Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2009, Vol. 45, No. 1, 7–57
DOI: 10.1214/07-AIHP146

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

Sparsity in penalized empirical risk minimization

Vladimir Koltchinskii[1]

School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA. E-mail: vlad@math.gatech.edu

Received 11 October 2006; revised 6 June 2007; accepted 10 September 2007

Abstract. Let (X, Y) be a random couple in $S \times T$ with unknown distribution P. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. copies of (X, Y), P_n being their empirical distribution. Let $h_1, \dots, h_N : S \mapsto [-1, 1]$ be a dictionary consisting of N functions. For $\lambda \in \mathbb{R}^N$, denote $f_\lambda := \sum_{j=1}^N \lambda_j h_j$. Let $\ell : T \times \mathbb{R} \mapsto \mathbb{R}$ be a given loss function, which is convex with respect to the second variable. Denote $(\ell \bullet f)(x, y) := \ell(y; f(x))$. We study the following penalized empirical risk minimization problem

$$\hat{\lambda}^\varepsilon := \operatorname*{argmin}_{\lambda \in \mathbb{R}^N} \bigl[P_n(\ell \bullet f_\lambda) + \varepsilon \|\lambda\|_{\ell_p}^p \bigr],$$

which is an empirical version of the problem

$$\lambda^\varepsilon := \operatorname*{argmin}_{\lambda \in \mathbb{R}^N} \bigl[P(\ell \bullet f_\lambda) + \varepsilon \|\lambda\|_{\ell_p}^p \bigr]$$

(here $\varepsilon \geq 0$ is a regularization parameter; λ^0 corresponds to $\varepsilon = 0$). A number of regression and classification problems fit this general framework. We are interested in the case when $p \geq 1$, but it is close enough to 1 (so that $p - 1$ is of the order $\frac{1}{\log N}$, or smaller). We show that the "sparsity" of λ^ε implies the "sparsity" of $\hat{\lambda}^\varepsilon$ and study the impact of "sparsity" on bounding the excess risk $P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^0})$ of solutions of empirical risk minimization problems.

Résumé. Soit (X, Y) un couple aléatoire à valeurs dans $S \times T$ et de loi P inconnue. Soient $(X_1, Y_1), \dots, (X_n, Y_n)$ des répliques i.i.d. de (X, Y), de loi empirique associée P_n. Soit $h_1, \dots, h_N : S \mapsto [-1, 1]$ un dictionnaire composé de N fonctions. Pour tout $\lambda \in \mathbb{R}^N$, on note $f_\lambda := \sum_{j=1}^N \lambda_j h_j$. Soit $\ell : T \times \mathbb{R} \mapsto \mathbb{R}$ fonction de perte donnée que l'on suppose convexe en la seconde variable. On note $(\ell \bullet f)(x, y) := \ell(y; f(x))$. On étudie le problème de minimisation du risque empirique pénalisé suivant

$$\hat{\lambda}^\varepsilon := \operatorname*{argmin}_{\lambda \in \mathbb{R}^N} \bigl[P_n(\ell \bullet f_\lambda) + \varepsilon \|\lambda\|_{\ell_p}^p \bigr],$$

qui correspond à la version empirique du problème

$$\lambda^\varepsilon := \operatorname*{argmin}_{\lambda \in \mathbb{R}^N} \bigl[P(\ell \bullet f_\lambda) + \varepsilon \|\lambda\|_{\ell_p}^p \bigr]$$

(ici $\varepsilon \geq 0$ est un paramètre de régularisation; λ^0 correspond au cas $\varepsilon = 0$). Ce cadre général englobe un certain nombre de problèmes de régression et de classification. On s'intéresse au cas où $p \geq 1$, mais reste proche de 1 (de sorte que $p - 1$ soit de l'ordre $\frac{1}{\log N}$, ou inférieur). On montre que la "sparsité" de λ^ε implique la "sparsité" de $\hat{\lambda}^\varepsilon$. En outre, on étudie les conséquences de la "sparsité" en termes de bornes supérieures sur l'excès de risque $P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^0})$ des solutions obtenues pour les différents problèmes de minimisation du risque empirique.

[1]Partially supported by NSF Grants DMS-0304861 and DMS-MSPA-0624841.

MSC: 62G99; 62J99; 62H30

Keywords: Empirical risk; Penalized empirical risk; ℓ_p-penalty; Sparsity; Oracle inequalities

1. Introduction

Let (X, Y) be a random couple in $S \times T$, where S and T are measurable spaces with σ-algebras $\mathcal{S}$ and $\mathcal{T}$, respectively. The distribution P of (X, Y) is unknown, but the training data $(X_1, Y_1), \dots, (X_n, Y_n)$ consisting of n i.i.d. copies of (X, Y) is available. The distribution of X will be denoted by Π.

Let $\ell : T \times \mathbb{R} \mapsto \mathbb{R}_+$ be a loss function. We will assume in what follows that, for all $y \in T$, $\ell(y, \cdot)$ is convex. For a function $f : S \mapsto \mathbb{R}$, denote $(\ell \bullet f)(x, y) := \ell(y, f(x))$. Let $\mathcal{F}$ be a convex class of measurable functions on S. Consider the following *convex risk minimization* problem:

$$\mathbb{E}\ell\big(Y, f(X)\big) = P(\ell \bullet f) \longrightarrow \min,\ f \in \mathcal{F}. \tag{1.1}$$

Since the distribution P is not known, the ℓ-risk $P(\ell \bullet f)$ is estimated by its empirical version (the empirical ℓ-risk) defined as

$$n^{-1} \sum_{j=1}^{n} \ell\big(Y_j, f(X_j)\big) = P_n(\ell \bullet f),$$

where P_n is the empirical distribution based on the sample $(X_1, Y_1), \dots, (X_n, Y_n)$, which leads to the following empirical version of problem (1.1):

$$P_n(\ell \bullet f) \longrightarrow \min,\ f \in \mathcal{F}. \tag{1.2}$$

In many applications, the random variable X represents an observable instance whereas Y is not observable and is to be predicted based on the observation of X. Binary classification, in which $T = \{-1, 1\}$, is a particular example of this problem. In this case, the loss function is often defined as $\ell(y, u) = \phi(yu)$ with a decreasing convex function ϕ satisfying the condition $\phi(u) \geq I_{(-\infty,0]}(u)$. This choice is typical in so called large margin classification methods such as boosting and kernel machines. If f_* denotes the function that minimizes the ℓ-risk over the set of all measurable functions f, then under very mild assumptions $\operatorname{sign}(f_*)$ is a classifier that minimizes the generalization error over the set of all binary classifiers (called the Bayes classifier). Another example is regression with random design. In this case, $T := \mathbb{R}$ and one can use, for instance, $\ell(y, u) := (y - u)^2$ (L_2-regression or the least squares method), or $\ell(y, u) := |y - u|$ (L_1-regression or the least absolute deviations method).

In the case of very large function classes, the empirical risk minimization can easily lead to overfitting and to avoid this various techniques of penalization for complexity of function f have been suggested.

Let $\mathcal{H} := \{h_1, \dots, h_N\}$ be a given set of functions from S into $[-1, 1]$ called *a dictionary.* Given $\lambda \in \mathbb{R}^N$, denote

$$f_\lambda := \sum_{j=1}^{N} \lambda_j h_j.$$

The set

$$\text{l.s.}(\mathcal{H}) := \big\{ f_\lambda \colon \lambda \in \mathbb{R}^N \big\}$$

is the linear span of $\mathcal{H}$. Our main interest is in the case when the cardinality N of the dictionary $\mathcal{H}$ is very large (larger or much larger than the sample size n), and direct minimization of the empirical ℓ-risk over $\text{l.s.}(\mathcal{H})$ can result in overfitting. Because of this, one can consider instead the following penalized empirical risk minimization problem:

$$\hat{\lambda}^{\varepsilon} := \operatorname*{argmin}_{\lambda \in \mathbb{R}^N} \big[P_n(\ell \bullet f_\lambda) + \varepsilon \|\lambda\|_{\ell_p}^p \big], \tag{1.3}$$

where $p > 0$,

$$\|\lambda\|_{\ell_p}^p := \sum_{j=1}^N |\lambda_j|^p,$$

and $\varepsilon > 0$ is a regularization parameter.

Consider the following distribution dependent version of the problem (1.3):

$$\lambda^\varepsilon := \operatorname*{argmin}_{\lambda\in\mathbb{R}^N}\big[P(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_p}^p\big]. \tag{1.4}$$

In particular,

$$\lambda^0 := \operatorname*{argmin}_{\lambda\in\mathbb{R}^N} P(\ell \bullet f_\lambda),$$

and we assume in what follows that λ^0 exists. It is easy to see that for all $\varepsilon \geq 0$ $\|\lambda^\varepsilon\|_{\ell_p} \leq \|\lambda^0\|_{\ell_p}$, and it will be shown (under some assumptions on the loss ℓ) that the norm of the empirical solution $\hat{\lambda}^\varepsilon$ is dominated (up to a constant) by the norm of λ^0. This essentially means that true and empirical solutions are in an ℓ_p-ball of certain radius which will be involved as a parameter in some of our bounds.

For a given $\lambda \in \mathbb{R}^N$, the excess risk of f_λ will be defined as

$$\begin{aligned}\mathcal{E}(f_\lambda) &:= P(\ell \bullet f_\lambda) - P(\ell \bullet f_{\lambda^0}) \\ &= P(\ell \bullet f_\lambda) - \inf_{u\in\mathbb{R}^N} P(\ell \bullet f_u).\end{aligned}$$

In the cases when the true solution λ^0 is sparse in the sense that most of the coefficients λ_j^0 are equal to zero, or at least approximately sparse in the sense that the majority of the coefficients are very small and insignificant, the approach based on penalized empirical risk minimization (1.3) often leads to a satisfactory estimation of λ^0 or f_{λ^0}. The most straightforward choice of p, that provides very strong direct penalization for nonsparsity of the solution, is the limit case $p = 0$. In this case the complexity penalty formally becomes

$$\|\lambda\|_{\ell_p}^p := \sum_{j=1}^N |\lambda_j|^0 = \sum_{j=1}^N I\big(|\lambda_j| > 0\big),$$

which is just the number of nonzero coefficients. Equivalently, one can consider 2^N models corresponding to various choices of a subset $J \subset \{1, \dots, N\}$ and such that $\lambda_j = 0$, $j \notin J$. For each of the models, the empirical risk minimization is performed over the linear span of functions $\{h_j, j \in J\}$ and then one of the models is selected based on minimization of penalized empirical risk with complexity penalty depending on the dimension of the model

$$d(J) := \#(J).$$

It is not hard to analyze the performance of this method in sparse problems using general theory of excess risk bounds and model selection in empirical risk minimization (see, e.g., [1,16,19,20]). However, if N is very large, solving 2^N empirical risk minimization problems becomes computationally intractable (even when each of the problems is convex due to the choice of the loss function ℓ). This difficulty is well known in various model selection problems in statistics (for instance, variable selection in regression models). As an alternative, other values of p have been tried, the value $p = 1$ being by far the most popular. This is related to the fact that it is the smallest value of p that makes the optimization problem (1.3) convex and hence computationally feasible (we are not discussing here any specific algorithms of solving the convex optimization problem (1.3); as a general reference, see [3]). On the other hand, the ℓ_p-norms with small values of p penalize more for nonsparsity, which is the goal here. Also, the solution of (1.3) is inevitably in an ℓ_p-ball of a finite radius. When the dimension N is large and $p > 1$, the ℓ_p-balls are becoming very large while for $p = 1$ the balls are smaller and the empirical processes indexed by such sets are more manageable.

In statistics, the choice $p = 1$ leads to LASSO penalties in regression and is related to soft thresholding in non-parametric statistics (see, e.g., [25]). It possesses a number of attractive properties such as shrinkage of coefficients and "automatic" variable selection. Moreover, recently there has been intensive study of the problem of sparse signal recovery via ℓ_1-norm minimization and a number of interesting results, with some relationship to what is discussed below, have been proved (see [6–8,10–13,21,24]), and further references therein). A typical problem considered in these papers can be formulated (in our notations) as follows. Suppose that for some $\lambda^0 := \vec{\lambda}^0 \in \mathbb{R}^N$

$$Y_j = f_{\lambda^0}(X_j), \quad j = 1, \dots, n.$$

This can be rewritten as a linear system $\vec{y} = A\vec{\lambda}^0$, where $\vec{y}$ is a vector-column of dimension n with entries Y_j and A is an $n \times N$ matrix with entries $h_i(X_j)$. If $N > n$, the system is underdetermined. However, if it has a "sufficiently sparse solution," it was shown that this solution can be found by minimizing the ℓ_1-norm of $\lambda := \vec{\lambda}$ subject to constraints $\vec{y} = A\vec{\lambda}$ (which is a linear programming problem). Moreover, this result has been extended in various directions to allow the noise in the measurements Y_j (see [6] and references therein).

This problem can be also viewed as a version of optimal linear aggregation of given functions $h_1, \dots, h_N$ (that can be statistical estimates based on an independent data set, for instance, pre-trained base classifiers in classification setting). Aggregation techniques, especially in the case of L_2-regression, have been introduced in [23] and [28] and studied extensively in [4,5,9,26,29] among others. Bunea, Tsybakov and Wegkamp [4,5], Koltchinskii [15] and van de Geer [14] also use ℓ_1-type penalization and study the role of sparsity in this type of problems.

We will consider in what follows the values of $p > 1$, but close enough to 1. To be specific, suppose that $p > 1$ and $q > 1$ are conjugate numbers in the sense that $p^{-1} + q^{-1} = 1$. We define q_N such that $N^{1/q_N} = e$. As a result, $q_N = \log N$ and its conjugate

$$p_N = 1 + (\log N - 1)^{-1}.$$

We will use $p \in [1, p_N]$. Note that with this choice of p (by Hölder's inequality)

$$\|\lambda\|_{\ell_p} \le \|\lambda\|_{\ell_1} \le N^{1/q}\|\lambda\|_{\ell_p} = e\|\lambda\|_{\ell_p}, \tag{1.5}$$

so, essentially, ℓ_p-penalization is equivalent to ℓ_1-penalization. However, we will see that the fact that for $p > 1$ the penalty is a strictly convex smooth function of λ gives some advantages in the analysis of the problem. Some of the results below apply to all the values of $p \in [1, p_N]$, but others are true only for p such that $p - 1 \asymp 1/\log N$. It is easy to see that under very mild continuity assumption on the loss function the solution $\hat{\lambda}^{\varepsilon,p}$ of (1.3) converges as $p \to 1$ (for fixed ε, N) to $\hat{\lambda}^{\varepsilon,1}$, which is the LASSO-estimator. However, if $p - 1 \asymp 1/\log N$ (which is of special interest in this paper), then the difference $p - 1$ can be substantial even for large enough values of N. Hence, $\hat{\lambda}^{\varepsilon,p}$ can significantly deviate from the LASSO-estimator if N is large, but $\log N$ is not.

Our main goal is to give precise meaning to the following claim: the "sparsity" of λ^ε implies the "sparsity" of $\hat{\lambda}^\varepsilon$. We also would like to explore how the sparsity of the problem influences the excess risk of $f_{\hat{\lambda}^\varepsilon}$

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) = P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^0}),$$

and also the L_2-errors of estimation of f_{λ^0} and f_{λ^ε} by $f_{\hat{\lambda}^\varepsilon}$ and the ℓ_1-errors of estimation of λ^0 or λ^ε by $\hat{\lambda}^\varepsilon$.

It is not hard to imagine situations in which the solution λ^ε of the "true" problem is sparse. For instance, suppose there exists only a small subset $J \subset \{1, 2, \dots, N\}$ of "relevant features" for prediction of Y (the rest of the features being "irrelevant") such that

$$\mathbb{E}_{J^c} h_j(X) = 0, \quad j \in J^c,$$

where $\mathbb{E}_{J^c}$ denotes the conditional expectation given $Y, h_j(X)$, $j \in J$. In particular, this assumption holds if the sets of random variables

$$\big\{Y, h_j(X), j \in J\big\} \quad \text{and} \quad \big\{h_j(X), j \notin J\big\}$$

are independent and $\mathbb{E}h_j(X) = 0$, $j \notin J$. Then, using Jensen's inequality, we get

$$\begin{aligned}\mathbb{E}\ell\big(Y, f_\lambda(X)\big) + \varepsilon\|\lambda\|_{\ell_p}^p &= \mathbb{E}\mathbb{E}_{J^c}\ell\big(Y, f_\lambda(X)\big) + \varepsilon\|\lambda\|_{\ell_p}^p \geq \mathbb{E}\ell\big(Y, \mathbb{E}_{J^c} f_\lambda(X)\big) + \varepsilon\|\lambda\|_{\ell_p}^p \\ &= \mathbb{E}\ell\bigg(Y, \sum_{j\in J}\lambda_j h_j(X) + \mathbb{E}_{J^c}\sum_{j\in J^c}\lambda_j h_j(X)\bigg) + \varepsilon\sum_{j=1}^N |\lambda_j|^p \\ &\geq \mathbb{E}\ell\bigg(Y, \sum_{j\in J}\lambda_j h_j(X)\bigg) + \varepsilon\sum_{j\in J}|\lambda_j|^p,\end{aligned}$$

and since the last inequality is strict if $\lambda_j \neq 0$ for some $j \in J^c$, it follows that for all $\varepsilon > 0$, $\lambda_j^\varepsilon = 0$, $j \notin J$ (so, λ^ε is "sparse").

Now we introduce the definition of a "sparsity function" of vector $\lambda \in \mathbb{R}^N$: for $d = 0, 1, \ldots, N$, define

$$\gamma_d(\lambda) := \min\bigg\{\sum_{j\notin J}|\lambda_j| : \#(J) = d,\, J \subset \{1, 2, \ldots, N\}\bigg\} = \sum_{j=d+1}^N |\lambda_{[j]}|,$$

where

$$|\lambda_{[1]}| \geq |\lambda_{[2]}| \geq \cdots \geq |\lambda_{[N]}|$$

is a nonincreasing rearrangement of the coefficients. Obviously, $\gamma_d(\lambda)$ is a nonincreasing function of d. If $\gamma_d(\lambda) = 0$, then there are at most d nonzero coordinates of vector $\lambda \in \mathbb{R}^N$ and the corresponding subset $\{h_j\colon\ j \in J\}$ of the dictionary consists of "relevant" functions ("features"). In general, the "sparsity" of λ is characterized by the rate of decay of $\gamma_d(\lambda)$ as d increases and if $\gamma_d(\lambda)$ becomes small for not too large value of d (or, equivalently, if there exists $J \subset \{1, \ldots, N\}$ with $\#(J) = d$ such that $\sum_{j\notin J}|\lambda_j|$ is "small"), it makes sense to call λ "approximately sparse" and still consider sets $\{h_j\colon\ j \in J\}$ of functions ("features") that correspond to the d largest values of $|\lambda_j|$ and thus are "relevant" for representation of f_λ. Such quantities as $\gamma_d(\lambda)$ have been used before (along with some other, more sophisticated, approaches to measuring the complexities of linear or convex combinations) in so-called margin type bounds on generalization error in classification, see [17].

In Section 2, we provide a brief description of the main results of the paper before formulating and proving them in subsequent sections. In Section 3, it is shown that the empirical solution $\hat\lambda^\varepsilon$ belongs to a ball of radius comparable to $\|\lambda^0\|_{\ell_1}$ with a high probability and, based on this fact, preliminary bounds on the excess risk of $f_{\hat\lambda^\varepsilon}$ are established. In Section 4, oracle inequalities for the excess risk with rather strong dependence on the "well posedness" of the dictionary are presented. In Sections 5 and 6, we develop new bounds on excess risk, $L_2(\Pi)$-error and sparsity of the empirical solution with much weaker dependence on the properties of the dictionary. Several technical facts needed in the proofs (primarily, bounds on Rademacher processes) are given in the Appendix.

2. A brief description and discussion of the main results

2.1. *Global and local characteristics of linear independence*

The analysis of the role of sparsity in this type of problem usually requires the assumption of linear independence in $L_2(\Pi)$ of the "relevant" functions $\{h_j\colon\ j \in J\}$ and, moreover, even certain conditions about "weakness" of correlations between the "relevant" functions $\{h_j\colon\ j \in J\}$ and "irrelevant" functions $\{h_j\colon\ j \notin J\}$ in the dictionary (at least in some form). Various versions of such assumptions have been used in recent papers on sparse function reconstruction (see, for instance, "uniform uncertainty principle" or "restricted isometry" conditions in [6], which are assumptions on the $n \times N$ matrix $A = (h_j(X_i))_{i=1,n;\,j=1,N}$), on aggregation of regression estimates [4,5], on aggregation in sparse classification with convex loss [15], or on analysis of LASSO penalization in other risk minimization problems [14]. Obviously, the linear independence assumption (at least of the "relevant" part of the dictionary) is of importance if the vector λ^0 is to be estimated by $\hat\lambda^\varepsilon$: otherwise, λ^0 is not even well defined, so, the problem becomes unidentifiable.

Moreover, in this case, it is necessary to deal with some quantitative measures of identifiability (linear independence) to ensure reasonable accuracy of the solution. However, when the goal is not to recover the vector λ^0, but rather to estimate the function f_{λ^0} in $L_2(\Pi)$-norm, or to ensure that the excess risk of $f_{\hat{\lambda}^\varepsilon}$ is small, the role of the assumptions of this type is more questionable and we will see that some interesting results can be obtained without them.

We describe below several quantitative measures of linear independence that will be used in what follows. First, define for any $J \subset \{1, \dots, N\}$

$$\Gamma(J) := \inf\left\{ \left\| \sum_{j\in J} \alpha_j h_j \right\|_{L_2(\Pi)} : \sum_{j\in J} |\alpha_j| = 1 \right\},$$

which is strictly positive iff $\{h_j : j \in J\}$ are linearly independent in $L_2(\Pi)$.

