, we get the inequality

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \le \frac{9\lambda^2 s^\star \kappa}{4(\kappa - 16 s^\star \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N\|_\infty)^2} \le \frac{9\lambda^2 s^\star}{\kappa},$$

where the last inequality holds on the event $\mathscr{E} \cap \{32 s^\star \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N\|_\infty \le \kappa\}$. In view of the union bound, Hoeffding's inequality and Assumption (A2), we get for any $t > 0$,

$$\mathbf{P}\left(\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N\|_\infty \ge t\right) \le p^2 \max_{j,j' \in [p]} \mathbf{P}\left(|\sigma_{jj'} - \widehat{\sigma}_{jj'}| \ge t\right) \le 2p^2 \exp\left(-2Nt^2/B_X^4\right),$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})$ and $\widehat{\boldsymbol{\Sigma}}_N = (\widehat{\sigma}_{ij})$. Therefore, if

$$16 s^\star B_X^2 \Big(\frac{2\log(4p^2/\delta)}{N}\Big)^{1/2} \le \kappa,$$

then the event $\{32 s^\star \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N\|_\infty \le \kappa\}$ has a probability larger than $1 - (\delta/2)$. To bound the probability of $\mathscr{E}$, we use the fact that $\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)} = \bar{\boldsymbol{\zeta}}$ and the quantiles of the supremum norm of the random vector $\bar{\boldsymbol{\zeta}}$ have been assessed in Proposition 1. This implies that the choice

$$\lambda \ge 4B_Y\Big(\frac{\log(4p/\delta)}{n}\Big)^{1/2} + \frac{B_X B_Y \log(4p/\delta)}{n}$$

guarantees that $P(\mathscr{E}) = P(\|\boldsymbol{\zeta}\|_\infty \le \lambda/2) \ge 1 - (\delta/2)$. This completes the proof.

*7.2.2. Proof of Theorem 7*

We start by some auxiliary results before providing the proof of the theorem.

**Proposition 3.** *Let $J \subseteq [p]$ and let $\boldsymbol{\beta}$ be a minimizer of the right hand side of (15). On the event $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2 \cap \mathscr{E}_3$, where*

$$\mathscr{E}_1 = \big\{\|\boldsymbol{\zeta}^{(1)}\|_\infty \le \tfrac{\lambda}{4}\big\}, \quad \mathscr{E}_2 = \big\{\|\boldsymbol{\zeta}^{(2)}\|_\infty \le \tfrac{\lambda}{4}\big\}, \quad \textit{and}$$
$$\mathscr{E}_3 = \big\{\lambda_{\min}(\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}) \ge \tfrac{2}{3}\big\},$$

*we have*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) \le 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2|J|}{2\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}}(J,3)}.$$

*Proof.* Our starting point in this proof is (24). We first focus on bounding $\mathbf{T}_1$. On the event $\mathscr{E}_1 \cap \mathscr{E}_2$, we have

$$\mathbf{T}_1 \le 2\|\boldsymbol{\zeta}^{(1)}+\boldsymbol{\zeta}^{(2)}\|_\infty\|\boldsymbol{u}\|_1 + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \le \lambda\big(\|\boldsymbol{u}\|_1 + 2\|\boldsymbol{\beta}\|_1 - 2\|\widehat{\boldsymbol{\beta}}\|_1\big). \quad (31)$$

We now look for an upper bound of the term $\mathbf{T}_2$. On the event $\mathscr{E}_3$, for any $\boldsymbol{v} \in \mathbb{R}^p$,

$$\boldsymbol{v}^\top\big(2\mathbf{I}_p - 3\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{v} \le 0,$$

which leads to

$$\boldsymbol{v}^\top\big(\mathbf{I}_p - 2\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{v} \le -\tfrac{1}{2}(\boldsymbol{v}^\top\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2})\boldsymbol{v}. \quad (32)$$

Therefore, applying (32) to $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}$, it follows that on the event $\mathscr{E}_3$

$$\mathbf{T}_2 = \boldsymbol{v}^\top\big(\mathbf{I}_p - 2\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{v} \le -\tfrac{1}{2}\boldsymbol{v}^\top(\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2})\boldsymbol{v} = -\tfrac{1}{2}\|\mathbf{A}\boldsymbol{u}\|_2^2. \quad (33)$$

To sum up, equations (31) and (33) together imply that on the event $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2 \cap \mathscr{E}_3$,