Secondly, we will use

$$\Gamma_2(J) := \inf\left\{ \left\| \sum_{j\in J} \alpha_j h_j \right\|_{L_2(\Pi)} : \sum_{j\in J} |\alpha_j|^2 = 1 \right\}.$$

Clearly, $\kappa(J) := \Gamma_2^2(J)$ is equal to the smallest eigenvalue of the nonnegatively definite Gram matrix $(\langle h_i, h_j\rangle_{L_2(\Pi)})_{i,j\in J}$.

We will also use a modification of $\Gamma(J)$ that will be denoted $\tilde{\Gamma}(J)$ and is defined as

$$\tilde{\Gamma}(J) := \inf\left\{ \left\| \sum_{j=1}^{N} \alpha_j h_j \right\|_{L_2(\Pi)} : \sum_{j\in J} |\alpha_j| = 1 \right\}.$$

Note that $\tilde{\Gamma}(J)$ depends on the whole dictionary, not just on $\{h_j : j \in J\}$, as $\Gamma(J)$ does. Because of this, one can call $\tilde{\Gamma}(J)$ a *global characteristic* of the dictionary whereas $\Gamma(J)$ and $\Gamma_2(J)$ are its *local characteristics*.

Denote the linear span of $\{h_i : i \in J\}$ by L_J. We will need the following quantity which characterizes the "correlation" (or the angle) between the subspaces L_J and L_{J^c} (compare with the definition of canonical correlations in multivariate statistical analysis):

$$\rho(J) := \sup_{f\in L_J, g\in L_{J^c}, f,g\neq 0} \left| \frac{\langle f, g\rangle_{L_2(\Pi)}}{\|f\|_{L_2(\Pi)}\|g\|_{L_2(\Pi)}} \right|.$$

Finally, we will use one more quantity defined as follows. Given $J \subset \{1, \dots, N\}$, let P_J be the orthogonal projector (in the Hilbert space $L_2(\Pi)$) on L_J. For $j = 1, \dots, N$, let $h'_j := P_J h_j$ and

$$U(J) := \max\left\{ \|h_j - h'_j\|_\infty : 1 \le j \le N \right\}$$

(essentially, $U(J)$ is the maximal L_∞-error of L_2-optimal linear predictors of h_j, $j = 1, \dots, N$, based on $\{h_j : j \in J\}$).

The following simple proposition describes some relationships between these quantities used below (its proof is given in the Appendix).

Proposition 1. *For all nonempty $J \subset \{1, \dots, N\}$,*

(i) $\Gamma(J) \le \frac{1}{\sqrt{d(J)}}$;

(ii) $\Gamma(J) \ge \frac{\Gamma_2(J)}{\sqrt{d(J)}} = \sqrt{\frac{\kappa(J)}{d(J)}}$;

(iii) $1 \ge \Gamma(J) \ge \tilde{\Gamma}(J) \ge \Gamma(J)\sqrt{1-\rho^2(J)}$;

(iv) *for $\kappa = \kappa(\{1, \dots, N\})$,* $\tilde{\Gamma}(J) \ge \sqrt{\frac{\kappa}{d(J)}}$;

(v) $U(J) \le \frac{1}{\Gamma(J)} + 1 \le \frac{\sqrt{d(J)}}{\Gamma_2(J)} + 1 = \sqrt{\frac{d(J)}{\kappa(J)}} + 1$;

(vi) *If h_j, $j = 1, \dots, N$, are orthogonal, then $U(J) \le 1$.*

Note that despite the fact that $U(J)$ is a global characteristic of the dictionary, it is bounded from above by local characteristics (see bound (v) in Proposition 1).

2.2. *Oracle inequalities*

It has become common in Nonparametric Statistics to express optimality properties of statistical estimators in the form of so-called oracle inequalities. In the context of this paper, the ℓ_p penalization with $p \geq 1$ is being used as a "convex relaxation" of ℓ_0-penalization that is the most suitable penalization to recover sparse solutions of the problem, but it is computationally intractable. Therefore, it is desirable to prove oracle inequalities showing that the solution $\hat{\lambda}^\varepsilon$ (for $p = 1$ or for $p > 1$, but close to 1) "mimicks" the ℓ_0-oracle. This is the approach taken by Bunea, Tsybakov and Wegkamp [4,5] in the case of L_2-regression and by van de Geer [14] in the case of other loss functions. We describe some of the results of this type in Section 4. In particular, we show (see Theorem 2) that for $p \in [1, p_N]$ and for all $\varepsilon > 0$

$$\mathcal{E}(f_{\lambda^\varepsilon}) \leq \inf_{\lambda \in \mathbb{R}^N} \left[3\mathcal{E}(f_\lambda) + \frac{2p^2 \|\lambda\|_{\ell_p}^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1} \vee e\|\lambda^0\|_{\ell_1}) \tilde{\Gamma}^2(J_\lambda)} \varepsilon^2 \right] =: \mathcal{O}(\ell_0; \varepsilon),$$

where

$$J_\lambda := \operatorname{supp}(\lambda) = \{j\colon \lambda_j \neq 0\}$$

and τ is defined in (3.1) (the dependence of τ on $\|\lambda\|_{\ell_1} \vee e\|\lambda^0\|_{\ell_1}$ is related to the fact that this quantity provides an upper bound on uniform norms of the functions f_{λ^ε}, f_λ). Moreover, if $\bar{\lambda}$ denotes the value of λ for which the infimum in the expression for $\mathcal{O}(\ell_0; \varepsilon)$ is attained, then

$$\varepsilon \sum_{j \notin J_{\bar{\lambda}}} |\lambda_j^\varepsilon|^p \leq \mathcal{O}(\ell_0; \varepsilon)$$

(which means "approximate sparsity" of λ^ε provided that $\mathcal{O}(\ell_0; \varepsilon)$ is small and the cardinality of $J_{\bar{\lambda}}$ is not very large).

Since (see Proposition 1, (ii) and (iii))

$$\frac{1}{\tilde{\Gamma}^2(J_\lambda)} \leq \frac{d(J_\lambda)}{\kappa(J_\lambda)(1 - \rho^2(J_\lambda))},$$

this result can be interpreted as follows. Suppose there exists a vector $\lambda \in \mathbb{R}^N$ such that

(a) λ is sparse, i.e. $d(J_\lambda)$ is small;
(b) $\kappa(J_\lambda)$ is not very small, i.e. functions $\{h_j : j \in J_\lambda\}$ have a "good" degree of linear independence;
(c) $\rho^2(J_\lambda)$ is bounded away from 1, i.e. subspaces L_{J_λ} and $L_{J_\lambda^c}$ are not "too correlated";
(d) the excess risk $\mathcal{E}(f_\lambda)$ is small.

Then $\mathcal{O}(\ell_0; \varepsilon)$ is small and the solution λ^ε obtained via ℓ_p-penalization has small excess risk and is approximately sparse.

Theorem 3 shows that roughly the same is true, under the assumption that $\varepsilon \geq D\sqrt{\frac{A \log N}{n}}$ and with probability at least $1 - N^{-A}$, for the solution $\hat{\lambda}^\varepsilon$ of the empirical risk minimization problem (1.3) with ℓ_p-penalty, $p \in [1, p_N]$, which provides an oracle inequality on the excess risk of empirical solution. In addition, Corollary 2 gives an upper bound on ℓ_1-distance from the empirical solution to the "oracle" solution.

However, the major drawback of these results is that $\mathcal{O}(\ell_0; \varepsilon)$ cannot be, strictly speaking, viewed as an ℓ_0-oracle since the error term depends on $\frac{1}{\tilde{\Gamma}^2(J_\lambda)}$, or in a more apparent version on $\frac{d(J_\lambda)}{\kappa(J_\lambda)(1-\rho^2(J_\lambda))}$, and not just on the ℓ_0-norm of λ that is equal to $d(J_\lambda)$. Note that in the case of ℓ_0-penalization the error term in the corresponding oracle inequality would be of the order $\frac{d(J_\lambda)}{n}$ (up to a logarithmic factor) which is known to be optimal (see, e.g., [26] for minimax lower bounds in the case of L_2-regression). On the other hand, the oracle of Theorem 3 favors solutions that are not necessarily sparse provided that they are "well posed" in the sense that $\kappa(J_\lambda)$ is not too small and $\rho^2(J_\lambda)$ is not too

close to 1. It is not clear (at least to the author and at least at the moment) to which extent this "well posedness" is really needed and to which extent it is related to the existing methods of the proof of oracle inequalities. Such quantities as $\tilde{\Gamma}(J)$, $\kappa(J)$, $\rho(J)$ heavily depend on the unknown distribution P which makes the performance of the method unpredictable unless one tries to estimate these quantities based on the data. But even if one does it and discovers that the oracle inequalities do not guarantee that the excess risk is small, the question remains whether it means that the ℓ_1-penalization yields a bad solution, or the oracle inequalities available so far are not good enough.

2.3. *Excess risk bounds with weak dependence on the dictionary*

In Sections 5 and 6 (which are the main part of the paper), we explore another approach to the problem based on analyzing separately two parts of the excess risk of $f_{\hat{\lambda}^\varepsilon}$: the random error $P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^\varepsilon})$ and the approximation error $P(\ell \bullet f_{\lambda^\varepsilon}) - P(\ell \bullet f_{\lambda^0})$. It happens that the analysis of the random error relies much less on such characteristics as $\tilde{\Gamma}(J)$, $\kappa(J)$, $\rho(J)$, etc. and it is possible to bound this error by a quantity of optimal order. However, the analysis becomes more complicated and is based on extracting some information about the sparsity of empirical solution $\hat{\lambda}^\varepsilon$ from necessary conditions of extrema in optimization problems (1.3) and (1.4) defining $\hat{\lambda}^\varepsilon$ and λ^ε, respectively. So far, we have been able to complete this analysis only for $p \in (1, p_N]$, excluding a very important case of $p = 1$ (and in some of the statements it is actually required that $p - 1 \asymp (\log N - 1)^{-1}$). It remains to be understood whether these restrictions are really needed and there is some advantage in using the ℓ_p-penalties with $p > 1$, or it is only related to our method of proof (the current version of the method definitely does not work for $p = 1$).

More specifically, for all $p \in (1, p_N]$, we prove (see Theorem 5) that, for $d = d(J)$ and for

$$\varepsilon \geq D\bigl(1 + L\|\lambda^0\|_{\ell_1}\bigr)\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right), \tag{2.1}$$

the assumption $\lambda^\varepsilon_j = 0$, $j \notin J$ implies that with probability at least $1 - N^{-A}$

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\bigr| \leq K^2\frac{d + A\log N}{n} + K\sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau(e\|\lambda^0\|_{\ell_1})}}\sqrt{\frac{d + A\log N}{n}} \tag{2.2}$$

(here D, L are constants depending only on ℓ, $L = 0$ if $\ell'_u(y,u)$ is uniformly bounded, and K is a constant depending on ℓ and on $\|\lambda^0\|_{\ell_1}$). The only dependence of this bound on the "well-posedness" of the dictionary is through the quantity $U(J)$ involved in the expressions for the threshold of ε in (2.1). Moreover, for large enough ε, namely, for

$$\varepsilon \geq D\bigl(1 + L\|\lambda^0\|_{\ell_1}\bigr)\left(\sqrt{\frac{A\log N}{n}} \vee \sqrt{\frac{d}{n}}\right),$$

there is no dependence on the well-posedness of the dictionary whatsoever. Also, for all

$$\varepsilon \geq D\bigl(1 + L\|\lambda^0\|_{\ell_1}\bigr)\sqrt{\frac{A\log N}{n}},$$

the assumption $U(J) \leq \sqrt{\frac{An}{\log N}}$ implies condition (2.1). Since $U(J) \leq \sqrt{\frac{d(J)}{\kappa(J)}} + 1$, this is the case, for instance, if

$$\kappa(J) \geq \frac{d\log N}{n}\left(1 - \sqrt{\frac{\log N}{n}}\right)^{-1},$$

so things can get really bad only if $\kappa(J)$ is really small. Note also that there is no dependence at all on the "irrelevant part" of the dictionary $\{h_j\colon\ j \notin J\}$ (which is the case in Theorem 3 because of the presence of global characteristics, such as $\tilde{\Gamma}(J)$ or $\rho(J)$, in the bounds).

Essentially, these results show that, in the case of ℓ_p-penalization with $p > 1$, the possibility to achieve the optimal size of the error in sparse problems ($\frac{d}{n}$ up to logarithmic factors) is related only to the behavior of the approximation error $\mathcal{E}(f_{\lambda^\varepsilon})$ (if it behaves nicely, so does the excess risk of the empirical solution).

Combining this with simple bounds on approximation error (see Lemma 1) yields the following excess risk bound (see also Corollary 4)

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) \le K^{2p}\frac{d(J)}{\kappa(J)}\frac{A\log N}{n},$$

that holds with probability at least $1 - N^{-A}$ for

$$\varepsilon = D\bigl(1 + L\|\lambda^0\|_{\ell_1}\bigr)\sqrt{\frac{A\log N}{n}}$$

under the assumption that $\lambda^\varepsilon_j = \lambda^0_j = 0$, $j \notin J$.

The assumption that $\lambda^\varepsilon_j = 0$, $j \notin J$ means sparsity of λ^ε (provided that the set J is small). It is more realistic, however, to expect that λ^ε is only "approximately sparse." This situation is analyzed in Section 6. In particular, Theorem 7 provides, for $p = p_N$, the bound

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\bigr| \le \Omega^2 + \Omega\sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau(e\|\lambda^0\|_{\ell_1})}},$$

where

$$\Omega^2 := K^2\frac{d + A\log N}{n} \vee K\sum_{j\notin J}|\lambda^\varepsilon_j|\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right),$$

that holds with probability at least $1 - N^{-A}$ for arbitrary $A \ge 1$ and for all

$$\varepsilon \ge D\bigl(1 + L\|\lambda^0\|_{\ell_1}\bigr)\log N\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right) \tag{2.3}$$

(as before, $d = d(J)$). Note that in (2.3) the threshold of ε is now multiplied by an extra factor of the order $\log N$. It can be seen from the proofs that this $\log N$ is, in fact, $q - 1$, where q is such that $\frac{1}{p} + \frac{1}{q} = 1$. This is what prevented us from extending the result to the values of p closer to 1: when $p \to 1$, the factor $q - 1$ and, thus, the threshold tend to infinity. This factor is annoying since increasing the value of ε might also increase the approximation error $P(\ell \bullet f_{\lambda^\varepsilon}) - P(\ell \bullet f_{\lambda^0})$, which could lead to an extra log-factor in the overall excess risk bound. However, eliminating this factor seems to be beyond the scope of our method. It is not even clear whether it can be eliminated without imposing some "well-posedness" assumptions on the dictionary, something we are trying to avoid here.

The above bound can be combined with the approximation error bound of Theorem 2 to give an oracle inequality close to the one of Theorem 3 (see Corollary 6). However, Theorem 7 does contain more information than this. Essentially, it tells us that if the true solution λ^ε (a) has small excess risk and (b) is "approximately" sparse, then the empirical solution $\hat{\lambda}^\varepsilon$ also possesses these properties, and this basic fact is true with almost no need to assume "well-posedness" of the dictionary. From this point of view, Theorem 2 provides just one possible way to quantify properties (a) and (b) of λ^ε (and this particular way does require some (rather strong) form of "well-posedness" of the dictionary).

The key ingredients of the proofs of these results are general bounds on $\gamma_d(\hat{\lambda}^\varepsilon)$ in terms of $\gamma_d(\lambda^\varepsilon)$ and vice versa, and also the bounds on the L_2-error $\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}$. In particular, we show (Theorem 4) that for $p \in (1, p_N]$, for all $A \ge 1$, for $d = d(J)$, and for all ε satisfying the condition (2.1), the assumption $\lambda^\varepsilon_j = 0$, $j \notin J$, implies that with probability at least $1 - N^{-A}$

$$\gamma_d\bigl(\hat{\lambda}^\varepsilon\bigr) \le \sum_{j\notin J}|\hat{\lambda}^\varepsilon_j| \le K\bigl(1 + L\|\lambda^0\|_{\ell_1}\bigr)\frac{d + A\log N}{n\varepsilon}$$

and

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le K\sqrt{\frac{d + A\log N}{n}}.$$

For $p = p_N$, we extend these "sparsity bounds" to the case when λ^ε is "approximately sparse" (see Theorem 6).

At the end of Section 5, we provide some bounds on the ℓ_1-error of empirical solution. However, a serious study of this problem is beyond the scope of this paper. The problem of estimation of the vector of coefficients λ^0 is not the same as estimation of the function f_{λ^0} (one can expect much stronger dependence on the "well-posedness" of the dictionary in the first problem).

In addition, it is possible to prove that for somewhat larger values of ε with a very high probability the condition $\lambda_j^\varepsilon = 0$ implies that $\hat{\lambda}_j^\varepsilon$ is of the order N^{-B}, where B can be arbitrarily large, and that the coefficients of $\hat{\lambda}^\varepsilon$ and of λ^ε, in some sense, follow the same pattern (see Corollaries 3 and 5).

The methods used in the proofs of the main results are based on Talagrand's concentration inequalities and on the properties of empirical and Rademacher processes, as it has become common in learning theory literature (see, e.g., [2,16]).

Remark. *It is worth mentioning that the form of the inequalities proved in the paper is related to the assumption that the functions in the dictionary are uniformly bounded by* 1. *It is possible to take into account the size of the $L_2(\Pi)$-norm of these functions in the bounds and their $L_2(\Pi_n)$-norms in the definition of the penalty, but we are not pursuing it for the sake of simplicity. More interestingly, it is possible* (*using moment inequalities for Rademacher sums*) *to replace the logarithmic term $A \log N$ involved in the bounds by the following expression*

$$A(q-1)\left(\sum_{j=1}^N \|h_j\|_{L_q(\Pi)}^q\right)^{2/q},$$

where $q = \frac{p}{p-1}$. Note that if the functions h_j are bounded by 1 *and $p = p_N$, then this expression is dominated by $e^2 A \log N$. If the sum $\sum_{j=1}^N \|h_j\|_{L_q(\Pi)}^q$ is of the order N, one is forced to use in the penalty the values of $p \le p_N$ since for large values of p*

$$\left(\sum_{j=1}^N \|h_j\|_{L_q(\Pi)}^q\right)^{2/q}$$

becomes of the order $N^{2/q}$, which is prohibitively large. However, in the case when $\sum_{j=1}^N \|h_j\|_{L_q(\Pi)}^q$ is much smaller than N (*for instance, because many functions in the dictionary have small enough $L_q(\Pi)$-norms*) *it becomes possible to use in the penalty the values of p that are significantly larger than* 1 *and still control the sparsity of the solution. In principle, this opens a possibility of using the empirical $L_q(\Pi_n)$-norms in order to choose the right value of p in the penalty. The analysis of this circle of problems is beyond the scope of the paper.*

3. Preliminary bounds

In what follows, we assume the following properties of the loss function ℓ: for all $y \in T$, $\ell(y,\cdot)$ is twice differentiable, ℓ''_u is a uniformly bounded function in $T \times \mathbb{R}$,

$$\sup_{y\in T} \ell(y;0) < +\infty, \qquad \sup_{y\in T} \left|\ell'_u(y;0)\right| < +\infty$$

and

$$\tau(R) := \frac{1}{2} \inf_{y\in T} \inf_{|u|\le R} \ell''_u(y,u) > 0, \quad R > 0. \tag{3.1}$$

Without loss of generality, we assume that, for all R, $\tau(R) \le 1$ (otherwise, it can be replaced by a lower bound). In particular, these assumptions imply that

$$\left|\ell'_u(y,u)\right| \le L_1 + L|u|, \quad y \in T, u \in \mathbb{R},$$

with some constants $L_1, L \geq 0$, the fact frequently used in what follows. If ℓ'_u is uniformly bounded, one can set $L = 0$.

One of the consequences of these assumptions is that with some constant $C > 0$ depending only on ℓ

$$\tau\big(\|\lambda\|_{\ell_1} \vee \big\|\lambda^0\big\|_{\ell_1}\big)\|f_\lambda - f_{\lambda^0}\|^2_{L_2(\Pi)} \leq \mathcal{E}(f_\lambda) \leq C\|f_\lambda - f_{\lambda^0}\|^2_{L_2(\Pi)},$$

which is also used in the future. There are many important examples of loss functions satisfying these assumptions, most notably, the quadratic loss $\ell(y,u) := (y-u)^2$ in the case when $T \subset \mathbb{R}$ is a bounded set. In this case, $\tau(R) = 1$ for all R. In regression problems with a bounded response variable, one can also consider more general loss functions of the form $\ell(y,u) := \phi(y-u)$, where ϕ is an even nonnegative convex twice continuously differentiable function with ϕ'' uniformly bounded in $\mathbb{R}$, $\phi(0) = 0$ and $\phi''(u) > 0, u \in \mathbb{R}$. In classification setting, one can choose the loss $\ell(y,u) = \phi(yu)$ with ϕ being a nonnegative decreasing convex twice continuously differentiable function such that, again, ϕ'' is uniformly bounded in $\mathbb{R}$ and $\phi''(u) > 0, u \in \mathbb{R}$. The loss function $\phi(u) = \log_2(1 + \mathrm{e}^{-u})$, often called the logit loss, is a typical example.

Many of the results below (for instance, the excess risk bounds of Section 4) are also true in a more general situation, when the minimum of $P(\ell \bullet f_\lambda)$ is not attained in $\mathbb{R}^N$ and the excess risk is defined as

$$\mathcal{E}(f_\lambda) = P(\ell \bullet f_\lambda) - P(\ell \bullet f_*),$$

where f_* is the minimum of the risk function $f \mapsto P(\ell \bullet f)$ over all measurable real valued functions f, as it is common in learning theory literature (see, e.g., [16]). In this case the analysis goes through with some additional work (see [14] where this approach to the ℓ_1-penalization problem is taken).

Finally, it will be assumed throughout the paper that, for some $\gamma > 0$, $N \geq n^\gamma$ and also that $\log N \leq n$ (the last assumption is related only to the fact that in the error terms of the bounds given in the paper the fraction $\frac{\log N}{n}$ is often involved). Some of the constants below might depend on γ (we do not always mention this further). Without these assumptions, the results are still true with some extra terms. Since we are interested in the case of large N, we prefer to impose the condition on N rather than to deal with these minor complications.

Warning. *Throughout the paper, C usually is a constant depending on ℓ and, sometimes, on γ (such that $N \geq n^\gamma$) whose value might change from line to line without further notice.*

We start with several simple observations.

1. For all $\varepsilon_2 \geq \varepsilon_1 \geq 0$,

$$\big\|\lambda^{\varepsilon_2}\big\|_{\ell_p} \leq \big\|\lambda^{\varepsilon_1}\big\|_{\ell_p}.$$

In particular, for all $\varepsilon \geq 0$, $\|\lambda^\varepsilon\|_{\ell_p} \leq \|\lambda^0\|_{\ell_p}$.

Indeed, according to the definitions,

$$\begin{aligned} P(\ell \bullet f_{\lambda^{\varepsilon_2}}) + \varepsilon_2\big\|\lambda^{\varepsilon_2}\big\|^p_{\ell_p} &\leq P(\ell \bullet f_{\lambda^{\varepsilon_1}}) + \varepsilon_2\big\|\lambda^{\varepsilon_1}\big\|^p_{\ell_p} \\ &= P(\ell \bullet f_{\lambda^{\varepsilon_1}}) + \varepsilon_1\big\|\lambda^{\varepsilon_1}\big\|^p_{\ell_p} + (\varepsilon_2 - \varepsilon_1)\big\|\lambda^{\varepsilon_1}\big\|^p_{\ell_p} \\ &\leq P(\ell \bullet f_{\lambda^{\varepsilon_2}}) + \varepsilon_1\big\|\lambda^{\varepsilon_2}\big\|^p_{\ell_p} + (\varepsilon_2 - \varepsilon_1)\big\|\lambda^{\varepsilon_1}\big\|^p_{\ell_p}, \end{aligned}$$

which immediately implies the bound.

2. For all $\varepsilon_2 \geq \varepsilon_1 \geq 0$

$$P(\ell \bullet f_{\lambda^{\varepsilon_2}}) - P(\ell \bullet f_{\lambda^{\varepsilon_1}}) \leq \varepsilon_2\big(\big\|\lambda^{\varepsilon_1}\big\|^p_{\ell_p} - \big\|\lambda^{\varepsilon_2}\big\|^p_{\ell_p}\big) \leq \varepsilon_2\big\|\lambda^{\varepsilon_1}\big\|^p_{\ell_p}.$$

In particular, for all $\varepsilon \geq 0$

$$\mathcal{E}(f_{\lambda^\varepsilon}) = P(\ell \bullet f_{\lambda^\varepsilon}) - P(\ell \bullet f_{\lambda^0}) \leq \varepsilon\big\|\lambda^0\big\|^p_{\ell_p}.$$

The proof immediately follows from the previous bounds (note that statements 1 and 2 are also true for $\hat{\lambda}^{\varepsilon}$ with exactly the same proofs). Under sparsity of $\lambda^{\varepsilon}, \lambda^0$, the last bound can be improved.

Lemma 1. *Suppose that for some $\varepsilon > 0$, some $p \in [1, p_N]$ and for some $J \subset \{1, \dots, N\}$*

$$\lambda_j^{\varepsilon} = \lambda_j^0 = 0, \quad j \notin J.$$

Then

$$\sum_{j \in J} \bigl|\lambda_j^{\varepsilon} - \lambda_j^0\bigr| \le p \frac{\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1}) \Gamma^2(J)} \varepsilon, \tag{3.2}$$

$$\mathcal{E}(f_{\lambda^{\varepsilon}}) = P(\ell \bullet f_{\lambda^{\varepsilon}}) - P(\ell \bullet f_{\lambda^0}) \le p^2 \frac{\|\lambda^0\|_{\ell_p}^{2(p-1)}}{\tau(e\|\lambda^0\|_{\ell_1}) \Gamma^2(J)} \varepsilon^2, \tag{3.3}$$

and

$$\|f_{\lambda^{\varepsilon}} - f_{\lambda^0}\|_{L_2(\Pi)} \le p \frac{\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1}) \Gamma(J)} \varepsilon, \tag{3.4}$$

where the function τ is defined in condition (3.1).

Proof. Indeed,

$$\begin{aligned} P(\ell \bullet f_{\lambda^{\varepsilon}}) - P(\ell \bullet f_{\lambda^0}) &\le \varepsilon\bigl(\bigl\|\lambda^0\bigr\|_{\ell_p}^p - \bigl\|\lambda^{\varepsilon}\bigr\|_{\ell_p}^p\bigr) \le \varepsilon \sum_{j \in J} \bigl(\bigl|\lambda_j^0\bigr|^p - \bigl|\lambda_j^{\varepsilon}\bigr|^p\bigr) \\ &\le \varepsilon p \sum_{j \in J} \bigl|\lambda_j^0\bigr|^{p-1} \bigl|\lambda_j^{\varepsilon} - \lambda_j^0\bigr| \le \varepsilon p \bigl\|\lambda^0\bigr\|_{\ell_p}^{p-1} \sum_{j \in J} \bigl|\lambda_j^{\varepsilon} - \lambda_j^0\bigr|. \end{aligned}$$

Note that for all $\lambda \in \mathbb{R}^N$,

$$\bigl|f_{\lambda}(x)\bigr| \le \sum_{j=1}^N |\lambda_j| \bigl|h_j(x)\bigr| \le \|\lambda\|_{\ell_1} \le e\|\lambda\|_{\ell_p}$$

(we used that, for all j, $\|h_j\|_{\infty} \le 1$). Since λ^0 is the minimal point of $\lambda \mapsto P(\ell \bullet f_{\lambda})$, we have

$$P\bigl(\ell' \bullet f_{\lambda^0}\bigr) h_k = 0, \quad k = 1, \dots, N,$$

and the second-order Taylor expansion and condition (3.1) imply that

$$P(\ell \bullet f_{\lambda^{\varepsilon}}) - P(\ell \bullet f_{\lambda^0}) \ge \tau\bigl(e\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr) \|f_{\lambda^{\varepsilon}} - f_{\lambda^0}\|_{L_2(\Pi)}^2$$

(where we also used that $\|\lambda^{\varepsilon}\|_{\ell_p} \le \|\lambda^0\|_{\ell_p}$). By the definition of $\Gamma(J)$,

$$\|f_{\lambda^{\varepsilon}} - f_{\lambda^0}\|_{L_2(\Pi)} \ge \Gamma(J) \sum_{j \in J} \bigl|\lambda_j^{\varepsilon} - \lambda_j^0\bigr|.$$

Combining these bounds gives

$$\varepsilon p \bigl\|\lambda^0\bigr\|_{\ell_p}^{p-1} \sum_{j \in J} \bigl|\lambda_j^{\varepsilon} - \lambda_j^0\bigr| \ge P(\ell \bullet f_{\lambda^{\varepsilon}}) - P(\ell \bullet f_{\lambda^0}) \ge \tau\bigl(e\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr) \Gamma^2(J) \Biggl(\sum_{j \in J} \bigl|\lambda_j^{\varepsilon} - \lambda_j^0\bigr|\Biggr)^2,$$

which immediately implies (3.2)–(3.4). □

Remarks. 1. *Assuming only sparsity of* λ^0, *i.e., that* $\lambda^0_j = 0$, $j \notin J$, *it is easy to get the following bounds*:

$$\mathcal{E}(f_{\lambda^\varepsilon}) = P(\ell \bullet f_{\lambda^\varepsilon}) - P(\ell \bullet f_{\lambda^0}) \le p^2 \frac{\|\lambda^0\|_{\ell_p}^{2(p-1)}}{\tau(e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J)}\varepsilon^2,$$

$$\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|_{L_2(\Pi)} \le p \frac{\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}(J)}\varepsilon$$

and

$$\|\lambda^\varepsilon - \lambda^0\|_{\ell_1} \le C \frac{\|\lambda^0\|_{\ell_p}^{2(p-1)} \vee \|\lambda^0\|_{\ell_p}^{2(1-1/p)}}{\tau(e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J)}\varepsilon$$

with some constant C.

2. *In the case when* $\lambda^\varepsilon, \lambda^0$ *are only "approximately sparse" the following version of the statements of Lemma* 1 (*with a slight modification of the proof*) *will be of interest. Suppose that* $J \subset \{1, \dots, N\}$ *is such that*

$$\sum_{j \notin J} |\lambda^\varepsilon_j| \le \varepsilon \quad \text{and} \quad \sum_{j \notin J} |\lambda^0_j| \le \varepsilon.$$

Let $d = d(J)$. *Then*

$$\|\lambda^\varepsilon - \lambda^0\|_{\ell_1} \le 6\left[\frac{p\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})\Gamma^2(J)} \vee \left(\frac{p\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})\Gamma^2(J)}\right)^{1/2} \vee \frac{1}{\Gamma(J)} + 2\right]\varepsilon, \tag{3.5}$$

$$P(\ell \bullet f_{\lambda^\varepsilon}) - P(\ell \bullet f_{\lambda^0}) \le 6p\|\lambda^0\|_{\ell_p}^{p-1}\left[\frac{p\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})\Gamma^2(J)} \vee \left(\frac{p\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})\Gamma^2(J)}\right)^{1/2} \vee \frac{1}{\Gamma(J)} + 2\right]\varepsilon^2 \tag{3.6}$$

and

$$\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|_{L_2(\Pi)} \le \sqrt{\frac{6p\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})}\left[\frac{p\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})\Gamma^2(J)} \vee \left(\frac{p\|\lambda^0\|_{\ell_p}^{p-1}}{\tau(e\|\lambda^0\|_{\ell_1})\Gamma^2(J)}\right)^{1/2} \vee \frac{1}{\Gamma(J)} + 2\right]}^{1/2}\varepsilon. \tag{3.7}$$

Our next goal is to derive bounds on $\|\hat{\lambda}^\varepsilon\|_{\ell_1}$ (*which in view of* (1.5) *is the same as to bound* $\|\hat{\lambda}^\varepsilon\|_{\ell_p}$) *and to provide our first bounds on the excess risk that are of the order* $O(\varepsilon)$.

Theorem 1. *There exist constant* $D, L > 0$ *depending only on* ℓ *such that, for all* $p \in [1, p_N]$, *for all* $A \ge 1$ *and for all*

$$\varepsilon \ge D\big(1 + L\|\lambda^0\|_{\ell_1}\big)\sqrt{\frac{A\log N}{n}},$$

the following bound holds

$$\mathbb{R}\big\{\|\hat{\lambda}^\varepsilon\|_{\ell_p} \ge 3^{1/p}\|\lambda^0\|_{\ell_p}\big\} \le N^{-A}. \tag{3.8}$$

Moreover, for some c *depending only on* ℓ *and for all*

$$\varepsilon \ge D\big(1 + L\|\lambda^{\varepsilon/c}\|_{\ell_1}\big)\sqrt{\frac{A\log N}{n}},$$

the following bounds hold

$$\mathbb{P}\big\{\|\hat{\lambda}^\varepsilon\|_{\ell_p} \ge 3^{1/p}\|\lambda^{\varepsilon/c}\|_{\ell_p}\big\} \le N^{-A} \tag{3.9}$$

and

$$\mathbb{R}\bigl\{\bigl\|\hat{\lambda}^{\varepsilon}\bigr\|_{\ell_p} \le 3^{-1/p}\bigl\|\lambda^{c\varepsilon}\bigr\|_{\ell_p}\bigr\} \le N^{-A}. \tag{3.10}$$

As a consequence, with some C depending only on ℓ

$$\mathbb{P}\bigl\{\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) \ge C\bigl(1 + L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\bigl\|\lambda^0\bigr\|_{\ell_1}\varepsilon\bigr\} \le N^{-A},$$

and a similar bound also holds for $P(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) - P(\ell \bullet f_{\lambda^{\varepsilon}})$.

Remarks. 1. *If the loss function ℓ has a bounded derivative with respect to the second variable, one can take $L = 0$.*

2. *It easily follows from the proof that in the condition on ε, needed for (3.8) to hold, one can replace the factor $1 + L\|\lambda^0\|_{\ell_1}$ by the uniform norm $\|\ell' \bullet f_{\lambda^0}\|_{\infty}$ (or even by some Orlicz norms of the same function such as ψ_1-norm). For instance, in the case of L_2-loss in regression problems with bounded noise, this would result in the following condition on ε:*

$$\varepsilon \ge D\bigl(1 + \|f_{\lambda^0} - f_*\|_{\infty}\bigr)\sqrt{\frac{A\log N}{n}},$$

where f_* is the regression function. The difference might be of importance in statistical problems, but we ignore it here and in what follows to simplify the formulation of the results.

Proof of Theorem 1. It follows from the definition of $\hat{\lambda}^{\varepsilon}$ that

$$P_n(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) + \varepsilon\bigl\|\hat{\lambda}^{\varepsilon}\bigr\|_{\ell_p}^p \le P_n(\ell \bullet f_{\lambda}) + \varepsilon\|\lambda\|_{\ell_p}^p, \quad \lambda \in \mathbb{R}^N.$$

By convexity of the function $\lambda \mapsto P_n(\ell \bullet f_{\lambda})$,

$$P_n(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) - P_n(\ell \bullet f_{\lambda}) \ge P_n\bigl(\ell' \bullet f_{\lambda}\bigr)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda}).$$

Therefore,

$$\begin{aligned}\varepsilon\bigl\|\hat{\lambda}^{\varepsilon}\bigr\|_{\ell_p}^p &\le \varepsilon\|\lambda\|_{\ell_p}^p + P_n\bigl(\ell' \bullet f_{\lambda}\bigr)(f_{\lambda} - f_{\hat{\lambda}^{\varepsilon}})\\ &\le \varepsilon\|\lambda\|_{\ell_p}^p + \max_{1\le k\le N}\bigl|P_n\bigl(\ell' \bullet f_{\lambda}\bigr)h_k\bigr|\bigl\|\hat{\lambda}^{\varepsilon} - \lambda\bigr\|_{\ell_1},\end{aligned}$$

which implies

$$\Bigl(\varepsilon - \max_{1\le k\le N}\bigl|P_n\bigl(\ell' \bullet f_{\lambda}\bigr)h_k\bigr|\Bigr)\bigl\|\hat{\lambda}^{\varepsilon}\bigr\|_{\ell_p}^p \le \Bigl(\varepsilon + \max_{1\le k\le N}\bigl|P_n\bigl(\ell' \bullet f_{\lambda}\bigr)h_k\bigr|\Bigr)\|\lambda\|_{\ell_p}^p.$$

Under the assumption

$$\varepsilon > \max_{1\le k\le N}\bigl|P_n\bigl(\ell' \bullet f_{\lambda}\bigr)h_k\bigr|,$$

we get

$$\bigl\|\hat{\lambda}^{\varepsilon}\bigr\|_{\ell_p}^p \le \frac{\varepsilon + \max_{1\le k\le N}|P_n(\ell' \bullet f_{\lambda})h_k|}{\varepsilon - \max_{1\le k\le N}|P_n(\ell' \bullet f_{\lambda})h_k|}\|\lambda\|_{\ell_p}^p. \tag{3.11}$$

Let us set $\lambda := \lambda^0$ in (3.11). It follows from the necessary conditions of the minimum of $\lambda \mapsto P(\ell \bullet f_{\lambda})$ at the point λ^0 that

$$P\bigl(\ell' \bullet f_{\lambda^0}\bigr)h_k = 0, \quad k = 1, \dots, N.$$

Therefore,

$$\max_{1\le k\le N}\bigl|P_n\bigl(\ell'\bullet f_{\lambda^0}\bigr)h_k\bigr| = \max_{1\le k\le N}\bigl|(P_n-P)\bigl(\ell'\bullet f_{\lambda^0}\bigr)h_k\bigr|,$$

which, using Bernstein's inequality, can be bounded further by

$$C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\biggl(\sqrt{\frac{A\log N}{n}}\vee\frac{A\log N}{n}\biggr)$$

with probability at least $1-N^{-A}$. (Recall that under the assumptions on the loss

$$\bigl|\ell'(y;u)\bigr|\le L_1+L|u|$$

with constants L_1, L depending only on ℓ and, hence, the functions $(\ell'\bullet f_{\lambda^0})h_k, k=1,\dots,N$, are uniformly bounded by $L_1+L\|\lambda^0\|_{\ell_1}$.)

As soon as

$$\varepsilon\ge 2C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\biggl(\sqrt{\frac{A\log N}{n}}\vee\frac{A\log N}{n}\biggr),$$

the above bounds yield that with probability at least $1-N^{-A}$

$$\bigl\|\hat\lambda^{\varepsilon}\bigr\|_{\ell_p}^p\le\frac{\varepsilon+\varepsilon/2}{\varepsilon-\varepsilon/2}\bigl\|\lambda^0\bigr\|_{\ell_p}^p=3\bigl\|\lambda^0\bigr\|_{\ell_p}^p.$$

Alternatively, we can use in (3.11) $\lambda:=\lambda^{\varepsilon/c}$. Then, by the necessary conditions of extremum in the definition of $\lambda^{\varepsilon/c}$,

$$\bigl|P\bigl(\ell'\bullet f_{\lambda^{\varepsilon/c}}\bigr)h_k\bigr|\le p\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}^{p-1}\frac{\varepsilon}{c},\quad k=1,\dots,N,$$

which implies

$$\max_{1\le k\le N}\bigl|P_n\bigl(\ell'\bullet f_{\lambda^{\varepsilon/c}}\bigr)h_k\bigr|\le\max_{1\le k\le N}\bigl|(P_n-P)\bigl(\ell'\bullet f_{\lambda^{\varepsilon/c}}\bigr)h_k\bigr|+p\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}^{p-1}\frac{\varepsilon}{c}.$$

Applying Bernstein's inequality to the first term in the right-hand side, we get with probability at least $1-N^{-A}$

$$\max_{1\le k\le N}\bigl|P_n\bigl(\ell'\bullet f_{\lambda^{\varepsilon/c}}\bigr)h_k\bigr|\le C\bigl(1+L\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}\bigr)\biggl(\sqrt{\frac{A\log N}{n}}\vee\frac{A\log N}{n}\biggr)+p\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}^{p-1}\frac{\varepsilon}{c}.$$

Next

$$P(\ell\bullet f_{\lambda^\varepsilon})+\varepsilon\bigl\|\lambda^\varepsilon\bigr\|_{\ell_p}^p\le\mathbb{E}\ell(Y_j,0)\le\sup_y\ell(y,0)=:c_1$$

(since $\mathbb{E}\ell(Y_j,0)$ is the value of the penalized risk at $\lambda=0$). Therefore

$$\bigl\|\lambda^\varepsilon\bigr\|_{\ell_p}\le\biggl(\frac{c_1}{\varepsilon}\biggr)^{1/p}. \tag{3.12}$$

Since $p\le p_N$, $\varepsilon\ge n^{-1/2}$ and $N\ge n^\gamma$, we get

$$p\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}^{p-1}\frac{\varepsilon}{c}\le p\biggl(\frac{e^pc_1c}{\varepsilon}\biggr)^{(p-1)/p}\frac{\varepsilon}{c}\le p\bigl(e^pc_1cn^{1/2}\bigr)^{1/\log N}\frac{\varepsilon}{c}\le\frac{\varepsilon}{4},$$

provided that c is a large enough constant (depending only on ℓ and γ). As soon as

$$\varepsilon \geq 4C\big(1+L\big\|\lambda^{\varepsilon/c}\big\|_{\ell_1}\big)\bigg(\sqrt{\frac{A\log N}{n}}\vee\frac{A\log N}{n}\bigg),$$

this yields

$$\max_{1\leq k\leq N}\big|P_n\big(\ell'\bullet f_{\lambda^{\varepsilon/c}}\big)h_k\big|\leq\frac{\varepsilon}{2},$$

and we can conclude from (3.11) that with probability at least $1-N^{-A}$

$$\big\|\hat{\lambda}^{\varepsilon}\big\|_{\ell_p}^p\leq 3\big\|\lambda^{\varepsilon/c}\big\|_{\ell_p}^p. \tag{3.13}$$

Finally, by the definition of $\lambda^{c\varepsilon}$

$$P(\ell\bullet f_{\lambda^{c\varepsilon}})+c\varepsilon\big\|\lambda^{c\varepsilon}\big\|_{\ell_p}^p\leq P(\ell\bullet f_{\hat{\lambda}^{\varepsilon}})+c\varepsilon\big\|\hat{\lambda}^{\varepsilon}\big\|_{\ell_p}^p.$$

By convexity of the function $\lambda\mapsto P(\ell\bullet f_{\lambda})$,

$$P(\ell\bullet f_{\lambda^{c\varepsilon}})-P(\ell\bullet f_{\hat{\lambda}^{\varepsilon}})\geq P\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)(f_{\lambda^{c\varepsilon}}-f_{\hat{\lambda}^{\varepsilon}}).$$