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) \le \lambda\big(\|\boldsymbol{u}\|_1 + 2\|\boldsymbol{\beta}\|_1 - 2\|\widehat{\boldsymbol{\beta}}\|_1\big) - \tfrac{1}{2}\|\mathbf{A}\boldsymbol{u}\|_2^2.$$

The desired result follows from this inequality and Lemma 2 with $\mu = \lambda$ and $\gamma = 2$. □

Note that according to Proposition 1,

$$\mathbf{P}\left(\|\boldsymbol{\zeta}^{(1)}\|_\infty \le 2B_Y\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\left[1 + \frac{B_X}{3}\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\right]\right) \ge 1 - \frac{\delta}{3}. \quad (34)$$

The next two lemmas provide bounds for the probabilities of the events $\mathscr{E}_2$ and $\mathscr{E}_3$ introduced in Proposition 3.

**Lemma 3.** *Let assumption (A2) be fulfilled. Let $J \subseteq [p]$ and let $\boldsymbol{\beta}$ be a minimizer of the right hand side of (15). Then, for all $\delta \in (0,1)$, the inequality*

$$\|\boldsymbol{\zeta}^{(2)}\|_\infty \geq B_X B_Y \Big(\frac{2\log(6p/\delta)}{N}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{2p\|\boldsymbol{\Sigma}^{-1}\|\log(6p/\delta)}{N}\Big)^{1/2}\Big]$$

*holds with probability at most $\delta/3$, where the random vector $\boldsymbol{\zeta}^{(2)}$ is defined in eq. (20).*

*Proof.* Note that $\boldsymbol{\zeta}^{(2)} = (1/N)\sum_{i=1}^N \boldsymbol{U}_i$, where $\boldsymbol{U}_i = \boldsymbol{X}_i(\boldsymbol{X}_i^\top\boldsymbol{\beta}) - \mathbb{E}[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\beta})]$. The random vectors $\boldsymbol{U}_i$ are independent and, for all $i \in [N]$ and all $j \in [p]$, the $j$-th component $U_{ij} = X_{ij}(\boldsymbol{X}_i^\top\boldsymbol{\beta}) - \mathbb{E}[X_j(\boldsymbol{X}^\top\boldsymbol{\beta})]$ of $\boldsymbol{U}_i$ satisfies, almost surely,

$$|U_{ij}| \leq 2B_X^2\|\boldsymbol{\beta}\|_1 \leq 2B_X^2\sqrt{p}\|\boldsymbol{\beta}\|_2,$$

where we have used that $|\boldsymbol{X}^\top\boldsymbol{\beta}| \leq \|\boldsymbol{X}\|_\infty\|\boldsymbol{\beta}\|_1 \leq B_X\|\boldsymbol{\beta}\|_1$ with probability 1. Then, noticing that $\|\boldsymbol{\beta}\|_2 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 \leq \|\boldsymbol{\Sigma}^{-1/2}\|\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 = \|\boldsymbol{\Sigma}^{-1}\|^{1/2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2$, we deduce that

$$|U_{ij}| \leq 2B_X^2(p\|\boldsymbol{\Sigma}^{-1}\|)^{1/2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2,$$

almost surely. Since $\boldsymbol{\beta}$ minimizes the term on the right hand side of (15), by Lemma 5 below, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 \leq B_Y$. Thus for all $i \in [N]$ and all $j \in [p]$, $|U_{ij}| \leq 2B_X^2B_Y(p\|\boldsymbol{\Sigma}^{-1}\|)^{1/2}$. Furthermore, according to the previous lines, it holds $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[X_{ij}^2(\boldsymbol{X}_i^\top\boldsymbol{\beta})^2] \leq B_X^2B_Y^2$. Proposition 4 and the union bound complete the proof. □

**Lemma 4.** *Under assumption (A2), the smallest eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2})$ of the matrix $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}$ satisfies*

$$\mathbf{P}\Big\{\lambda_{\min}(\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}) \geq 1-\Big(\frac{2B_X^2p\|\boldsymbol{\Sigma}^{-1}\|\log(p/\delta)}{N}\Big)^{1/2}\Big\} \geq 1-\delta, \quad (35)$$

*for all $\delta \in (0,1)$ such that $2B_X^2p\|\boldsymbol{\Sigma}^{-1}\|\log(p/\delta) \leq N$.*