It follows that

$$c\varepsilon\big\|\lambda^{c\varepsilon}\big\|_{\ell_p}^p\leq c\varepsilon\big\|\hat{\lambda}^{\varepsilon}\big\|_{\ell_p}^p+P\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)(f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{c\varepsilon}})\leq c\varepsilon\big\|\hat{\lambda}^{\varepsilon}\big\|_{\ell_p}^p+\max_{1\leq k\leq N}\big|P\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|\big\|\hat{\lambda}^{\varepsilon}-\lambda^{c\varepsilon}\big\|_{\ell_1},$$

which implies

$$\Big(c\varepsilon-\max_{1\leq k\leq N}\big|P\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|\Big)\big\|\lambda^{c\varepsilon}\big\|_{\ell_p}^p\leq\Big(c\varepsilon+\max_{1\leq k\leq N}\big|P\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|\Big)\big\|\hat{\lambda}^{\varepsilon}\big\|_{\ell_p}^p.$$

If

$$c\varepsilon>\max_{1\leq k\leq N}\big|P\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|,$$

then

$$\big\|\lambda^{c\varepsilon}\big\|_{\ell_p}^p\leq\frac{c\varepsilon+\max_{1\leq k\leq N}|P(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}})h_k|}{c\varepsilon-\max_{1\leq k\leq N}|P(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}})h_k|}\big\|\hat{\lambda}^{\varepsilon}\big\|_{\ell_p}^p. \tag{3.14}$$

Next

$$\max_{1\leq k\leq N}\big|P\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|\leq\max_{1\leq k\leq N}\big|P_n\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|+\max_{1\leq k\leq N}\big|(P_n-P)\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|.$$

To bound the first term, note that similarly to (3.12) we have

$$\big\|\hat{\lambda}^{\varepsilon}\big\|_{\ell_p}\leq\bigg(\frac{c_1}{\varepsilon}\bigg)^{1/p} \tag{3.15}$$

(with c_1 depending only on ℓ) and repeat the argument already used to bound

$$\max_{1\leq k\leq N}\big|P\big(\ell'\bullet f_{\lambda^{\varepsilon/c}}\big)h_k\big|.$$

This gives

$$\max_{1\leq k\leq N}\big|P_n\big(\ell'\bullet f_{\hat{\lambda}^{\varepsilon}}\big)h_k\big|\leq c\frac{\varepsilon}{2},$$

with a properly chosen constant c (depending only on ℓ and on γ). To bound the second term, use Lemma 2 from the Appendix and bound (3.13). This gives (with a proper adjustment of the constants) that with probability at least $1 - N^{-A}$

$$\max_{1\le k\le N}\big|(P_n - P)\big(\ell' \bullet f_{\hat\lambda^\varepsilon}\big)h_k\big| \le C\big(1 + L\big\|\lambda^{\varepsilon/c}\big\|_{\ell_1}\big)\sqrt{\frac{A\log N}{n}}.$$

If

$$c\varepsilon \ge 4C\big(1 + L\big\|\lambda^{\varepsilon/c}\big\|_{\ell_1}\big)\sqrt{\frac{A\log N}{n}},$$

we are getting

$$c\varepsilon > 2\max_{1\le k\le N}\big|P\big(\ell' \bullet f_{\hat\lambda^\varepsilon}\big)h_k\big|,$$

and hence (3.14) implies that with probability at least $1 - N^{-A}$

$$\big\|\lambda^{c\varepsilon}\big\|_{\ell_p}^p \le 3\big\|\hat\lambda^\varepsilon\big\|_{\ell_p}^p.$$

To prove the last bound, note that

$$\begin{aligned}
&P(\ell \bullet f_{\hat\lambda^\varepsilon}) - P(\ell \bullet f_{\lambda^0})\\
&\quad\le P_n(\ell \bullet f_{\hat\lambda^\varepsilon}) + \varepsilon\big\|\hat\lambda^\varepsilon\big\|_{\ell_p}^p - P_n(\ell \bullet f_{\lambda^0}) - \varepsilon\big\|\lambda^0\big\|_{\ell_p}^p + \varepsilon\big\|\lambda^0\big\|_{\ell_p}^p + 2\xi_n\big(\big\|\hat\lambda^\varepsilon\big\|_{\ell_p} \vee \big\|\lambda^0\big\|_{\ell_p}\big)\\
&\quad\le \varepsilon\big\|\lambda^0\big\|_{\ell_1}^p + 2\xi_n\big(\big\|\hat\lambda^\varepsilon\big\|_{\ell_p} \vee \big\|\lambda^0\big\|_{\ell_p}\big),
\end{aligned}$$

where

$$\xi_n(R) := \sup_{\|\lambda\|_{\ell_p}\le R}\big|(P_n - P)(\ell \bullet f_\lambda - \ell \bullet f_0)\big|$$

(here f_0 is f_λ for $\lambda = 0$). Combining the bound of Lemma 2 on $\xi_n(R)$ with the bound $\|\hat\lambda^\varepsilon\|_{\ell_1} \le 3^{1/p}\|\lambda^0\|$ and adjusting the constants, we get that, with probability at least $1 - N^{-A}$, the excess risk is dominated by

$$C\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\big\|\lambda^0\big\|_{\ell_1}\varepsilon + C\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\big\|\lambda^0\big\|_{\ell_1}\sqrt{\frac{A\log N}{n}} \le 2C\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\big\|\lambda^0\big\|_{\ell_1}\varepsilon,$$

implying the result. (More details: if $\|\lambda^0\|_{\ell_1}^p \le C\|\lambda^0\|_{\ell_1}$, the result is immediate. Otherwise, $\|\lambda^0\|_{\ell_1} \ge C^{q-1} \ge N^{\log C}/C$. It easily follows that, for large enough C, $\|\lambda^0\|_{\ell_1}\varepsilon \ge CN$. On the other hand, under the conditions on the loss function ℓ

$$\mathcal{E}(f_{\hat\lambda^\varepsilon}) \le P(\ell \bullet f_{\hat\lambda^\varepsilon}) \le \mathbb{E}\ell(Y;0) + \big(L_1 + L\big\|\hat\lambda^\varepsilon\big\|_{\ell_1}\big)\|\hat\lambda^\varepsilon\|_{\ell_1},$$

which is smaller than CN in view of (3.15) and conditions on ε.) □

4. Oracle inequalities

Throughout the section, it is assumed that $p \in [1, p_N]$. We start with a bound on the approximation error

$$\mathcal{E}(f_{\lambda^\varepsilon}) := P(\ell \bullet f_{\lambda^\varepsilon}) - P(\ell \bullet f_{\lambda^0})$$

(Theorem 2) showing that the ℓ_p-penalization, in some sense, mimics another penalty whose main term is

$$\frac{1}{\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2 \le \frac{d(J_\lambda)}{\kappa(J_\lambda)(1-\rho^2(J_\lambda))}\varepsilon^2,$$

which is closely related to the ℓ_0-penalty, but, at the same time, it heavily depends on the "well-posedness" of the dictionary. After this, we will prove a similar bound on the excess risk of $f_{\hat{\lambda}^\varepsilon}$ (Theorem 3) establishing an oracle inequality in statistical sense. The approximation error bound of Theorem 2 can be viewed as a limit version of the oracle inequality of Theorem 3 when $n \to \infty$ and the empirical solution $\hat{\lambda}^\varepsilon$ becomes λ^ε (however, the constants in Theorem 2 are better). In addition, the theorem shows that the regularized solution λ^ε is "approximately sparse."

Recall that

$$J_\lambda := \operatorname{supp}(\lambda) = \{j\colon \lambda_j \neq 0\}.$$

Theorem 2. *For all $\varepsilon > 0$,*

$$\mathcal{E}(f_{\lambda^\varepsilon}) \le \inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{2p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\right].$$

Moreover, if $\bar{\lambda}$ denotes the value of λ for which the infimum in the right-hand side is attained, then

$$\varepsilon\sum_{j\notin J_{\bar{\lambda}}}\bigl|\lambda_j^\varepsilon\bigr|^p \le \frac{1}{2}\inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{2p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\right].$$

Proof. By the definition of λ^ε, for all $\lambda\in\mathbb{R}^N$,

$$P(\ell\bullet f_{\lambda^\varepsilon}) + \varepsilon\bigl\|\lambda^\varepsilon\bigr\|_{\ell_p}^p \le P(\ell\bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_p}^p,$$

which implies

$$\begin{aligned}\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon\sum_{j\notin J_\lambda}\bigl|\lambda_j^\varepsilon\bigr|^p &\le \mathcal{E}(f_\lambda) + \varepsilon\sum_{j\in J_\lambda}\bigl(|\lambda_j|^p - \bigl|\lambda_j^\varepsilon\bigr|^p\bigr)\\ &\le \mathcal{E}(f_\lambda) + p\varepsilon\|\lambda\|_{\ell_p}^{p-1}\sum_{j\in J_\lambda}\bigl|\lambda_j - \lambda_j^\varepsilon\bigr|.\end{aligned}$$

Then, using the bounds

$$\sum_{j\in J_\lambda}\bigl|\lambda_j - \lambda_j^\varepsilon\bigr| \le \frac{1}{\tilde{\Gamma}(J_\lambda)}\|f_\lambda - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le \frac{1}{\tilde{\Gamma}(J)}\bigl(\|f_\lambda - f_{\lambda^0}\|_{L_2(\Pi)} + \|f_{\lambda^\varepsilon} - f_{\lambda^0}\|_{L_2(\Pi)}\bigr)$$

and (see the proof of Lemma 1)

$$\mathcal{E}(f_\lambda) \ge \tau\|f_\lambda - f_{\lambda^0}\|^2, \qquad \mathcal{E}(f_{\lambda^\varepsilon}) \ge \tau\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|^2,$$

where

$$\tau := \tau\bigl(\|\lambda\|_{\ell_1}\vee e\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr),$$

we get

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon\sum_{j\notin J_\lambda}\bigl|\lambda_j^\varepsilon\bigr|^p \le \mathcal{E}(f_\lambda) + \frac{p\|\lambda\|_{\ell_p}^{p-1}}{\tilde{\Gamma}(J_\lambda)}\varepsilon\left(\sqrt{\frac{\mathcal{E}(f_\lambda)}{\tau}} + \sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau}}\right).$$

Using the inequality $ab \le a^2/2 + b^2/2$, we get

$$\frac{p\|\lambda\|_{\ell_p}^{p-1}}{\sqrt{\tau}\tilde{\Gamma}(J_\lambda)}\varepsilon\sqrt{\mathcal{E}(f_\lambda)} \le \frac{\mathcal{E}(f_\lambda)}{2} + \frac{p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{2\tau\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2$$

and, similarly,

$$\frac{p\|\lambda\|_{\ell_p}^{p-1}}{\sqrt{\tau}\tilde{\Gamma}(J_\lambda)}\varepsilon\sqrt{\mathcal{E}(f_{\lambda^\varepsilon})} \le \frac{\mathcal{E}(f_{\lambda^\varepsilon})}{2} + \frac{p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{2\tau\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2.$$

As a consequence,

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon\sum_{j\notin J_\lambda}|\lambda_j^\varepsilon|^p \le \mathcal{E}(f_\lambda) + \frac{\mathcal{E}(f_{\lambda^\varepsilon})}{2} + \frac{p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{2\tau\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2 + \frac{\mathcal{E}(f_\lambda)}{2} + \frac{p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{2\tau\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2,$$

which implies

$$\mathcal{E}(f_{\lambda^\varepsilon}) + 2\varepsilon\sum_{j\notin J_\lambda}|\lambda_j^\varepsilon|^p \le 3\mathcal{E}(f_\lambda) + \frac{2p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{\tau\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2.$$

Since it holds for all $\lambda \in \mathbb{R}^N$, we get

$$\mathcal{E}(f_{\lambda^\varepsilon}) \le \inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{2p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{\tau\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\right]$$

and also

$$\varepsilon\sum_{j\notin J_{\tilde{\lambda}}}|\lambda_j^\varepsilon|^p \le \frac{1}{2}\inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{2p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{\tau\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\right],$$

which implies the result. □

Remark. *It easily follows from the second bound of Theorem 2 that with some constant c*

$$\varepsilon\sum_{j\notin J_{\tilde{\lambda}}}|\lambda_j^\varepsilon| \le c\inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{2p^2(\|\lambda\|_{\ell_p}\vee 1)^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\right]. \tag{4.1}$$

Indeed, assuming that the infimum in the right-hand side is attained at $\tilde{\lambda}$, we get

$$\sum_{j\notin J_{\tilde{\lambda}}}|\lambda_j^\varepsilon| \le N^{1/q}\left(\sum_{j\notin J_{\tilde{\lambda}}}|\lambda_j^\varepsilon|^p\right)^{1/p} \le \frac{e}{2^{1/p}}\left[3\mathcal{E}(f_{\tilde{\lambda}}) + \frac{2p^2(\|\tilde{\lambda}\|_{\ell_p}\vee 1)^{2(p-1)}}{\tau(\|\tilde{\lambda}\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J_{\tilde{\lambda}})}\varepsilon^2\right]^{1/p}.$$

Since

$$\tau\big(\|\tilde{\lambda}\|_{\ell_1}\vee e\big\|\lambda^0\big\|_{\ell_1}\big)\tilde{\Gamma}^2(J_{\tilde{\lambda}}) \le 1,$$

the term in brackets is larger than $(2p^2\|\tilde{\lambda}\|_{\ell_p}\vee 1)^{2(p-1)}\varepsilon^2$. Therefore

$$\sum_{j\notin J_{\tilde{\lambda}}}|\lambda_j^\varepsilon| \le \frac{e}{2^{1/p}}\big(2p^2\big(\|\tilde{\lambda}\|_{\ell_p}\vee 1\big)^{2(p-1)}\varepsilon^2\big)^{1/p-1}\left[3\mathcal{E}(f_{\tilde{\lambda}}) + \frac{2p^2(\|\tilde{\lambda}\|_{\ell_p}\vee 1)^{2(p-1)}}{\tau(\|\tilde{\lambda}\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J_{\tilde{\lambda}})}\varepsilon^2\right].$$

It is easily seen that under the assumptions on n, N, p, ε

$$\frac{e}{2^{1/p}}\bigl(2p^2\bigl(\|\tilde{\lambda}\|_{\ell_p}\vee 1\bigr)^{2(p-1)}\varepsilon^2\bigr)^{1/p-1}\le c$$

with some constant c, which implies the claim. The following corollary is now immediate:

Corollary 1. *For all $\varepsilon>0$ and for $\bar{\lambda}$ for which the infimum in the right-hand side of (4.2) is attained,*

$$\varepsilon\bigl\|\lambda^{\varepsilon}-\bar{\lambda}\bigr\|_{\ell_1}\le C\inf_{\lambda\in\mathbb{R}^N}\biggl[3\mathcal{E}(f_\lambda)+\frac{2p^2(\|\lambda\|_{\ell_p}\vee 1)^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\biggr] \tag{4.2}$$

with some constant $C>0$.

Proof. It is enough to write

$$\sum_{j\in J_{\bar{\lambda}}}\bigl|\bar{\lambda}_j-\lambda^{\varepsilon}_j\bigr|\le\frac{1}{\tilde{\Gamma}(J_{\bar{\lambda}})}\|f_{\bar{\lambda}}-f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}\le\frac{1}{\tilde{\Gamma}(J_{\bar{\lambda}})}\bigl(\|f_{\bar{\lambda}}-f_{\lambda^0}\|_{L_2(\Pi)}+\|f_{\lambda^{\varepsilon}}-f_{\lambda^0}\|_{L_2(\Pi)}\bigr)$$

and to bound it further as in the proof of Theorem 2 to get

$$\sum_{j\in J_{\bar{\lambda}}}\bigl|\bar{\lambda}_j-\lambda^{\varepsilon}_j\bigr|\le\frac{1}{2}\biggl(\frac{\mathcal{E}(f_{\bar{\lambda}})}{\varepsilon}+\frac{\varepsilon}{\tau\tilde{\Gamma}^2(J_{\bar{\lambda}})}+\frac{\mathcal{E}(f_{\lambda^{\varepsilon}})}{\varepsilon}+\frac{\varepsilon}{\tau\tilde{\Gamma}^2(J_{\bar{\lambda}})}\biggr),$$

where

$$\tau:=\tau\bigl(\|\bar{\lambda}\|_{\ell_1}\vee e\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr).$$

It remains to use the first inequality of Theorem 2, to combine it with (4.1) and to choose the constant C properly. □

Next we extend the bounds of Theorem 2 to the empirical solution $\hat{\lambda}^{\varepsilon}$. For sufficiently large values of ε,

$$\varepsilon\ge D\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\sqrt{\frac{A\log N}{n}}$$

(with constants D, L depending only on ℓ and $L=0$ in the case of the loss with bounded derivative), this provides an oracle inequality in the spirit of [4,5] and [14] (the current version is especially close to the one of [14]), and also the bound showing the "approximate sparsity" of $\hat{\lambda}^{\varepsilon}$. In Section 6, a different approach to this type of result is discussed.

Theorem 3. *Let $p\in[1,p_N]$. There exist constants $D, C, L>0$, depending only on ℓ, such that for all $A\ge 1$ and for all*

$$\varepsilon\ge D\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\sqrt{\frac{A\log N}{n}},$$

the following bound holds with probability at least $1-N^{-A}$:

$$\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}})\le\inf_{\lambda\in\mathbb{R}^N}\biggl[3\mathcal{E}(f_\lambda)+\frac{\mathcal{U}(\lambda,\lambda^0,C)}{\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\biggr], \tag{4.3}$$

where

$$\mathcal{U}\bigl(\lambda,\lambda^0,C\bigr):=\frac{2(p\|\lambda\|_{\ell_p}^{p_N-1}+C(1+L\|\lambda\|_{\ell_1}/(1+L\|\lambda^0\|_{\ell_1})))^2}{\tau(\|\lambda\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})}.$$

Moreover, if $\bar{\lambda}$ denotes the value of λ for which the infimum in the right-hand side is attained, then with the same probability

$$\varepsilon \sum_{j\notin J_{\bar{\lambda}}} |\hat{\lambda}_j^{\varepsilon}| \le C \inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{\mathcal{U}(\lambda,\lambda^0,C)}{\Gamma^2(J_\lambda)}\varepsilon^2\right].$$

Proof. By the definition of $\hat{\lambda}^{\varepsilon}$, for all $\lambda\in\mathbb{R}^N$,

$$P_n(\ell\bullet f_{\hat{\lambda}^{\varepsilon}}) + \varepsilon\|\hat{\lambda}^{\varepsilon}\|_{\ell_p}^p \le P_n(\ell\bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_p}^p,$$

which implies

$$\begin{aligned}\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) + \varepsilon\sum_{j\notin J_\lambda}|\hat{\lambda}_j^{\varepsilon}|^p &\le \mathcal{E}(f_\lambda) + \varepsilon\sum_{j\in J_\lambda}\bigl(|\lambda_j|^p - |\hat{\lambda}_j^{\varepsilon}|^p\bigr) + \bigl|(P_n-P)(\ell\bullet f_{\hat{\lambda}^{\varepsilon}} - \ell\bullet f_\lambda)\bigr| \\ &\le \mathcal{E}(f_\lambda) + p\varepsilon\|\lambda\|_{\ell_p}^{p-1}\sum_{j\in J_\lambda}|\lambda_j - \hat{\lambda}_j^{\varepsilon}| + \bigl|(P_n-P)(\ell\bullet f_{\hat{\lambda}^{\varepsilon}} - \ell\bullet f_\lambda)\bigr|.\end{aligned}$$

We will apply this bound to $\lambda=\bar{\lambda}$. By Lemma 3, we have with probability at least $1-N^{-A}$

$$\begin{aligned}\bigl|(P_n-P)(\ell\bullet f_{\hat{\lambda}^{\varepsilon}} - \ell\bullet f_{\bar{\lambda}})\bigr| &\le C\bigl(1+L\bigl(\|\lambda^0\|_{\ell_1}\vee\|\bar{\lambda}\|_{\ell_1}\bigr)\bigr)\Biggl(\sum_{j\in J_{\bar{\lambda}}}|\hat{\lambda}_j^{\varepsilon}-\bar{\lambda}_j^{\varepsilon}|\vee e^{-N}\Biggr)\sqrt{\frac{A\log N}{n}} \\ &\quad + C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\Biggl(\sum_{j\notin J_{\bar{\lambda}}}|\hat{\lambda}_j^{\varepsilon}|\vee e^{-N}\Biggr)\sqrt{\frac{A\log N}{n}},\end{aligned}$$

provided that

$$\|\hat{\lambda}^{\varepsilon}-\bar{\lambda}\|_{\ell_1}\le e^N.$$

If the constant D in the condition on ε is large enough, then, under the assumptions on N,n and ε the right-hand side is dominated by

$$\frac{\varepsilon}{C}\sum_{j\notin J_{\bar{\lambda}}}|\hat{\lambda}_j^{\varepsilon}| + C\Biggl(\frac{1+L\|\bar{\lambda}\|_{\ell_1}}{(1+L\|\lambda^0\|_{\ell_1})}\Biggr)\varepsilon\sum_{j\in J_{\bar{\lambda}}}|\hat{\lambda}_j^{\varepsilon}-\bar{\lambda}_j^{\varepsilon}| + C\varepsilon^2$$

with some constant C.

Therefore, we have

$$\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) + \varepsilon\sum_{j\notin J_{\bar{\lambda}}}|\hat{\lambda}_j^{\varepsilon}|^p \le \mathcal{E}(f_{\bar{\lambda}}) + \Biggl(p\|\bar{\lambda}\|_{\ell_p}^{p-1} + C\Biggl(\frac{1+L\|\bar{\lambda}\|_{\ell_1}}{1+L\|\lambda^0\|_{\ell_1}}\Biggr)\Biggr)\varepsilon\sum_{j\in J_{\bar{\lambda}}}|\bar{\lambda}_j-\hat{\lambda}_j^{\varepsilon}| + \frac{\varepsilon}{C}\sum_{j\notin J_{\bar{\lambda}}}|\hat{\lambda}_j^{\varepsilon}| + C\varepsilon^2,$$

which implies

$$\begin{aligned}&\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) + \varepsilon\sum_{j\notin J_{\bar{\lambda}}}\Biggl(|\hat{\lambda}_j^{\varepsilon}|^p - \frac{1}{C}|\hat{\lambda}_j^{\varepsilon}|\Biggr) \\ &\quad\le \mathcal{E}(f_{\bar{\lambda}}) + \Biggl(p\|\bar{\lambda}\|_{\ell_p}^{p-1} + C\Biggl(1+\frac{L\|\bar{\lambda}\|_{\ell_1}}{1+L\|\lambda^0\|_{\ell_1}}\Biggr)\Biggr)\varepsilon\sum_{j\in J_{\bar{\lambda}}}|\bar{\lambda}_j-\hat{\lambda}_j^{\varepsilon}| + C\varepsilon^2.\end{aligned}$$

Note that

$$\sum_{j\notin J_{\bar\lambda}}\Big(|\hat\lambda_j^\varepsilon|^p-\frac{1}{C}|\hat\lambda_j^\varepsilon|\Big)\geq \sum_{j\notin J_{\bar\lambda},|\hat\lambda_j^\varepsilon|\geq N^{-B}}\Big(|\hat\lambda_j^\varepsilon|^p-\frac{1}{C}|\hat\lambda_j^\varepsilon|\Big)-\frac{1}{C}N^{-B+1}$$

$$\geq \sum_{j\notin J_{\bar\lambda},|\hat\lambda_j^\varepsilon|\geq N^{-B}}\big(N^{-B(p-1)}-C^{-1}\big)|\hat\lambda_j^\varepsilon|-\frac{1}{C}N^{-B+1}$$

$$\geq \big(\mathrm{e}^{-B(\log N)/(\log N-1)}-C^{-1}\big)\sum_{j\notin J_{\bar\lambda},|\hat\lambda_j^\varepsilon|\geq N^{-B}}|\hat\lambda_j^\varepsilon|-\frac{2}{C}N^{-B+1}$$

$$\geq \frac{1}{C}\sum_{j\notin J_{\bar\lambda}}|\hat\lambda_j^\varepsilon|-\frac{2}{C}N^{-B+1},$$

with a proper choice of large enough constants B, C such that

$$\mathrm{e}^{-B(\log N)/(\log N-1)}-C^{-1}\geq C^{-1}.$$

Therefore,

$$\mathcal{E}(f_{\hat\lambda^\varepsilon})+\frac{\varepsilon}{C}\sum_{j\notin J_{\bar\lambda}}|\hat\lambda_j^\varepsilon|\leq \mathcal{E}(f_{\bar\lambda})+\bigg(p\|\bar\lambda\|_{\ell_p}^{p-1}+C\bigg(\frac{1+L\|\bar\lambda\|_{\ell_1}}{(1+L\|\lambda^0\|_{\ell_1})}\bigg)\bigg)\varepsilon\sum_{j\in J_{\bar\lambda}}|\bar\lambda_j-\hat\lambda_j^\varepsilon|+C\varepsilon^2.$$

It remains to repeat the argument of Theorem 2 with minor modifications of the constants. In the case when

$$\big\|\hat\lambda^\varepsilon-\bar\lambda\big\|_{\ell_1}>e^N,$$

we have

$$\|\bar\lambda\|_{\ell_1}>e^N-\big\|\hat\lambda^\varepsilon\big\|_{\ell_1}$$

and since (see bound (3.15) in the proof of Theorem 1)

$$\big\|\hat\lambda^\varepsilon\big\|_{\ell_1}\leq e\bigg(\frac{c_1}{\varepsilon}\bigg)^{1/p},$$

we easily get under the conditions on ε, n, N (in particular, under the assumption that $N\geq n^\gamma$) that

$$\|\bar\lambda\|_{\ell_1}>\frac{e^N}{c}$$

(with a proper choice of c). Then a simple computation shows that under the conditions on the loss function ℓ

$$\mathcal{E}(f_{\hat\lambda^\varepsilon})\leq P(\ell\bullet f_{\hat\lambda^\varepsilon})\leq \mathbb{E}\ell(Y;0)+C\big(1+L\big\|\hat\lambda^\varepsilon\big\|_{\ell_1}\big)\big\|\hat\lambda^\varepsilon\big\|_{\ell_1},$$

which, again by the bound (3.15) and the conditions on ε is bounded by an expression of the form $C_1+C_2n^{1/p}$. It is now easy to show that this is smaller than

$$\frac{\mathcal{U}(\bar\lambda,\lambda^0,C)}{\tilde\Gamma^2(J_{\bar\lambda})}\varepsilon^2.$$

Indeed, since

$$\tau\big(\|\bar\lambda\|_{\ell_1}\vee e\big\|\lambda^0\big\|_{\ell_1}\big)\leq 1,\qquad \tilde\Gamma^2(J_{\bar\lambda})\leq 1,\qquad \varepsilon^2\geq n^{-1}$$

and since $N \geq n^{\gamma}$, we have

$$\frac{\mathcal{U}(\bar{\lambda}, \lambda^0, C)}{\tilde{\Gamma}^2(J_{\bar{\lambda}})}\varepsilon^2 \geq \|\bar{\lambda}\|_{\ell_p}^{2(p_N-1)} n^{-1} \geq \exp\left\{\frac{2N}{(\log N - 1)}\right\} c^{-2(p_N-1)} n^{-1} \geq C_1 + C_2 n^{1/p}$$

with a proper choice of constants in the above bounds. This implies that (4.3) also holds when $\|\hat{\lambda}^{\varepsilon} - \bar{\lambda}\|_{\ell_1} > e^N$. The last inequality is trivial in this case. □

Corollary 2. *Under the assumptions of Theorem 3, with probability at least* $1 - N^{-A}$

$$\varepsilon \|\hat{\lambda}^{\varepsilon} - \bar{\lambda}\|_{\ell_1} \leq C \inf_{\lambda \in \mathbb{R}^N} \left[3\mathcal{E}(f_{\lambda}) + \frac{\mathcal{U}(\lambda, \lambda^0, C)}{\tilde{\Gamma}^2(J_{\lambda})}\varepsilon^2 \right] \tag{4.4}$$

with some constant $C > 0$ *and with* $\bar{\lambda}$ *that minimizes the infimum in the right-hand side of* (4.4).

Proof. See the proof of Corollary 1. Use Theorem 3 instead of Theorem 2. □

5. Sparsity and excess risk bounds

In this section, we derive the bounds on $P(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) - P(\ell \bullet f_{\lambda^{\varepsilon}})$ and on $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$ under the assumption that λ^{ε} is sparse, i.e. there exists a "small" set $J \subset \{1, \dots, N\}$ such that $\lambda_j^{\varepsilon} = 0$, $j \notin J$. It will be assumed throughout the section that $p \in (1, p_N]$ (so, the results do not apply to the case of $p = 1$). Our first goal is to provide bounds on $\sum_{j \notin J} |\hat{\lambda}_j^{\varepsilon}|$ and on $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$.

The following relationships, that are based on necessary conditions of minima in minimization problems defining λ^{ε} and $\hat{\lambda}^{\varepsilon}$, will be crucial in our derivation of sparsity bounds in this and in the next sections:

$$P\big(\ell' \bullet f_{\lambda^{\varepsilon}}\big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}) + p\varepsilon \sum_{j=1}^{N} |\lambda_j^{\varepsilon}|^{p-1} \operatorname{sign}\big(\lambda_j^{\varepsilon}\big)\big(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon}\big) = 0 \tag{5.1}$$

and

$$P_n\big(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}) + p\varepsilon \sum_{j=1}^{N} |\hat{\lambda}_j^{\varepsilon}|^{p-1} \operatorname{sign}\big(\hat{\lambda}_j^{\varepsilon}\big)\big(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon}\big) = 0. \tag{5.2}$$

They imply that

$$\begin{aligned}\mathcal{D}\big(\hat{\lambda}^{\varepsilon}; \lambda^{\varepsilon}\big) &:= P\big(\big(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\big) - \big(\ell' \bullet f_{\lambda^{\varepsilon}}\big)\big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}) + p\varepsilon \sum_{j=1}^{N} \big(|\hat{\lambda}_j^{\varepsilon}|^{p-1} \operatorname{sign}\big(\hat{\lambda}_j^{\varepsilon}\big) - |\lambda_j^{\varepsilon}|^{p-1} \operatorname{sign}\big(\lambda_j^{\varepsilon}\big)\big)\big(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon}\big) \\ &= (P - P_n)\big(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}). \end{aligned} \tag{5.3}$$

Note that $\mathcal{D}(\hat{\lambda}^{\varepsilon}; \lambda^{\varepsilon}) \geq 0$ since $\ell'(y; \cdot)$ is nondecreasing (by convexity of $\ell(y, \cdot)$) and $\lambda \mapsto p|\lambda|^{p-1} \operatorname{sign}(\lambda)$ is also nondecreasing (by convexity of $\lambda \mapsto |\lambda|^p$ for $p > 1$).

Quite similarly,

$$\begin{aligned} 0 \leq \check{\mathcal{D}}\big(\hat{\lambda}^{\varepsilon}; \lambda^{\varepsilon}\big) &:= P_n\big(\big(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\big) - \big(\ell' \bullet f_{\lambda^{\varepsilon}}\big)\big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}) \\ &\quad + p\varepsilon \sum_{j=1}^{N} \big(|\hat{\lambda}_j^{\varepsilon}|^{p-1} \operatorname{sign}\big(\hat{\lambda}_j^{\varepsilon}\big) - |\lambda_j^{\varepsilon}|^{p-1} \operatorname{sign}\big(\lambda_j^{\varepsilon}\big)\big)\big(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon}\big) \\ &= (P_n - P)\big(\ell' \bullet f_{\lambda^{\varepsilon}}\big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}). \end{aligned} \tag{5.4}$$

We will have to analyze more carefully the expression for $\mathcal{D}(\hat{\lambda}^\varepsilon;\lambda^\varepsilon)$ and for $\check{\mathcal{D}}(\hat{\lambda}^\varepsilon;\lambda^\varepsilon)$ since this is where the information about the sparsity of $\hat{\lambda}^\varepsilon$ is contained.

First note that $\|\lambda^\varepsilon\|_{\ell_1}\le e\|\lambda^\varepsilon\|_{\ell_p}\le e\|\lambda^0\|_{\ell_p}\le e\|\lambda^0\|_{\ell_1}$. Therefore, $|f_{\lambda^\varepsilon}(x)|\le e\|\lambda^0\|_{\ell_1}$ (recall that h_j are bounded by 1). Similarly, if $\|\hat{\lambda}^\varepsilon\|_{\ell_1}\le C\|\lambda^0\|_{\ell_1}$ (which by Theorem 1 holds with probability $\ge 1-N^{-A}$), we have $|f_{\hat{\lambda}^\varepsilon}(x)|\le C\|\lambda^0\|_{\ell_1}$. Using condition (3.1) on the loss ℓ, this allows us to show that

$$P\big(\big(\ell'\bullet f_{\hat{\lambda}^\varepsilon}\big)-\big(\ell'\bullet f_{\lambda^\varepsilon}\big)\big)(f_{\hat{\lambda}^\varepsilon}-f_{\lambda^\varepsilon})\ge b\|f_{\hat{\lambda}^\varepsilon}-f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)}$$

and

$$P_n\big(\big(\ell'\bullet f_{\hat{\lambda}^\varepsilon}\big)-\big(\ell'\bullet f_{\lambda^\varepsilon}\big)\big)(f_{\hat{\lambda}^\varepsilon}-f_{\lambda^\varepsilon})\ge b\|f_{\hat{\lambda}^\varepsilon}-f_{\lambda^\varepsilon}\|^2_{L_2(\Pi_n)}$$

with $b=2\tau(C\|\lambda^0\|_{\ell_1})$ (see condition (3.1)).

Next, if $\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j\le 0$, then

$$\big(|\hat{\lambda}^\varepsilon_j|^{p-1}\operatorname{sign}\big(\hat{\lambda}^\varepsilon_j\big)-|\lambda^\varepsilon_j|^{p-1}\operatorname{sign}\big(\lambda^\varepsilon_j\big)\big)\big(\hat{\lambda}^\varepsilon_j-\lambda^\varepsilon_j\big)=\big(|\hat{\lambda}^\varepsilon_j|^{p-1}+|\lambda^\varepsilon_j|^{p-1}\big)\big(|\hat{\lambda}^\varepsilon_j|+|\lambda^\varepsilon_j|\big)\ge|\hat{\lambda}^\varepsilon_j|^p+|\lambda^\varepsilon_j|^p.$$

On the other hand, if $\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j>0$ and $|\hat{\lambda}^\varepsilon_j|\ge 2|\lambda^\varepsilon_j|$, we have

$$\big(|\hat{\lambda}^\varepsilon_j|^{p-1}\operatorname{sign}\big(\hat{\lambda}^\varepsilon_j\big)-|\lambda^\varepsilon_j|^{p-1}\operatorname{sign}\big(\lambda^\varepsilon_j\big)\big)\big(\hat{\lambda}^\varepsilon_j-\lambda^\varepsilon_j\big)\ge\frac{1-2^{-(p-1)}}{2}|\hat{\lambda}^\varepsilon_j|^p\ge\frac{(p-1)\log 2}{4}|\hat{\lambda}^\varepsilon_j|^p.$$

Therefore,

$$\begin{aligned}\mathcal{D}\big(\hat{\lambda}^\varepsilon;\lambda^\varepsilon\big)&\ge b\|f_{\hat{\lambda}^\varepsilon}-f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)}+p\varepsilon\sum_{j=1}^N\big(|\hat{\lambda}^\varepsilon_j|^p+|\lambda^\varepsilon_j|^p\big)I\big(\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j\le 0\big)\\&\quad+\frac{p(p-1)\log 2}{4}\varepsilon\sum_{j=1}^N|\hat{\lambda}^\varepsilon_j|^pI\big(\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j>0,\,|\hat{\lambda}^\varepsilon_j|\ge 2|\lambda^\varepsilon_j|\big)\\&\quad+\frac{p(p-1)\log 2}{4}\varepsilon\sum_{j=1}^N|\lambda^\varepsilon_j|^pI\big(\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j>0,\,|\lambda^\varepsilon_j|\ge 2|\hat{\lambda}^\varepsilon_j|\big).\end{aligned}\tag{5.5}$$

A similar bound holds for $\check{\mathcal{D}}(\hat{\lambda}^\varepsilon;\lambda^\varepsilon)$ with $\|\cdot\|_{L_2(\Pi)}$ replaced by $\|\cdot\|_{L_2(\Pi_n)}$:

$$\begin{aligned}\check{\mathcal{D}}\big(\hat{\lambda}^\varepsilon;\lambda^\varepsilon\big)&\ge b\|f_{\hat{\lambda}^\varepsilon}-f_{\lambda^\varepsilon}\|^2_{L_2(\Pi_n)}+p\varepsilon\sum_{j=1}^N\big(|\hat{\lambda}^\varepsilon_j|^p+|\lambda^\varepsilon_j|^p\big)I\big(\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j\le 0\big)\\&\quad+\frac{p(p-1)\log 2}{4}\varepsilon\sum_{j=1}^N|\hat{\lambda}^\varepsilon_j|^pI\big(\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j>0,\,|\hat{\lambda}^\varepsilon_j|\ge 2|\lambda^\varepsilon_j|\big)\\&\quad+\frac{p(p-1)\log 2}{4}\varepsilon\sum_{j=1}^N|\lambda^\varepsilon_j|^pI\big(\hat{\lambda}^\varepsilon_j\lambda^\varepsilon_j>0,\,|\lambda^\varepsilon_j|\ge 2|\hat{\lambda}^\varepsilon_j|\big).\end{aligned}\tag{5.6}$$

Theorem 4. *Let $p\in(1,p_N]$. There exist constants $D>0$, $L>0$ and $C>0$, depending only on ℓ, such that, for all $J\subset\{1,\dots,N\}$ with $d:=d(J)$, for all $A\ge 1$ and for all*

$$\varepsilon\ge D\big(1+L\|\lambda^0\|_{\ell_1}\big)\bigg(\sqrt{\frac{A\log N}{n}}\vee\bigg(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\bigg)\bigg),$$

the assumption $\lambda_j^\varepsilon = 0$, $j \notin J$, *implies that with probability at least* $1 - N^{-A}$

$$\gamma_d\big(\hat{\lambda}^\varepsilon\big) \le \sum_{j\notin J} \big|\hat{\lambda}_j^\varepsilon\big| \le K\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\frac{d + A\log N}{n\varepsilon}$$

and

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le K\sqrt{\frac{d + A\log N}{n}},$$

where

$$K := K\big(C; \big\|\lambda^0\big\|_{\ell_1}\big) := \frac{C(1 + \|\lambda^0\|_{\ell_1})}{\tau(C\|\lambda^0\|_{\ell_1})}. \tag{5.7}$$

Remark. *The values of* ε *in Theorem 4 are supposed to be above the threshold that depends on* $U(J)$. *If one drops the term depending on* $U(J)$ *and assumes just that*

$$\varepsilon \ge D\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\bigg(\sqrt{\frac{A\log N}{n}} \vee \sqrt{\frac{d}{n}}\bigg),$$

then, for such ε*'s we get the bounds*

$$\mathbb{P}\bigg\{\gamma_d\big(\hat{\lambda}^\varepsilon\big) \ge K\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\sqrt{\frac{d + A\log N}{n}}\bigg\} \le N^{-A}$$

and

$$\mathbb{P}\bigg\{\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \ge K\sqrt{\frac{d + A\log N}{n}}\bigg\} \le N^{-A}.$$

However, unless $U(J)$ *is very large, namely,*

$$U(J) \ge \frac{\sqrt{nd}}{\log N},$$

the $U(J)$*-term provides the bounds for a broader range of values of* ε. *Since* $U(J) \le \sqrt{\frac{d}{\kappa}} + 1$, *where* $\kappa = \kappa(J) = \Gamma_2^2(J)$ (*see Section 1*), *this means that as soon as*

$$\kappa \ge \frac{(\log N)^2}{n},$$

there is an advantage in a more complicated form of the threshold.

Remark. *In the definition of constant* K, $K = K(C; \|\lambda^0\|_{\ell_1})$ (*see* (5.7)), *one can replace* $\|\lambda^0\|_{\ell_1}$ *by* $\|\lambda^{\varepsilon/c}\|_{\ell_1}$.

Remark. *If the first derivative of the loss function* $\ell'_u(y,u)$ *is uniformly bounded, then the constant* L *in the bounds of the theorem can be chosen as* $L = 0$.

Proof of Theorem 4. Assume that J is a subset of $\{1,\dots,N\}$ with $\#(J) = d$ and such that $\lambda_j^\varepsilon = 0$, $j \notin J$. Since

$$\gamma_d\big(\hat{\lambda}^\varepsilon\big) \le \sum_{j\notin J} \big|\hat{\lambda}_j^\varepsilon\big|,$$

it will be enough to bound $\sum_{j\notin J}|\hat\lambda^\varepsilon_j|$. Denote

$$A(\delta;\Delta):=A_J(\delta;\Delta):=\Big\{\lambda\in\mathbb{R}^N:\|\lambda\|_{\ell_1}\le C\big\|\lambda^{\varepsilon/c}\big\|_{\ell_1},\|f_\lambda-f_{\lambda^\varepsilon}\|_{L_2(\Pi)}\le\delta,\sum_{j\notin J}|\lambda_j|^p\le\Delta^p\Big\},$$

$$\alpha_n(\delta;\Delta):=\sup\Big\{\big|(P_n-P)\big((\ell'\bullet f_\lambda)(f_\lambda-f_{\lambda^\varepsilon})\big)\big|:\ \lambda\in A(\delta;\Delta)\Big\}$$

and

$$\check\alpha_n(\delta;\Delta):=\sup\Big\{\big|(P_n-P)\big((\ell'\bullet f_{\lambda^\varepsilon})(f_\lambda-f_{\lambda^\varepsilon})\big)\big|:\ \lambda\in A(\delta;\Delta)\Big\}$$

(α_n and $\check\alpha_n$ depend on J that at the moment is fixed). It follows from Theorem 1 that with probability at least $1-N^{-A}$

$$\big\|\hat\lambda^\varepsilon\big\|_{\ell_1}\le C\big\|\lambda^{\varepsilon/c}\big\|_{\ell_1}.$$

Together with (5.3)–(5.6) this implies that with probability at least $1-N^{-A}$

$$b\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)}+p\varepsilon\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\le\mathcal{D}\big(\hat\lambda^\varepsilon;\lambda^\varepsilon\big)\le\alpha_n\bigg(\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)};\bigg(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\bigg)^{1/p}\bigg)$$

and

$$b\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|^2_{L_2(\Pi_n)}+p\varepsilon\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\le\check{\mathcal{D}}\big(\hat\lambda^\varepsilon;\lambda^\varepsilon\big)\le\check\alpha_n\bigg(\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)};\bigg(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\bigg)^{1/p}\bigg)$$

(recall that here $b=2\tau(C\|\lambda^0\|)$.