*Proof.* For all $i \in [N]$, $\lambda_{\max}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Sigma}^{-1/2}) = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|^2 \leq pB_X^2\|\boldsymbol{\Sigma}^{-1}\|$ and the matrix $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Sigma}^{-1/2}$ is positive semi-definite. Applying the first Chernoff matrix inequality given in Remark 5.3 of Tropp (2012) to the sequence of matrices $\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Sigma}^{-1/2} : i \in [N]\}$ with

$$t = 1-\Big(\frac{2B_X^2p\|\boldsymbol{\Sigma}^{-1}\|\log(p/\delta)}{N}\Big)^{1/2}, \qquad R = pB_X^2, \qquad \delta = p\exp\Big\{-\frac{(1-t)^2N}{2R\|\boldsymbol{\Sigma}^{-1}\|}\Big\}$$

yields (35). □

**Lemma 5.** *Let $\mathrm{pen} : \mathbb{R}^p \to [0,+\infty)$ be a convex function such that $\mathrm{pen}(\mathbf{0}_p) = 0$. Let $\bar{\boldsymbol{\beta}}$ be a minimizer of the function*

$$\Phi(\boldsymbol{\beta}) = \mathbb{E}[(\boldsymbol{\beta}^\top\boldsymbol{X} - Y)^2] + \mathrm{pen}(\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbb{R}^p.$$

*Then $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \leq \mathbb{E}[Y^2]$ and, if Assumption (A2) is fulfilled, $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \leq B_Y^2$.*

*Proof.* We apply Lemma 1 with $\mathbf{A} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]^{1/2}$, $n = 1$, $\boldsymbol{Y} = 1$ and $\mathbf{X}_{\rm lab} = \mathbb{E}[Y\boldsymbol{X}]$ so that $\frac{1}{n}\boldsymbol{Y}^\top\mathbf{X}_{\rm lab} = \mathbb{E}[Y\boldsymbol{X}]$. Inequality (16) with $\boldsymbol{\beta} = \mathbf{0}_p$ yields

$$\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \le 2\mathbb{E}[Y(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})] - \mathrm{pen}(\bar{\boldsymbol{\beta}}) - \mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2].$$

Rearranging the terms and using that $\mathrm{pen}(\bar{\boldsymbol{\beta}}) \ge 0$, we get $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \le \mathbb{E}[Y(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})]$. In view of the Cauchy-Schwarz inequality, $(\mathbb{E}[Y(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})])^2 \le \mathbb{E}[Y^2]\,\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2]$, which implies that $(\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2])^2 \le \mathbb{E}[Y^2]\,\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2]$. It now suffices to divide both sides of the last inequality by $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2]$ to obtain the claim of the lemma. □

*Proof of theorem 7.* Under the conditions of the theorem, we have

$$\Big(\frac{2B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|\log(3p/\delta)}{N}\Big)^{1/2} \le \frac{1}{3}.$$

Therefore, Lemma 4 implies that $\mathbf{P}(\mathscr{E}_3) \ge 1 - \delta/3$. On the other hand, in view of eq. (34) and Lemma 3, the conditions

$$\lambda \ge 8B_Y\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\Big[1 + \frac{B_X}{3}\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\Big],$$
$$\lambda \ge 4B_XB_Y\Big(\frac{2\log(6p/\delta)}{N}\Big)^{1/2}\Big[1 + \frac{B_X}{3}\Big(\frac{2p\|\boldsymbol{\Sigma}^{-1}\|\log(6p/\delta)}{N}\Big)^{1/2}\Big]$$

imply that $\mathbf{P}(\mathscr{E}_1) \ge 1 - \delta/3$ and $\mathbf{P}(\mathscr{E}_2) \ge 1 - \delta/3$. One can easily check that under the conditions of the theorem, the two inequalities of the last display are satisfied. Therefore, we have $\mathbf{P}(\mathscr{E}_1 \cap \mathscr{E}_2 \cap \mathscr{E}_3) \ge 1 - \delta$. Finally, applying Proposition 3 we get the claim of the theorem. □

### *7.2.3. Proof of the oracle inequality in expectation*

Let $\delta$ be a positive number smaller than 1 to be chosen later. We have already seen in Corollary 1 that on an event $\mathscr{E}$ of probability $1 - \delta$, we have

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \le \inf_{J\subseteq[p]}\inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \tfrac{27\|\boldsymbol{\Sigma}^{-1}\|}{4}\lambda^2|J|\Big\}.$$

On the other hand, using the fact that $\widehat{\boldsymbol{\beta}}$ minimises the function $\psi(\boldsymbol{\beta}) = \|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{\beta}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_n\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1$, we have $\psi(\widehat{\boldsymbol{\beta}}) \le \psi(\mathbf{0}_p)$, which yields

$$\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\widehat{\boldsymbol{\beta}}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_n\widehat{\boldsymbol{\beta}} + 2\lambda\|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\widehat{\boldsymbol{\beta}} - \tfrac{1}{n}\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top\boldsymbol{Y}\|_2^2$$
$$- \frac{1}{n^2}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top\boldsymbol{Y}\|_2^2 + 2\lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le 0.$$

Note that $\widehat{\boldsymbol{\Sigma}}_N^{-1/2}$ is understood as the Moore-Penrose pseudo-inverse and all the expressions involving this quantity are well defined since $N\widehat{\boldsymbol{\Sigma}}_N \succeq n\widehat{\boldsymbol{\Sigma}}_n =$

$\mathbf{X}_n^\top \mathbf{X}_n$. This implies that $2\lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{1}{n^2}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top \boldsymbol{Y}\|_2^2 \le \frac{1}{n^2}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top\|^2\|\boldsymbol{Y}\|_2^2 = \frac{1}{n}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\|\,\|\boldsymbol{Y}\|_2^2$, which entails

$$\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{B_Y^2}{2\lambda}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\| \le \frac{B_Y^2 N}{2n\lambda}. \tag{36}$$

It is also true that for every $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\mathcal{E}(f_{\boldsymbol{\beta}}) = \mathbb{E}[(f^\star(\boldsymbol{X}) - \boldsymbol{X}^\top\boldsymbol{\beta})^2] \le 2\mathbb{E}[f^\star(\boldsymbol{X})^2] + 2\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} \le 2B_Y^2 + 2\|\boldsymbol{\beta}\|_1^2.$$

Therefore, we have $\mathbb{E}[\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}})\mathbb{1}_{\mathscr{E}^c}] \le 2B_Y^2\mathbf{P}(\mathscr{E}^c) + 2\mathbb{E}[\|\widehat{\boldsymbol{\beta}}\|_1^2\mathbb{1}_{\mathscr{E}^c}] = 2\delta B_Y^2 + 2\mathbb{E}[\|\widehat{\boldsymbol{\beta}}\|_1^2\mathbb{1}_{\mathscr{E}^c}]$. Combining this inequality with (36), we get

$$\mathbb{E}[\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}})\mathbb{1}_{\mathscr{E}^c}] \le 2\delta B_Y^2 + \frac{\delta B_Y^4 N^2}{2n^2\lambda^2}.$$

Setting $\delta = N^{-2}$, we get the claim of the theorem.

## Appendix A: Bernstein inequality

The next result follows from (Massart, 2007, Proposition 2.9).

**Proposition 4.** *Let $Z_1, \dots, Z_N$ be independent real-valued random variables satisfying, for all $i \in [N]$ and for some constant $b$, $\mathbb{E}[Z_i^2] < +\infty$ and $|Z_i - \mathbb{E}Z_i| \le b$ almost surely. Denote $\bar{Z}_N = \frac{1}{N}\sum_{i=1}^N Z_i$ and $\sigma_N^2 = (1/N)\sum_{i=1}^N \mathbb{E}[Z_i^2 - (\mathbb{E}Z_i)^2]$. Then, for all $\delta \in (0,1)$, inequality*

$$|\bar{Z}_N - \mathbb{E}[\bar{Z}_N]| \le \sigma_N\Big(\frac{2\log(2/\delta)}{N}\Big)^{1/2}\Big[1 + \frac{b}{6N\sigma_N}\Big(\frac{2\log(2/\delta)}{N}\Big)^{1/2}\Big],$$

*holds with probability at least $1-\delta$.*

*Proof.* Define, for all $i \in [N]$, the random variable $X_i = (Z_i - \mathbb{E}[Z_i])/N$. Denote as well

$$v = \sum_{i=1}^N \mathbb{E}[X_i^2] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[Z_i^2 - (\mathbb{E}Z_i)^2\right] = \frac{u}{N}.$$