The idea of the proof is to construct nonrandom upper bounds $\beta_n(\delta;\Delta)$ and $\check\beta_n(\delta;\Delta)$ on $\alpha_n(\delta;\Delta)$ and $\check\alpha_n(\delta;\Delta)$, respectively, that hold uniformly in all δ,Δ satisfying conditions (A.1) with a high probability. The bounds are given in Lemma 4 (see Appendix). If now we take

$$\delta=\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)},\qquad \Delta=\bigg(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\bigg)^{1/p}$$

and assume that they satisfy conditions (A.1), then, with the same probability,

$$b\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)}\le\beta_n\bigg(\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)};\bigg(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\bigg)^{1/p}\bigg)\tag{5.8}$$

and

$$p\varepsilon\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\le\check\beta_n\bigg(\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)};\bigg(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p\bigg)^{1/p}\bigg),\tag{5.9}$$

which will provide upper bounds on $\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)}$ and on $(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p)^{1/p}$ just by bounding the solutions of the inequalities

$$p\varepsilon\Delta^p\le\check\beta_n(\delta;\Delta),\qquad b\delta^2\le\beta_n(\delta;\Delta).\tag{5.10}$$

Note that the upper bounds on δ and Δ of conditions (A.1) can be safely assumed in view of Theorem 1. If $(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p)^{1/p}\le n^{-1/2}$ and $\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)}\le n^{-1/2}$, then the bounds of the theorem hold trivially. If $(\sum_{j\notin J}|\hat\lambda^\varepsilon_j|^p)^{1/p}>n^{-1/2}$, but $\|f_{\hat\lambda^\varepsilon}-f_{\lambda^\varepsilon}\|_{L_2(\Pi)}\le n^{-1/2}$, or another way around, then the quantity that is smaller

than $n^{-1/2}$ can be replaced in the right-hand sides of inequalities (5.8) and (5.9) by the upper bound $n^{-1/2}$, leading to the same outcome as below, but in a simpler way.

Thus, we can and do assume that δ and Δ satisfy conditions (A.1) and to complete the proof it is enough to bound the solutions of inequalities (5.10). Bounding the values of Δ satisfying the first of the inequalities (5.10),

$$p\varepsilon\Delta^p \le \check{\beta}_n(\delta;\Delta),$$

reduces to solving separately the following three inequalities

$$p\varepsilon\Delta^p \le C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\delta\sqrt{\frac{d+A\log N}{n}},$$

$$p\varepsilon\Delta^p \le C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}$$

and

$$p\varepsilon\Delta^p \le C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\Delta\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)$$

with respect to Δ (δ being fixed) and then taking the maximum of the solutions. The first inequality gives an upper bound on Δ

$$\biggl(C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\frac{\delta}{p\varepsilon}\sqrt{\frac{d+A\log N}{n}}\biggr)^{1/p},$$

which for our choice of $p\le p_N$ can be bounded from above by

$$C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\frac{\delta}{\varepsilon}\sqrt{\frac{d+A\log N}{n}}$$

(reminder: C is a constant that depends only on ℓ; its value may be different in different places!). Similarly, the second inequality yields the bound

$$\frac{C(1+L\|\lambda^0\|_{\ell_1})\|\lambda^0\|_{\ell_1}}{\varepsilon}\frac{A\log N}{n}.$$

The third inequality gives the bound

$$\Delta^{1/(q-1)}=\Delta^{p-1}\le\frac{C(1+L\|\lambda^0\|_{\ell_1})}{\varepsilon}\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr),$$

implying, under our condition on ε,

$$\begin{aligned}\Delta&\le\biggl(\frac{C(1+L\|\lambda^0\|_{\ell_1})}{\varepsilon}\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)\biggr)^{q-1}\\&\le\biggl(\frac{C}{D}\biggr)^{\log N-1}\le e^{\gamma_1}N^{-\gamma_1}\le e^{\gamma_1}n^{-1}\end{aligned}$$

(where $\gamma_1:=\log\frac{D}{C}\ge\frac{1}{\gamma}$ (γ is such that $N\ge n^\gamma$) is a sufficiently large positive constant for a proper choice of D). As a result, under the assumptions on n, N,

$$\Delta\le C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\frac{\delta}{\varepsilon}\sqrt{\frac{d+A\log N}{n}}\vee C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\bigl\|\lambda^0\bigr\|_{\ell_1}\frac{\|\lambda^0\|_{\ell_1}+1}{\varepsilon}\frac{A\log N}{n}=:\Delta(\delta).\tag{5.11}$$

It remains now to plug in this bound on Δ to the inequality

$$b\delta^2 \le \alpha_n(\delta;\Delta) \le \beta_n(\delta;\Delta),$$

which holds with probability at least $1-2N^{-A}$, to get

$$b\delta^2 \le C\bigl(1+\|\lambda^0\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d+A\log N}{n}} \vee \Delta(\delta)\biggl(\sqrt{\frac{A\log N}{n}} \vee \biggl(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\biggr)\biggr)\biggr]$$
$$\vee C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}.$$

Here again $b=2\tau(C\|\lambda^0\|_{\ell_1})$. This easily implies that with $K=K(C;\|\lambda^0\|_{\ell_1})$ defined in (5.7) and under the assumption on ε

$$\delta \le K\sqrt{\frac{d+A\log N}{n}},$$

which yields the second bound of the theorem. Finally, we can plug in this bound on δ back to the right hand side of the bound (5.11) on Δ, which implies

$$\Delta \le K\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\frac{d+A\log N}{n\varepsilon},$$

where again $K=K(C;\|\lambda^0\|_{\ell_1})$ with some constant C depending on ℓ. Inequality (5.9) now implies that with probability at least $1-N^{-A}$

$$\sum_{j\notin J}|\hat\lambda_j^\varepsilon| \le e\biggl(\sum_{j\notin J}|\hat\lambda_j^\varepsilon|^p\biggr)^{1/p} \le K\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\frac{d+A\log N}{n\varepsilon},$$

and the result readily follows. □

The first bound of Theorem 4 essentially means that the empirical solution $\hat\lambda^\varepsilon$ preserves "the sparsity pattern" of the true solution λ^ε. The next corollary shows that for a little bit larger values of ε the sum $\sum_{j\notin J}|\hat\lambda_j^\varepsilon|$ is much smaller than the bound of Theorem 4 suggests: in fact, it can be made as small as an arbitrary negative power of N.

Corollary 3. *Suppose the assumptions and notations of Theorem 4 hold, in particular, that $p\in(1,p_N]$. Let $q:=p/(p-1)$, let $B\ge 1$ and let J be a subset of $\{1,\dots,N\}$ such that $d(J)=d$. There exist constants $C,D\ge 1$ depending only on ℓ such that, for ε satisfying the condition*

$$\varepsilon \ge De^B K\sqrt{\frac{d+A\log N}{n}}$$

and for $K=K(C;\|\lambda^0\|_{\ell_1})$ given by (5.7), the assumption

$$\lambda_j^\varepsilon = 0, \quad j\notin J,$$

implies that with probability at least $1-N^{-A}$

$$\max_{j\notin J}|\hat\lambda_j^\varepsilon| \le e^{-B(q-1)} \le N^{-B}.$$

Moreover, for any $\beta>0$, the constant D can be chosen in such a way that for all

$$\varepsilon \ge De^B(q-1)K\sqrt{\frac{d+A\log N}{n}},$$

we have with probability at least $1 - N^{-A}$ *for all* $j = 1, \dots, N$

$$\bigl|\hat{\lambda}_j^{\varepsilon}\bigr| \le e^{\delta}\bigl(\bigl|\lambda_j^{\varepsilon}\bigr| \vee N^{-\beta}\bigr)$$

and

$$\bigl|\lambda_j^{\varepsilon}\bigr| \le e^{\delta}\bigl(\bigl|\hat{\lambda}_j^{\varepsilon}\bigr| \vee N^{-\beta}\bigr),$$

where $\delta = e^{-B}$.

Proof. By the necessary conditions of extrema, we have

$$P\bigl(\ell' \bullet f_{\lambda^{\varepsilon}}\bigr)h_j = -\varepsilon p\bigl|\lambda_j^{\varepsilon}\bigr|^{p-1}\operatorname{sign}\bigl(\lambda_j^{\varepsilon}\bigr)$$

and

$$P_n\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j = -\varepsilon p\bigl|\hat{\lambda}_j^{\varepsilon}\bigr|^{p-1}\operatorname{sign}\bigl(\hat{\lambda}_j^{\varepsilon}\bigr).$$

For $j \notin J$, this yields

$$\begin{aligned}\varepsilon p\bigl|\hat{\lambda}_j^{\varepsilon}\bigr|^{p-1} &= \bigl|P_n\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j\bigr| \\ &\le \bigl|(P_n - P)\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j\bigr| + \bigl|P\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j - P\bigl(\ell' \bullet f_{\lambda^{\varepsilon}}\bigr)h_j\bigr|.\end{aligned}$$

In view of Theorem 1, the first term in the right-hand side can be bounded with probability at least $1 - N^{-A}$ by

$$\sup\Bigl\{\bigl|(P_n - P)\bigl(\ell' \bullet f_{\lambda}\bigr)h_j\bigr| \colon \|\lambda\|_{\ell_1} \le C\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}\Bigr\}.$$

Using the second inequality of Lemma 2, one can show that again with probability at least $1 - N^{-A}$

$$\bigl|(P_n - P)\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j\bigr| \le C\bigl(1 + L\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}\bigr)\sqrt{\frac{A\log N}{n}}.$$

On the other hand, for the second term, using the Lipschitz condition on ℓ', we get

$$\bigl|P\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j - P\bigl(\ell' \bullet f_{\lambda^{\varepsilon}}\bigr)h_j\bigr| \le C\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)},$$

and we can use now the second bound of Theorem 4 (which also dominates the first term). This gives

$$\varepsilon p\bigl|\hat{\lambda}_j^{\varepsilon}\bigr|^{p-1} \le K\sqrt{\frac{d + A\log N}{n}}$$

with probability at least $1 - N^{-A}$ (and with $K = K(C; \|\lambda^0\|_{\ell_1})$ defined by (5.7), C being a constant depending only on ℓ). Moreover, by using the union bound and changing the value of C it is easy to show that the last inequality holds with probability at least $1 - N^{-A}$ for all $j \notin J$ simultaneously. Using the condition on ε, we get

$$\bigl|\hat{\lambda}_j^{\varepsilon}\bigr|^{p-1} \le \frac{1}{De^{B}}.$$

For $p \le p_N$, $q - 1 = (p-1)^{-1} \ge \log N - 1$ and for $D > 1$, this implies

$$\bigl|\hat{\lambda}_j^{\varepsilon}\bigr| \le \left(\frac{1}{De^{B}}\right)^{q-1} \le e^{-B(q-1)} \le N^{-B}.$$

To prove the second statement, the above derivation is modified as follows. First note that for all j

$$\begin{aligned}\varepsilon p|\hat{\lambda}_j^{\varepsilon}|^{p-1} &= \bigl|P_n\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j\bigr| \\ &\le \bigl|(P_n - P)\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j\bigr| + \bigl|P\bigl(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}\bigr)h_j - P\bigl(\ell' \bullet f_{\lambda^{\varepsilon}}\bigr)h_j\bigr| + \bigl|P\bigl(\ell' \bullet f_{\lambda^{\varepsilon}}\bigr)h_j\bigr| \\ &\le \varepsilon p|\lambda_j^{\varepsilon}|^{p-1} + K\sqrt{\frac{d + A\log N}{n}}.\end{aligned}$$

Therefore, we have with probability at least $1 - N^{-A}$ for all j

$$\left(\frac{|\hat{\lambda}_j^{\varepsilon}|}{|\lambda_j^{\varepsilon}| \vee N^{-\beta}}\right)^{p-1} \le 1 + \frac{K}{p\varepsilon N^{-\beta(p-1)}}\sqrt{\frac{d + A\log N}{n}}.$$

Since $1 \ge N^{-\beta(p-1)} \ge c_\beta$, where c_β is a constant depending only on β, it is easy to choose D such that, under the assumption on ε,

$$\left(\frac{|\hat{\lambda}_j^{\varepsilon}|}{|\lambda_j^{\varepsilon}| \vee N^{-\beta}}\right)^{p-1} \le 1 + \frac{e^{-B}}{q-1} = 1 + \frac{\delta}{q-1}.$$

This implies that

$$\frac{|\hat{\lambda}_j^{\varepsilon}|}{|\lambda_j^{\varepsilon}| \vee N^{-\beta}} \le \left(1 + \frac{\delta}{q-1}\right)^{q-1} \le e^{\delta}.$$

The proof of the remaining bound is similar. □

Now we provide general bounds on the quantity

$$\bigl|P(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) - P(\ell \bullet f_{\lambda^{\varepsilon}})\bigr| = \bigl|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})\bigr|,$$

which plays the role of random error in the penalized empirical risk minimization problem.

Denote $\bar{P}_J$ the orthogonal projector on the subspace generated by the functions h_j, $j \in J$, in the space $L_2(P)$.

Theorem 5. *Let $p \in (1, p_N]$. There exist constants $D > 0$, $L > 0$ and $C > 0$, depending only on ℓ, such that, for all $J \subset \{1, \dots, N\}$ with $d := d(J) \ge 1$, for arbitrary $A \ge 1$ and for all*

$$\varepsilon \ge D\bigl(1 + L\|\lambda^0\|_{\ell_1}\bigr)\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right),$$

the assumption

$$\lambda_j^{\varepsilon} = 0, \quad j \notin J,$$

implies that with probability at least $1 - N^{-A}$ the following bounds hold:

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})\bigr| \le K^2\frac{d + A\log N}{n} + K\bigl\|\bar{P}_J\bigl(\ell' \bullet f_{\lambda^{\varepsilon}}\bigr)\bigr\|_{L_2(P)}\sqrt{\frac{d + A\log N}{n}},$$

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})\bigr| \le K^2\frac{d + A\log N}{n} + K\|f_{\lambda^{\varepsilon}} - f_{\lambda^0}\|_{L_2(\Pi)}\sqrt{\frac{d + A\log N}{n}},$$

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})\bigr| \le K^2\frac{d + A\log N}{n} + K\sqrt{\frac{\mathcal{E}(f_{\lambda^{\varepsilon}})}{\tau(e\|\lambda^0\|_{\ell_1})}}\sqrt{\frac{d + A\log N}{n}}$$

and

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\bigr| \le K^2 \frac{d + A\log N}{n} + \frac{K}{\Gamma(J)}\varepsilon\sqrt{\frac{d + A\log N}{n}},$$

where $K = K(C; \|\lambda^0\|_{\ell_1})$ (*see* (5.7)).

Proof. We use the Taylor expansion and the boundedness of the second derivative of ℓ to get

$$\ell\bigl(y; f_{\hat{\lambda}^\varepsilon}(x)\bigr) - \ell\bigl(y; f_{\lambda^\varepsilon}(x)\bigr) = \ell'_u\bigl(y; f_{\lambda^\varepsilon}(x)\bigr)\bigl(f_{\hat{\lambda}^\varepsilon}(x) - f_{\lambda^\varepsilon}(x)\bigr) + R,$$

with the remainder R satisfying

$$|R| \le C\bigl(f_{\hat{\lambda}^\varepsilon}(x) - f_{\lambda^\varepsilon}(x)\bigr)^2$$

with C depending only on ℓ. Integrating the Taylor expansion with respect to P yields

$$\bigl|P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^\varepsilon}) - P\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr)(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon})\bigr| \le C\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)}.$$

By the necessary conditions of extremum in the optimization problem defining λ^ε,

$$\bigl|P\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr)h_j\bigr| = \varepsilon p\bigl|\lambda^\varepsilon_j\bigr|^{p-1},$$

we have

$$P\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr)h_j = 0, \quad j \notin J.$$

Hence, the function $\ell' \bullet f_{\lambda^\varepsilon}$ is orthogonal to h_j, $j \notin J$ in the space $L_2(P)$. Since also

$$f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon} \in \text{l.s.}\bigl(\{h_1, \dots, h_N\}\bigr),$$

we get

$$\begin{aligned}\bigl|P\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr)(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon})\bigr| &= \bigl|\bigl\langle\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr), f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\bigr\rangle_{L_2(P)}\bigr| = \bigl|\bigl\langle\bar{P}_J\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr), f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\bigr\rangle_{L_2(P)}\bigr| \\ &\le \bigl\|\bar{P}_J\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr)\bigr\|_{L_2(P)}\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}.\end{aligned}$$

This gives the following bound

$$\bigl|P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^\varepsilon})\bigr| \le C\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)} + \bigl\|\bar{P}_J\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr)\bigr\|_{L_2(P)}\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}$$

that holds with some constant C depending only on ℓ. The first bound of Theorem 5 now follows from the bound on $\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}$ of Theorem 4.

To prove the second inequality, note that the function $\ell' \bullet f_{\lambda^0}$ is orthogonal to h_j, $j = 1, \dots, N$, since, by the necessary conditions of minimum at λ^0

$$\bigl\langle\bigl(\ell' \bullet f_{\lambda^0}\bigr), h_j\bigr\rangle_{L_2(P)} = P\bigl(\ell' \bullet f_{\lambda^0}\bigr)h_j = 0, \quad j = 1, \dots, N.$$

This implies that

$$\begin{aligned}\bigl\|\bar{P}_J\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr)\bigr\|_{L_2(P)} &= \bigl\|\bar{P}_J\bigl(\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr) - \bigl(\ell' \bullet f_{\lambda^0}\bigr)\bigr)\bigr\|_{L_2(P)} \\ &\le \bigl\|\bigl(\ell' \bullet f_{\lambda^\varepsilon}\bigr) - \bigl(\ell' \bullet f_{\lambda^0}\bigr)\bigr\|_{L_2(P)} \le C\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|_{L_2(\Pi)}\end{aligned}$$

since ℓ' satisfies the Lipschitz condition with a constant depending only on ℓ.

The third inequality follows from the bound

$$\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|_{L_2(\Pi)} \le \sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau(e\|\lambda^0\|_{\ell_1})}},$$

which is obvious in view of the conditions on the loss.

To prove the last bound, recall again that $(\ell' \bullet f_{\lambda^\varepsilon})$ is orthogonal to h_j, $j \notin J$, and

$$\big|P\big(\ell' \bullet f_{\lambda^\varepsilon}\big)h_j\big| \le p\varepsilon\big\|\lambda^\varepsilon\big\|_{\ell_p}^{p-1}, \quad j \in J.$$

Since $\|\lambda^\varepsilon\|_{\ell_p} = \mathrm{O}(\varepsilon^{-1})$, under the conditions on ε and p we can bound $\|\lambda^\varepsilon\|_{\ell_p}^{p-1}$ by a constant. Hence, we have

$$\big|P\big(\ell' \bullet f_{\lambda^\varepsilon}\big)h_j\big| \le C\varepsilon, \quad j \in J.$$

Let

$$\sum_{j\in J} c_j h_j = P_J(f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}).$$

Then

$$\begin{aligned}
\big|P\big(\ell' \bullet f_{\lambda^\varepsilon}\big)(f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon})\big| &= \big|\big\langle\big(\ell' \bullet f_{\lambda^\varepsilon}\big), f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}\big\rangle_{L_2(P)}\big| = \big|\big\langle\big(\ell' \bullet f_{\lambda^\varepsilon}\big), P_J(f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon})\big\rangle_{L_2(P)}\big| \\
&= \left|\left\langle\big(\ell' \bullet f_{\lambda^\varepsilon}\big), \sum_{j\in J} c_j h_j\right\rangle_{L_2(P)}\right| \le \sum_{j=1}^N \big|P\big(\ell' \bullet f_{\lambda^\varepsilon}\big)h_j\big||c_j| \\
&\le C\varepsilon\sum_{j\in J}|c_j| \le C\varepsilon\frac{1}{\Gamma(J)}\big\|P_J(f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon})\big\|_{L_2(\Pi)} \le C\varepsilon\frac{1}{\Gamma(J)}\|f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)},
\end{aligned}$$

which easily leads to the result. □

Corollary 4. *Under the conditions and notations of Theorem 5, for*

$$\varepsilon = D\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\sqrt{\frac{A\log N}{n}},$$

the assumption

$$\lambda_j^\varepsilon = \lambda_j^0 = 0, \quad j \notin J,$$

implies that with probability at least $1 - N^{-A}$

$$\mathcal{E}(f_{\hat\lambda^\varepsilon}) \le K^{2p}\left[\frac{d + A\log N}{n} + \frac{1}{\Gamma^2(J)}\frac{A\log N}{n}\right],$$

which can be further bounded by the expression of the form

$$K^{2p}\frac{d}{\kappa}\frac{A\log N}{n}$$

with $\kappa = \kappa(J)$.

Proof. Assume $n \geq 4 \log N$ (otherwise, the bound holds trivially). First, consider the case when $\Gamma(J) \geq \frac{\sqrt{\log N}}{\sqrt{n}-\sqrt{\log N}}$ and, as a consequence,

$$U(J) \leq 1 + \frac{1}{\Gamma(J)} \leq \sqrt{\frac{n}{\log N}}.$$

Then ε satisfies the assumptions of Theorem 5, and we use the third bound of this theorem to get

$$\begin{aligned}\bigl|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\bigr| &\leq K^2 \frac{d + A\log N}{n} + K\sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau(e\|\lambda^0\|_{\ell_1})}}\sqrt{\frac{d + A\log N}{n}} \\ &\leq 3K^2 \frac{d + A\log N}{n} + 2\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau(e\|\lambda^0\|_{\ell_1})},\end{aligned}$$

which implies (since $\tau(e\|\lambda^0\|_{\ell_1}) \leq 1$)

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) \leq 3K^2 \frac{d + A\log N}{n} + 3\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau(e\|\lambda^0\|_{\ell_1})}.$$

Next we use Lemma 1 to bound $\mathcal{E}(f_{\lambda^\varepsilon})$. It gives

$$\begin{aligned}\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) &\leq 3K^2 \frac{d + A\log N}{n} + 3\frac{p^2\|\lambda^0\|_{\ell_p}^{2(p-1)}}{\tau^2(e\|\lambda^0\|_{\ell_1})}\frac{\varepsilon^2}{\Gamma^2(J)} \\ &= 3K^2 \frac{d + A\log N}{n} + 3\frac{D^2 p^2 \|\lambda^0\|_{\ell_p}^{2(p-1)}(1 + L\|\lambda^0\|_{\ell_1})^2}{\tau^2(e\|\lambda^0\|_{\ell_1})}\frac{1}{\Gamma^2(J)}\frac{A\log N}{n},\end{aligned}$$

which can be easily bounded by an expression of the form

$$K^{2p}\left[\frac{d + A\log N}{n} + \frac{1}{\Gamma^2(J)}\frac{A\log N}{n}\right]$$

with a proper adjustment of constant C in the definition of K.