For all $k \ge 3$, the assumptions imply that

$$\sum_{i=1}^N \mathbb{E}[(X_i)_+^k] \le v\left(\frac{b}{N}\right)^{k-2} \le \frac{k!}{2}v\left(\frac{b}{3N}\right)^{k-2},$$

where we have used the fact that $k!/3^{k-2} \ge 2$, for all $k \ge 3$. As a result, applying (Massart, 2007, Prop. 2.9), with $v = \sigma_N^2/N$ and $c = b/3N$, we get that for all $\delta \in (0,1)$, the inequality

$$\sum_{i=1}^N X_i > \sigma_N\sqrt{\frac{2\log(2/\delta)}{N}} + \frac{b\log(2/\delta)}{3N}$$

holds with probability less than $\delta/2$. Applying the same argument to the variables $-X_i$, we infer that for all $\delta \in (0,1)$, the inequality

$$\sum_{i=1}^{N} X_i < -\sigma_N \sqrt{\frac{2\log(2/\delta)}{N}} - \frac{b\log(2/\delta)}{3N},$$

holds with probability less than $\delta/2$, which completes the proof. □

## Acknowledgments

The work of Q. Paris was supported by the Russian Academic Excellence Project 5-100. The work of A. Dalalyan and P. Bellec was partially supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and the chair "LCL/GENES/Fondation du risque, Nouveaux enjeux pour nouvelles données".

## References

Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication. MR1641250

Maria-Florina Balcan, Avrim Blum, Patrick Pakyan Choi, John Lafferty, Brian Pantano, Mugizi R. Rwebangira, and Xiaojin Zhu. Person identification in webcam images: An application of semi-supervised learning. *ICML2005 Workshop on Learning with Partially Classified Training Data*, 2005.

Matthieu Guillaumin, Jakob J. Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010*, pages 902–909, 2010. URL http://dx.doi.org/10.1109/CVPR.2010.5540120.

Céline Brouard, Florence d'Alché-Buc, and Marie Szafranski. Semi-supervised penalized output kernel regression for link prediction. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28–July 2, 2011*, pages 593–600. Omnipress, 2011.

O. Chapelle, B. Shölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin – Madison, 2008.

Philippe Rigollet. Generalized error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392, 2007. MR2332435

Junhui Wang and Xiaotong Shen. Large margin semi-supervised learning. *J. Mach. Learn. Res.*, 8:1867–1891, 2007. MR2353822

John D. Lafferty and Larry A. Wasserman. Statistical analysis of semi-supervised regression. In *NIPS*, pages 801–808. Curran Associates, Inc., 2007.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006. MR2274444

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems 22*, pages 1330–1338. Curran Associates, Inc., 2009.

Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14:1229–1250, 2013. URL http://jmlr.org/papers/v14/niyogi13a.html. MR3081923

Shiliang Sun and John Shawe-Taylor. Sparse semi-supervised learning using conjugate functions. *J. Mach. Learn. Res.*, 11:2423–2455, 2010. MR2727770

David Azriel, Lawrence D Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *arXiv preprint arXiv:1612.02391*, 2016.

A. Chakrabortty and T. Cai. Efficient and Adaptive Linear Regression in Semi-Supervised Settings. *ArXiv e-prints*, January 2017. MR3819109

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. MR1379242

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications. MR2807761

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *Annals of Statistics*, 46(6B):3603–3642, 2018. URL https://projecteuclid.org/euclid.aos/1536631285. MR3852663

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011. MR2906869

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. MR2999166

Arnak S. Dalalyan, Mohamed Heibiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552-581, 2017. MR3556784

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the Lasso and Dantzig selector for the $\ell_q$ loss in $\ell_r$ balls. *J. Mach. Learn. Res.*, 11:3519–3540, 2010. MR2756192

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011. MR2882274

Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011. MR2816337

Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012. MR3025134

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. MR2533469

Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009. MR2576316

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 38. Springer, 2011. MR2829871

Pierre Alquier and Mohamed Hebiri. Transductive versions of the LASSO and the dantzig selector. *Journal of Statistical Planning and Inference*, 142(9):2485–2500, 2012. MR2922000

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. Technical Report 1601.05584, arXiv, January 2016. MR3782379

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010. MR2719855

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv preprint arXiv:1312.2903*, 2013. MR3568047

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013. MR3061256

Anatoli Juditsky and Arkadi Nemirovski. Accuracy guarantees for-recovery. *Information Theory, IEEE Transactions on*, 57(12):7818–7839, 2011. MR2895363

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 04 2014. URL http://dx.doi.org/10.1214/14-AOS1204. MR3210986

M. Pensky. Solution of linear ill-posed problems using overcomplete dictionaries. Technical Report 1408.3386, Ann. Statist., to appear, arXiv, August 2014. MR3519939

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *ArXiv e-prints*, November 2010. MR2963170

Bubacarr Bah and Jared Tanner. Bounds of restricted isometry constants in extreme asymptotics: formulae for Gaussian matrices. *Linear Algebra Appl.*, 441:88–109, 2014. MR3134338

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. MR2946459

Pascal Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*, volume 1896. Springer, 2007. MR2319879