On the other hand, if $\Gamma(J) \leq \frac{\sqrt{\log N}}{\sqrt{n}-\sqrt{\log N}}$, we can use the bound on

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\bigr|$$

of Theorem 1, which is of the order

$$C\bigl(1 + L\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}\bigr)\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}\varepsilon$$

and since $\|\lambda^{\varepsilon/c}\|_{\ell_1} \leq \frac{c_1}{\varepsilon}$, we have

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\bigr| \leq C\bigl(1 + L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)$$

with some constant C depending only on ℓ, which is clearly smaller than

$$\frac{K^{2p}}{\Gamma^2(J)}\frac{A\log N}{n},$$

so the inequality of the theorem holds in this case, too. □

Remark. *Although we are not attempting to derive in this paper the bounds on the ℓ_1-norm $\|\hat{\lambda}^{\varepsilon} - \lambda^0\|_{\ell_1}$ with optimal dependence on the parameters of the problem (such as d, κ, etc.), the following observations are straightforward. Suppose the conditions of Theorem 4 hold. First, we have*

$$\big\|\hat{\lambda}^{\varepsilon} - \lambda^{\varepsilon}\big\|_{\ell_1} = \sum_{j=1}^{N} |\hat{\lambda}^{\varepsilon}_j - \lambda^{\varepsilon}_j| \le \sum_{j\in J} |\hat{\lambda}^{\varepsilon}_j - \lambda^{\varepsilon}_j| + \sum_{j\notin J} |\hat{\lambda}^{\varepsilon}_j|.$$

On the other hand,

$$\begin{aligned}\sum_{j\in J} |\hat{\lambda}^{\varepsilon}_j - \lambda^{\varepsilon}_j| &\le \frac{1}{\Gamma(J)} \|f_{\hat{\lambda}^{\varepsilon},J} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)} \\ &\le \frac{1}{\Gamma(J)} \big(\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)} + \|f_{\hat{\lambda}^{\varepsilon},J} - f_{\hat{\lambda}^{\varepsilon}}\|_{L_2(\Pi)}\big) \\ &\le \frac{1}{\Gamma(J)} \Big(\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)} + \sum_{j\notin J} |\hat{\lambda}^{\varepsilon}_j|\Big). \end{aligned} \tag{5.12}$$

We can now use the bounds of Theorem 4 to show that, for

$$\varepsilon \ge D\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\bigg(\sqrt{\frac{A\log N}{n}} \vee \Big(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\Big)\bigg),$$

with probability at least $1 - N^{-A}$

$$\big\|\hat{\lambda}^{\varepsilon} - \lambda^{\varepsilon}\big\|_{\ell_1} \le \frac{K}{\Gamma(J)}\sqrt{\frac{d + A\log N}{n}} + K\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\Big(1 + \frac{1}{\Gamma(J)}\Big)\frac{d + A\log N}{n\varepsilon}.$$

This can be combined with Lemma 1 to get a bound on $\|\hat{\lambda}^{\varepsilon} - \lambda^0\|_{\ell_1}$ in terms of $\Gamma(J)$. Since $\frac{1}{\Gamma(J)} \le \sqrt{\frac{d}{\kappa}}$, one can then easily obtain for $\varepsilon = \sqrt{\frac{A\log N}{n}}$ the bound of the order

$$\Big(\frac{d}{\kappa} \vee \frac{(d)^{3/2}}{(\kappa)^{1/2}}\Big)\sqrt{\frac{\log N}{n}}.$$

As an alternative, one can use $\tilde{\Gamma}(J)$ instead of $\Gamma(J)$ leading to the following bound

$$\big\|\hat{\lambda}^{\varepsilon} - \lambda^{\varepsilon}\big\|_{\ell_1} \le \frac{K}{\tilde{\Gamma}(J)}\sqrt{\frac{d + A\log N}{n}} + K\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\frac{d + A\log N}{n\varepsilon}.$$

If κ denotes the smallest eigenvalue of the whole matrix $(\langle h_i, h_j\rangle_{L_2(\Pi)})_{i,j=1,N}$, this approach leads to the bound on $\|\hat{\lambda}^{\varepsilon} - \lambda^0\|_{\ell_1}$ of the order $\frac{d}{\kappa}\sqrt{\frac{\log N}{n}}$.

6. Approximate sparsity

Next we turn to bounding the sparsity function in the general case, when there is no knowledge about $\gamma_d(\lambda^{\varepsilon})$ being equal to 0 for some d. In this case, we need a more restrictive assumption on p, namely that $p - 1 \asymp (\log N - 1)^{-1}$. For simplicity, we just assume that $p = p_N$. We also have to assume a little more about the regularization parameter ε, namely, that

$$\varepsilon \ge D\big(1 + L\big\|\lambda^0\big\|_{\ell_1}\big)\log N\bigg(\sqrt{\frac{A\log N}{n}} \vee \Big(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\Big)\bigg).$$

The $\log N$-factor in the right-hand side is, in fact, just $q-1$, where

$$\frac{1}{p}+\frac{1}{q}=1.$$

This explains why we cannot obtain this result for the values of p closer to 1: the factor $q-1$ is becoming too large to keep the value of ε reasonably small (note that, for large values of ε, λ^{ε} does not provide a good approximation of λ^0).

Theorem 6. *Assume that $p=p_N$. There exist constants $L>0$, $D>0$ and $C>0$ depending only on ℓ (with $L=0$ in the case of loss function with bounded derivative) such that, for all $J\subset\{1,\dots,N\}$ with $d:=d(J)$, for all $A\geq 1$ and for all*

$$\varepsilon\geq D\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr),$$

the following bounds hold with $K=K(C;\|\lambda^0\|_{\ell_1})$ defined in (5.7) and with probability at least $1-N^{-A}$:

$$\gamma_d\bigl(\hat{\lambda}^{\varepsilon}\bigr)\leq\sum_{j\notin J}\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|\leq 9\sum_{j\notin J}\bigl|\lambda^{\varepsilon}_j\bigr|+K\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\frac{d+A\log N}{n\varepsilon}$$

and

$$\|f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}\leq\biggl[K\sqrt{\frac{d+A\log N}{n}}\vee K^{1/2}\biggl(\sum_{j\notin J}\bigl|\lambda^{\varepsilon}_j\bigr|\biggr)^{1/2}\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)^{1/2}\biggr].$$

If, in addition, the set J is such that

$$\gamma_d\bigl(\lambda^{\varepsilon}\bigr)=\sum_{j\notin J}\bigl|\lambda^{\varepsilon}_j\bigr|,$$

then with probability at least $1-N^{-A}$

$$\gamma_d\bigl(\lambda^{\varepsilon}\bigr)\leq 9\gamma_d\bigl(\hat{\lambda}^{\varepsilon}\bigr)+K\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\frac{d+A\log N}{n\varepsilon}.$$

Proof. Let J denote a subset of $\{1,\dots,N\}$ such that $\#(J)=d$ and the condition on ε holds. Recall the definitions of $\Lambda(\delta;\Delta)=\Lambda_J(\delta;\Delta)$ and of $\alpha_n(\delta;\Delta)$, $\check{\alpha}_n(\delta;\Delta)$.

We will first show that

$$\mathbb{P}\biggl\{\sum_{j\notin J}\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|\geq 9\sum_{j\notin J}\bigl|\lambda^{\varepsilon}_j\bigr|+K\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\frac{d+A\log N}{n\varepsilon}\biggr\}\leq N^{-A}.$$

Obviously,

$$\begin{aligned}\sum_{j\notin J}\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|^p&=\sum_{j\notin J}\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|^p I\bigl(\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|\leq 2\bigl|\lambda^{\varepsilon}_j\bigr|\bigr)+\sum_{j\notin J}\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|^p I\bigl(\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|>2\bigl|\lambda^{\varepsilon}_j\bigr|\bigr)\\&\leq(2\Lambda)^p+\sum_{j\notin J}\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|^p I\bigl(\bigl|\hat{\lambda}^{\varepsilon}_j\bigr|>2\bigl|\lambda^{\varepsilon}_j\bigr|\bigr),\end{aligned}$$

where

$$\Lambda:=\sum_{j\notin J}\bigl|\lambda^{\varepsilon}_j\bigr|.$$

Using bounds (5.5), (5.6), it is now easy to modify inequality (5.9) as follows:

$$\frac{p\log 2}{4}(p-1)\varepsilon\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j|^p \le \frac{p\log 2}{4}(p-1)\varepsilon(2\Lambda)^p + \check{\beta}_n\left(\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}; \left(\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j|^p\right)^{1/p}\right). \tag{6.1}$$

Here and in what follows $\check{\beta}_n$ and β_n are the bounds of Lemma 5. As in the proof of Theorem 4, (5.8) and (6.1) provide upper bounds on $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$ and on $(\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j|^p)^{1/p}$ just by bounding the solutions of the inequalities

$$\frac{p\log 2}{4}(p-1)\varepsilon\Delta^p \le \frac{p\log 2}{4}(p-1)\varepsilon(2\Lambda)^p + \check{\beta}_n(\delta;\Delta), \quad b\delta^2 \le \beta_n(\delta;\Delta). \tag{6.2}$$

If

$$\left(\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j|^p\right)^{1/p} \le 3\Lambda,$$

then

$$\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j| \le e\left(\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j|^p\right)^{1/p} \le 9\Lambda,$$

and the result follows. So, it is enough to consider the case

$$\left(\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j|^p\right)^{1/p} > 3\Lambda,$$

and in this case the upper bound on Δ's solving (6.2) must be larger than 3Λ. Denote

$$\bar{\Delta} := \Delta - 2\Lambda, \qquad \bar{\varepsilon} := \frac{\log 2}{4}(p-1)\varepsilon.$$

Then (6.2) implies

$$p\bar{\varepsilon}\bar{\Delta}^p + p\bar{\varepsilon}(2\Lambda)^p \le p\bar{\varepsilon}(2\Lambda)^p + \check{\beta}_n(\delta;\bar{\Delta}+2\Lambda) \tag{6.3}$$

and

$$b\delta^2 \le \beta_n(\delta;\bar{\Delta}+2\Lambda). \tag{6.4}$$

Recall that we are interested only in solutions $\bar{\Delta} \ge \Lambda$. For such values of $\bar{\Delta}$, it is easy to see that

$$\begin{aligned}
&\beta_n(\delta,\bar{\Delta}+2\Lambda)\\
&\quad\le C\big(1+\|\lambda^0\|_{\ell_1}\big)\left[\delta\sqrt{\frac{d+A\log N}{n}} \vee \bar{\Delta}\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right)\right]\\
&\qquad\vee C\big(1+L\|\lambda^0\|_{\ell_1}\big)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}
\end{aligned}$$

and

$$\begin{aligned}
&\check{\beta}_n(\delta,\bar{\Delta}+2\Lambda)\\
&\quad\le C\big(1+L\|\lambda^{\varepsilon}\|_{\ell_1}\big)\left[\delta\sqrt{\frac{d+A\log N}{n}} \vee \bar{\Delta}\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right)\right]\\
&\qquad\vee C\big(1+L\|\lambda^0\|_{\ell_1}\big)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}
\end{aligned}$$

with some value of constant C. We can use these upper bounds in inequalities (6.3) and (6.4) and (with a little abuse of notations) we denote them $\beta_n(\delta, \bar{\Delta})$ and $\check{\beta}_n(\delta, \bar{\Delta})$. Since they are given by exactly the same expressions as β_n and $\check{\beta}_n$ in the proof of Theorem 4 and since

$$\bar{\varepsilon} \ge D_1\big(1 + L\|\lambda^0\|_{\ell_1}\big)\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right)$$

(to see this recall that $p-1$ is of the order $\frac{1}{\log N}$ and also recall the assumption on ε in the theorem) we can use the argument of the last part of the proof of Theorem 4 to show that $\bar{\Delta}$'s solving inequalities (6.3) and (6.4) are bounded from above by

$$K\big(1 + L\|\lambda^0\|_{\ell_1}\big)\frac{d + A\log N}{n\bar{\varepsilon}},$$

where $K = K(C; \|\lambda^0\|_{\ell_1})$ with some C. Hence,

$$\Delta \le 3\Lambda + K\big(1 + L\|\lambda^0\|_{\ell_1}\big)\log N\frac{d + A\log N}{n\varepsilon},$$

implying that

$$\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j| \le e\left(\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j|^p\right)^{1/p} \le 9\Lambda + K\big(1 + L\|\lambda^0\|_{\ell_1}\big)\log N\frac{d + A\log N}{n\varepsilon}$$

with probability at least $1 - N^{-A}$ and with $K = K(C; \|\lambda^0\|_{\ell_1})$ for some C depending on ℓ, which implies the first bound.

To prove the remaining bounds, consider the set $\hat{J} \subset \{1, \dots, N\}$ such that $\#(\hat{J}) = d$ and

$$\gamma_d\big(\hat{\lambda}^\varepsilon\big) := \sum_{j\notin\hat{J}}|\hat{\lambda}^\varepsilon_j|.$$

Similarly to (6.1), we have the following bound

$$\frac{p\log 2}{4}(p-1)\varepsilon\sum_{j\notin\hat{J}}|\lambda^\varepsilon_j|^p \le \frac{p\log 2}{4}(p-1)\varepsilon\big(2\gamma_d\big(\hat{\lambda}^\varepsilon\big)\big)^p + \check{\beta}_n\left(\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}; \left(\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j|^p\right)^{1/p}\right). \tag{6.5}$$

As it was proved earlier, with probability at least $1 - N^{-A}$

$$\left(\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j|^p\right)^{1/p} \le 3\Lambda + K\big(1 + L\|\lambda^0\|_{\ell_1}\big)\log N\frac{d + A\log N}{n\varepsilon}$$

with $K = K(C; \|\lambda^0\|_{\ell_1})$. This can be plugged in the expression for $\check{\beta}_n$ in the right-hand side of bound (6.5) yielding with some C and under the assumption on ε

$$\begin{aligned}
&\check{\beta}_n\left(\delta; \left(\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j|^p\right)^{1/p}\right)\\
&\quad\le C\big(1 + L\|\lambda^\varepsilon\|_{\ell_1}\big)\left[\delta\sqrt{\frac{d + A\log N}{n}} \vee \Lambda\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right) \vee K\frac{d + A\log N}{n}\right]\\
&\qquad\vee C\big(1 + L\|\lambda^0\|_{\ell_1}\big)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n} =: \check{\beta}^1_n(\delta).
\end{aligned}$$

Similarly, we have

$$\beta_n\biggl(\delta;\biggl(\sum_{j\notin J}|\hat{\lambda}^{\varepsilon}_j|^p\biggr)^{1/p}\biggr)$$
$$\le C\bigl(1+\|\lambda^0\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d+A\log N}{n}}\vee\Lambda\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)\vee K\frac{d+A\log N}{n}\biggr]$$
$$\vee C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}=:\beta^1_n(\delta).$$

As before, we also must have (with $b=2\tau(C\|\lambda^0\|_{\ell_1})$)

$$b\|f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{\varepsilon}}\|^2_{L_2(\Pi)}\le\beta^1_n\bigl(\|f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}\bigr),$$

which, by solving the inequality, yields the bound

$$\|f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}\le\biggl[K\sqrt{\frac{d+A\log N}{n}}\vee K^{1/2}\Lambda^{1/2}\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)^{1/2}\biggr]\tag{6.6}$$

with $K=K(C;\|\lambda^0\|_{\ell_1})$, which is the second bound of the theorem. Finally, under the additional assumption that

$$\Lambda=\sum_{j\notin J}|\lambda^{\varepsilon}_j|=\gamma_d\bigl(\lambda^{\varepsilon}\bigr),$$

we plug this bound into $\check{\beta}^1_n(\|f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)})$, which yields

$$\check{\beta}^1_n\bigl(\|f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}\bigr)$$
$$\le K\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\frac{d+A\log N}{n}$$
$$\vee C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\gamma_d\bigl(\lambda^{\varepsilon}\bigr)\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr):=\tilde{\beta}.$$

To see this, note that the expression defining $\check{\beta}^1_n$ is the maximum of four terms. The third and the fourth terms are dominated by

$$K\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\frac{d+A\log N}{n}$$

with a proper adjustment of the constants. Then we substitute $\delta=\|f_{\hat{\lambda}^{\varepsilon}}-f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$ in the first term and use bound (6.6) and the inequality $ab\le(a^2+b^2)/2$. This gives

$$\delta\sqrt{\frac{d+A\log N}{n}}\le K\frac{d+A\log N}{n}+\Lambda\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)$$

(again, with an adjustment of constants). Finally, the second term in the definition of $\check{\beta}^1_n$ is of the same form as the second term in the expression for $\tilde{\beta}$.

It is easy to see that with some numerical constant C_1

$$\biggl(\frac{\tilde{\beta}}{(p-1)\varepsilon}\biggr)^{1/p}\le C_1\frac{\tilde{\beta}}{(p-1)\varepsilon}.$$

Therefore, we can obtain from (6.5) the following bound (changing one more time the value of constant C):

$$\begin{aligned}\gamma_d\bigl(\lambda^\varepsilon\bigr) &\le \sum_{j\notin\hat J}\bigl|\lambda_j^\varepsilon\bigr| \le e\biggl(\sum_{j\notin\hat J}\bigl|\lambda_j^\varepsilon\bigr|^p\biggr)^{1/p}\\ &\le 2e\gamma_d\bigl(\hat\lambda^\varepsilon\bigr) + \frac{K(1+L\|\lambda^0\|_{\ell_1})}{(p-1)\varepsilon}\frac{d+A\log N}{n}\\ &\quad + \frac{C(1+L\|\lambda^0\|_{\ell_1})}{(p-1)\varepsilon}\gamma_d\bigl(\lambda^\varepsilon\bigr)\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr),\end{aligned}$$

which under the assumption on ε with a proper choice of constants D and C yields

$$\gamma_d\bigl(\lambda^\varepsilon\bigr) \le 3e\gamma_d\bigl(\hat\lambda^\varepsilon\bigr) + K\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\frac{d+A\log N}{n\varepsilon},$$

and which holds with probability at least $1-N^{-A}$, implying the second bound of the theorem. □

The next statement shows that for larger values of ε the "pattern" of $\hat\lambda^\varepsilon$ is even closer to the "pattern" of λ^ε. The proof is similar to Corollary 3.

Corollary 5. *Suppose the assumptions and notations of Theorem 6 hold, in particular, that $p = p_N$. Let $q := p/(p-1) = \log N$, let $B \ge 1$ and let $\beta > 0$. There exists a constant D depending on ℓ and β such that, for all ε satisfying the condition (for some d)*

$$\varepsilon \ge De^B(q-1)\biggl[K\sqrt{\frac{d+A\log N}{n}}\vee K^{1/2}\bigl(\gamma_d\bigl(\lambda^\varepsilon\bigr)\bigr)^{1/2}\biggl(\frac{d+\log N}{n}\biggr)^{1/4}\biggr],$$

we have that with probability at least $1-N^{-A}$ for all $j=1,\dots,N$

$$\bigl|\hat\lambda_j^\varepsilon\bigr| \le e^\delta\Bigl(\bigl|\lambda_j^\varepsilon\bigr|\vee N^{-\beta}\Bigr)$$

and

$$\bigl|\lambda_j^\varepsilon\bigr| \le e^\delta\bigl(\bigl|\hat\lambda_j^\varepsilon\bigr|\vee N^{-\beta}\bigr),$$

where $\delta = e^{-B}$.

Remark. *It is known that ℓ_1-penalized least square regression estimators (LASSO) can be used for variable selection (see, e.g., [22,30] and references therein). It is not our goal here to provide a comprehensive analysis of this problem. However, the following observations are obvious. Suppose that for some $J\subset\{1,\dots,N\}$ with $\#(J)=d$ and for some $\Delta>0$*

$$\bigl|\lambda_j^0\bigr| \ge \Delta, \quad j\in J \quad \text{and} \quad \lambda_j^0 = 0, \quad j\notin J.$$

Suppose also that

$$\bigl\|\lambda^\varepsilon-\lambda^0\bigr\|_{\ell_1} = \mathrm{O}(\varepsilon), \quad \varepsilon\to 0$$

(see Lemma 1 for some conditions under which it is true). Finally, suppose that $N=N_n\ge n^\gamma$ and take the regularization parameter $\varepsilon=\varepsilon_n\to 0$ as $n\to\infty$ such that

$$\frac{(\log N_n)^{5/2}}{n^{1/2}} = \mathrm{o}(\varepsilon_n) \quad \text{as } n\to\infty.$$

Then

$$\gamma_d\bigl(\lambda^{\varepsilon_n}\bigr) \le \bigl\|\lambda^{\varepsilon_n} - \lambda^0\bigr\|_{\ell_1} = \mathrm{O}(\varepsilon_n)$$

and the condition on ε of Corollary 5 is satisfied for large enough n. Therefore, for large enough n with probability at least $1 - N_n^{-A}$

$$\max_{j\notin J}\bigl|\hat{\lambda}_j^{\varepsilon_n}\bigr| \le e^{\delta} N_n^{-\beta} = \mathrm{o}(\varepsilon_n)$$

and

$$\min_{j\in J}\bigl|\hat{\lambda}_j^{\varepsilon_n}\bigr| \ge e^{-\delta}\Delta \ge \varepsilon_n$$

(recall that $N = N_n \ge n^{\gamma}$ and assume that $\beta > 1/(2\gamma)$). Define

$$\hat{J} := \bigl\{j\colon\ \bigl|\hat{\lambda}_j^{\varepsilon_n}\bigr| \ge \varepsilon_n\bigr\}.$$

Then for large enough n

$$\mathbb{P}\{\hat{J} \ne J\} \le N_n^{-A} \to 0 \quad \text{as } n\to\infty,$$

showing consistency of variable selection. It is easy to extend this argument to the case of small $\Delta = \Delta_n \to 0$ and large $d = d_n \to \infty$ as $n\to\infty$.

In the case of "approximate sparsity," the following version of Theorem 5 holds (the proof is essentially the same as before, with Theorem 6 used instead of Theorem 4). Denote by $\bar{P}_{1,N}$ the orthogonal projector on the linear span of $\{h_1,\dots,h_N\}$ in the space $L_2(P)$ and

$$g^{\varepsilon} := \bar{P}_{1,N}\bigl(\ell' \bullet f_{\lambda^{\varepsilon}}\bigr).$$

Theorem 7. *Let $p = p_N$. There exist constants $D > 0, L > 0$ and $C > 0$, depending only on ℓ, such that, for all $J \subset \{1,\dots,N\}$ with $d := d(J) \ge 1$ for arbitrary $A \ge 1$ and for all*

$$\varepsilon \ge D\bigl(1 + L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right),$$

the following bound holds with probability at least $1 - N^{-A}$

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})\bigr| \le \Omega^2 + \bigl\|g^{\varepsilon}\bigr\|_{L_2(P)}\Omega,$$

where

$$\Omega := K\sqrt{\frac{d + A\log N}{n}} \vee K^{1/2}\left(\sum_{j\notin J}\bigl|\lambda_j^{\varepsilon}\bigr|\right)^{1/2}\left(\sqrt{\frac{A\log N}{n}} \vee \left(U(J)\frac{\log N}{n} \wedge \sqrt{\frac{d}{n}}\right)\right)^{1/2}$$

and $K = K(C; \|\lambda^0\|_{\ell_1})$ (as defined in (5.7)). This also implies that with a properly chosen constant C

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})\bigr| \le \Omega^2 + \|f_{\lambda^{\varepsilon}} - f_{\lambda^0}\|_{L_2(\Pi)}\Omega$$

and

$$\bigl|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})\bigr| \le \Omega^2 + \sqrt{\frac{\mathcal{E}(f_{\lambda^{\varepsilon}})}{\tau(e\|\lambda^0\|_{\ell_1})}}\Omega.$$

Remarks.

1. *If* $U(J) \le \sqrt{\frac{n}{\log N}}$, *then the bounds of Theorems* 5 *and* 7 *hold for all* $\varepsilon \ge D(1+L\|\lambda^0\|_{\ell_1})\sqrt{\frac{A\log N}{n}}$ *and* $\varepsilon \ge D(1+L\|\lambda^0\|_{\ell_1})\log N\sqrt{\frac{A\log N}{n}}$, *respectively. Note also that, by Theorem* 1, *for all* $\varepsilon > 0$,

$$\bigl|P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^\varepsilon})\bigr| \le C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\bigl\|\lambda^0\bigr\|_{\ell_1}\varepsilon,$$

which can be used when

$$U(J) > \sqrt{\frac{n}{\log N}},$$

as we did in the proof of Corollary 4.

2. *The bound of Theorem* 5 *clearly holds for all*

$$\varepsilon \ge D\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\sqrt{\frac{d+A\log N}{n}}$$

(*and all* $p \in (1, p_N]$) *implying that under the assumption* $\gamma_d(\lambda^\varepsilon) = 0$ *with probability at least* $1 - N^{-A}$

$$P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^0}) \le C\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|^2_{L_2(\Pi)} + K^2\frac{d+A\log N}{n}.$$

Similarly, for $p = p_N$ and

$$\varepsilon \ge D\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\sqrt{\frac{d+A\log N}{n}},$$

we have with the same probability (and with no assumption on $\gamma_d(\lambda^\varepsilon)$)

$$P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^0}) \le C\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|^2_{L_2(\Pi)} + \left(K^2\frac{d+A\log N}{n} + K\gamma_d\bigl(\lambda^\varepsilon\bigr)\sqrt{\frac{d+\log N}{n}}\right).$$

The derivation of these inequalities requires only the fact that

$$\mathcal{E}(f_{\lambda^\varepsilon}) \le C\|f_{\lambda^\varepsilon} - f_{\lambda^0}\|^2_{L_2(\Pi)},$$

which is true under the assumptions on the loss (with C depending only on ℓ).

For $\lambda \in \mathbb{R}^N$, recall the notation $J_\lambda = \operatorname{supp}(\lambda)$.

We now show how to obtain oracle inequalities of the same type as in Theorem 3 (with worse constants) as a corollary of Theorem 7 and approximation error bound of Theorem 2.

Corollary 6. *Under the assumptions and notations of Theorem* 7, *the following bound holds for all*

$$\varepsilon \in \left[D\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\log N\sqrt{\frac{A\log N}{n}}, 1\right]$$

with probability at least $1 - N^{-A}$ *and with some* $C > 0$:

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) \le K^2 \inf_{\lambda\in\mathbb{R}^N}\left[\mathcal{E}(f_\lambda) + \frac{(\|\lambda\|_{\ell_p} \vee 1)^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1} \vee e\|\lambda^0\|_{\ell_1})}\frac{1}{\tilde{\Gamma}^2(J_\lambda)}\varepsilon^2\right],$$

where $K = K(C; \|\lambda^0\|_{\ell_1})$ (*see* (5.7)).

Proof. It follows from Theorem 2 that

$$\mathcal{E}(f_{\lambda^\varepsilon}) \le \inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{2p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})}\frac{1}{\tilde\Gamma^2(J_\lambda)}\varepsilon^2\right] \tag{6.7}$$

and for $\bar\lambda$ defined in Theorem 2

$$\varepsilon\sum_{j\notin J_{\bar\lambda}}|\lambda_j^\varepsilon| \le c\inf_{\lambda\in\mathbb{R}^N}\left[3\mathcal{E}(f_\lambda) + \frac{2p^2\|\lambda\|_{\ell_p}^{2(p-1)}}{\tau(\|\lambda\|_{\ell_1}\vee e\|\lambda^0\|_{\ell_1})}\frac{1}{\tilde\Gamma^2(J_\lambda)}\varepsilon^2\right]. \tag{6.8}$$

We will use the set $J := J_{\bar\lambda}$ in the bounds of Theorem 7. If

$$\frac{1}{\tilde\Gamma(J)} \le \sqrt{\frac{n}{\log N}} - 1,$$

then (see Proposition 1)

$$U(J) \le \frac{1}{\Gamma(J)} + 1 \le \sqrt{\frac{n}{\log N}}.$$

Hence, the assumption

$$\varepsilon \ge D\big(1 + L\|\lambda^0\|_{\ell_1}\big)\log N\sqrt{\frac{A\log N}{n}}$$

immediately implies the assumption on ε of Theorem 7. In this case, Theorem 7 implies that with probability at least $1 - N^{-A}$

$$\begin{aligned}\big|\mathcal{E}(f_{\hat\lambda^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\big| &\le K^2\frac{d(\bar\lambda) + A\log N}{n} + K\varepsilon\sum_{j\notin J}|\lambda_j^\varepsilon| \\ &\quad + \sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau(e\|\lambda^0\|_{\ell_1})}}\left(K\sqrt{\frac{d(\bar\lambda) + A\log N}{n}} \vee K^{1/2}\varepsilon^{1/2}\left(\sum_{j\notin J}|\lambda_j^\varepsilon|\right)^{1/2}\right).\end{aligned}$$

It remains to combine this bound with (6.7) and to use (6.8) to bound the expression

$$\varepsilon\sum_{j\notin J}|\lambda_j^\varepsilon|$$

in order to complete the proof in the case

$$\frac{1}{\tilde\Gamma(J)} \le \sqrt{\frac{n}{\log N}} - 1.$$

It has to be also taken into account that

$$\frac{1}{\tilde\Gamma^2(J_{\bar\lambda})}\varepsilon^2 \ge \varepsilon^2 \ge \frac{A\log N}{n}$$

(since $\frac{1}{\Gamma(J_{\bar\lambda})} \ge 1$) and (see Proposition 1, (i) and (iii))

$$\frac{1}{\tilde\Gamma^2(J_{\bar\lambda})}\varepsilon^2 \ge \frac{d(\bar\lambda)}{n}.$$

If $\frac{1}{\tilde{\Gamma}(J_\lambda)} > \sqrt{\frac{n}{\log N}} - 1$ the excess risk bound of Theorem 1 is even tighter than the bound of the current theorem, so the result also holds. □

Appendix. Bounds on Rademacher processes and other auxiliary results

Proof of Proposition 1. (i) It follows from the definition of $\Gamma(J)$ that, for α_j equal to $+1$ or -1,

$$\Gamma(J)d(J) = \Gamma(J)\sum_{j\in J}|\alpha_j| \le \left\|\sum_{j\in J}\alpha_j h_j\right\|_{L_2(\Pi)}.$$

Hence, for i.i.d. Rademacher r.v. ε_j,

$$\begin{aligned}\Gamma(J)d(J) &\le \min_{|\alpha_j|=1}\left\|\sum_{j\in J}\alpha_j h_j\right\|_{L_2(\Pi)} \le \mathbb{E}\left\|\sum_{j\in J}\varepsilon_j h_j\right\|_{L_2(\Pi)} \le \mathbb{E}^{1/2}\left\|\sum_{j\in J}\varepsilon_j h_j\right\|^2_{L_2(\Pi)}\\ &\le \left(\sum_{j\in J}\|h_j\|^2_{L_2(\Pi)}\right)^{1/2} \le \sqrt{d(J)}\end{aligned}$$

(since h_j are bounded by 1).

(ii) It easily follows from Cauchy–Schwarz inequality.

(iii) The following inequality is obvious

$$\left\|\sum_{j\in J}\lambda_j h_j\right\|_{L_2(\Pi)} \le \bigl(1-\rho^2(J)\bigr)^{-1/2}\left\|\sum_{j=1}^N \lambda_j h_j\right\|_{L_2(\Pi)},$$

since for $f = \sum_{j\in J}\lambda_j h_j$ and $g = \sum_{j\notin J}\lambda_j h_j$, we have

$$\|f+g\|^2_{L_2(\Pi)} = \bigl(1-\cos^2(\alpha)\bigr)\|f\|^2_{L_2(\Pi)} + \bigl(\|f\|_{L_2(\Pi)}\cos(\alpha) + \|g\|_{L_2(\Pi)}\bigr)^2 \ge \bigl(1-\rho^2(J)\bigr)\|f\|^2_{L_2(\Pi)},$$

where α is the angle between f and g. This easily leads to the bound

$$\sum_{j\in J}|\alpha_j| \le \frac{1}{\Gamma(J)\sqrt{1-\rho^2(J)}}\left\|\sum_{j=1}^N \alpha_j h_j\right\|_{L_2(\Pi)},$$

which implies that

$$\tilde{\Gamma}(J) \ge \Gamma(J)\sqrt{1-\rho^2(J)}.$$

The inequality

$$1 \ge \Gamma(J) \ge \tilde{\Gamma}(J)$$

follows immediately from the definitions.

(iv) This is a straightforward application of the Cauchy–Schwarz inequality.

(v) and (vi) If I_J denotes the embedding operator from $(L_J; \|\cdot\|_{L_2(\Pi)})$ into $(L_J; \|\cdot\|_\infty)$, then it is easy to see that $U(J)$ is not greater than $\|I_J\| + 1$. If h_j, $j = 1,\dots,N$, are orthogonal, then $U(J) \le 1$. More generally, we have (since $\|h_j\|_\infty \le 1$)

$$\left\|\sum_{j\in J}\alpha_j h_j\right\|_\infty \le \sum_{j\in J}|\alpha_j| \le \frac{1}{\Gamma(J)}\left\|\sum_{j\in J}\alpha_j h_j\right\|_{L_2(\Pi)},$$

implying that

$$U(J) \le \|I_J\| + 1 \le \frac{1}{\Gamma(J)} + 1 \le \frac{\sqrt{d}}{\Gamma_2(J)} + 1 = \sqrt{\frac{d}{\kappa(J)}} + 1. \qquad \square$$

Below, we give several technical lemmas concerning Rademacher processes that are used in the proofs of the main results of the paper. They are relatively standard and have similar proofs. We give only the proofs of Lemma 2 and Lemma 4. We will frequently use the symmetrization inequality, contraction inequality and other standard facts about Rademacher processes that can be found in [18,27] and, in the context close to this paper, in [16].

Lemma 2. *Let $p \in [1, p_N]$. There exist constants C, L depending only on ℓ such that for all $A \ge 1$ and for all $R > 0$ with probability at least $1 - N^{-A}$ the following bound holds*

$$\sup_{\|\lambda\|_{\ell_p} \le R} \bigl|(P_n - P)(\ell \bullet f_\lambda - \ell \bullet f_0)\bigr| \le C(1 + LR)R\sqrt{\frac{A \log N}{n}}.$$

In addition, with the same probability

$$\max_{1 \le k \le N} \sup_{\|\lambda\|_{\ell_p} \le R} \bigl|(P_n - P)\bigl(\ell' \bullet f_\lambda\bigr)h_k\bigr| \le C(1 + LR)\sqrt{\frac{A \log N}{n}}.$$

Here f_0 is f_λ for $\lambda = 0$, so, f_0 is identically equal to 0.

Proof. First, use the bounded difference inequality to show that with probability $\ge 1 - e^{-t}$

$$\sup_{\|\lambda\|_{\ell_p} \le R} \bigl|(P_n - P)(\ell \bullet f_\lambda - \ell \bullet f_0)\bigr| \le \mathbb{E} \sup_{\|\lambda\|_{\ell_p} \le R} \bigl|(P_n - P)(\ell \bullet f_\lambda - \ell \bullet f_0)\bigr| + \frac{C(1 + LR)R\sqrt{t}}{\sqrt{n}}$$

(to apply the bounded difference inequality, note that, in view of the assumptions on ℓ, for $\|\lambda\|_{\ell_1} \le e\|\lambda\|_{\ell_p} \le eR$,

$$\bigl|(\ell \bullet f_\lambda)(x, y) - (\ell \bullet f_0)(x, y)\bigr| \le C(1 + LR)R$$

with constants $C, L > 0$ depending only on ℓ; the bounded difference inequality is then applied to the supremum of the empirical process on the class $\mathcal{G} := \{\frac{\ell \bullet f_\lambda}{C(1+LR)R} \colon \|\lambda\|_{\ell_p} \le R\}$). Next we have to bound

$$\mathbb{E} \sup_{\|\lambda\|_{\ell_p} \le R} \bigl|(P_n - P)(\ell \bullet f_\lambda - \ell \bullet f_0)\bigr|.$$

Denoting

$$R_n(g) := \sum_{j=1}^{n} \varepsilon_j g(X_j)$$

the Rademacher process (ε_j being i.i.d. Rademacher random variables) and using the symmetrization inequality, we get

$$\mathbb{E} \sup_{\|\lambda\|_{\ell_p} \le R} \bigl|(P_n - P)(\ell \bullet f_\lambda - \ell \bullet f_0)\bigr| \le 2\mathbb{E} \sup_{\|\lambda\|_{\ell_p} \le R} \bigl|R_n(\ell \bullet f_\lambda - \ell \bullet f_0)\bigr|,$$

which using the contraction inequality for Rademacher processes can be bounded further by

$$C(1 + LR)\mathbb{E} \sup_{\|\lambda\|_{\ell_1} \le eR} \bigl|R_n(f_\lambda)\bigr|.$$

Well-known bounds for the expectation of the maximum of Rademacher processes over a finite set of functions now yield

$$\mathbb{E}\sup_{\|\lambda\|_{\ell_1}\le eR}\big|R_n(f_\lambda)\big| = \mathbb{E}\sup_{\|\lambda\|_{\ell_1}\le eR}\left|R_n\left(\sum_{i=1}^N \lambda_i h_i\right)\right|$$
$$\le \mathbb{E}\sup_{\|\lambda\|_{\ell_1}\le eR}\|\lambda\|_{\ell_1}\max_{1\le i\le N}\big|R_n(h_i)\big| \le CR\sqrt{\frac{\log N}{n}}.$$

As a result, for all R we have the following bound that holds with probability $\ge 1-e^{-t}$ and with some constants C, L depending only on ℓ:

$$\sup_{\|\lambda\|_{\ell_p}\le R}\big|(P_n-P)(\ell\bullet f_\lambda-\ell\bullet f_0)\big| \le C(1+LR)R\sqrt{\frac{\log N+t}{n}}.$$

It remains to plug in $t = A\log N$ and to adjust the constants one more time to complete the proof.

The proof of another inequality is quite similar. □

Lemma 3. *There exist constants $C, L>0$ that depends only on ℓ such that for any $\bar\lambda\in\mathbb{R}^N$ with probability at least $1-N^{-A}$ for all $\lambda\in\mathbb{R}^N$ such that $\|\lambda-\bar\lambda\|_{\ell_1}\le e^N$*

$$\big|(P_n-P)(\ell\bullet f_\lambda-\ell\bullet f_{\bar\lambda})\big|$$
$$\le C\big(1+L\big(\|\lambda\|_{\ell_1}\vee\|\bar\lambda\|_{\ell_1}\big)\big)\Big(\sum_{j\in J_{\bar\lambda}}|\lambda_j-\bar\lambda_j|\Big)\sqrt{\frac{A\log N}{n}} + C\big(1+L\|\lambda\|_{\ell_1}\big)\Big(\sum_{j\notin J_{\bar\lambda}}|\lambda_j|\Big)\sqrt{\frac{A\log N}{n}}.$$

In the next lemma, we are assuming the conditions and using the notations of Theorem 4 and its proof (recall, in particular, the definitions of $\alpha_n(\delta;\Delta)$ and $\check\alpha_n(\delta;\Delta)$).

Lemma 4. *Let $C_1>0$. There exist constants $L, C, c>0$ that depend only on ℓ such that with probability at least $1-N^{-A}$, for all*

$$n^{-1/2}\le\delta\le C_1\big\|\lambda^{\varepsilon/c}\big\|_{\ell_1}\quad\text{and}\quad n^{-1/2}\le\Delta\le C_1\big\|\lambda^{\varepsilon/c}\big\|_{\ell_1}, \tag{A.1}$$

the following bounds hold:

$$\alpha_n(\delta;\Delta) \le \beta_n(\delta;\Delta)$$
$$:= C\big(1+\big\|\lambda^0\big\|_{\ell_1}\big)\Big[\delta\sqrt{\frac{d+A\log N}{n}}\vee\Delta\Big(\sqrt{\frac{\log N}{n}}\vee\Big(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\Big)\Big)\Big]$$
$$\vee C\big(1+L\big\|\lambda^0\big\|_{\ell_1}\big)\big\|\lambda^0\big\|_{\ell_1}\frac{A\log N}{n} \tag{A.2}$$

and

$$\check\alpha_n(\delta;\Delta) \le \check\beta_n(\delta;\Delta)$$
$$:= C\big(1+L\big\|\lambda^\varepsilon\big\|_{\ell_1}\big)\Big[\delta\sqrt{\frac{d+A\log N}{n}}\vee\Delta\Big(\sqrt{\frac{\log N}{n}}\vee\Big(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\Big)\Big)\Big]$$
$$\vee\, C\big(1+L\big\|\lambda^0\big\|_{\ell_1}\big)\big\|\lambda^0\big\|_{\ell_1}\frac{A\log N}{n}. \tag{A.3}$$

Proof. First note that, by Talagrand's concentration inequality, with probability at least $1 - e^{-t}$

$$\alpha_n(\delta; \Delta) \le 2\left[\mathbb{E}\alpha_n(\delta; \Delta) + L\delta\sqrt{\frac{t}{n}} + \frac{C(1 + L\|\lambda^0\|_{\ell_1})\|\lambda^0\|_{\ell_1} t}{n}\right].$$

Next, by the symmetrization inequality,

$$\mathbb{E}\alpha_n(\delta; \Delta) \le 2\mathbb{E}\sup\Big\{\big|R_n\big((\ell' \bullet f_\lambda)(f_\lambda - f_{\lambda^\varepsilon})\big)\big|\colon \lambda \in \Lambda(\delta; \Delta)\Big\}.$$

Now we write

$$\ell'\big(f_\lambda(\cdot)\big)\big(f_\lambda(\cdot) - f_{\lambda^\varepsilon}(\cdot)\big) = \ell'\big(f_{\lambda^\varepsilon}(\cdot) + u\big)u|_{u = f_\lambda(\cdot) - f_{\lambda^\varepsilon}(\cdot)}$$

and use the fact that the function

$$[-R, R] \ni u \mapsto \ell'\big(f_{\lambda^\varepsilon}(\cdot) + u\big)u$$

is Lipschitz with constant $C(1 + \|\lambda^\varepsilon\|_{\ell_1} + R)$ to prove that by the Rademacher contraction inequality

$$\mathbb{E}\alpha_n(\delta; \Delta) \le C\big(1 + \big\|\lambda^{\varepsilon/c}\big\|_{\ell_1}\big)\mathbb{E}\sup\Big\{\big|R_n(f_\lambda - f_{\lambda^\varepsilon})\big|\colon \lambda \in \Lambda(\delta; \Delta)\Big\}.$$

To bound the last expectation, for $\lambda \in \mathbb{R}^N$, denote

$$f_{\lambda, J} := \sum_{j \in J} \lambda_j h_j.$$

Then, for λ's in the supremum,

$$\|f_{\lambda, J} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le \|f_\lambda - f_{\lambda, J}\|_{L_2(\Pi)} + \|f_\lambda - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le \delta + e\Delta,$$

where we used the fact that

$$\|f_\lambda - f_{\lambda, J}\| \le \sum_{j \notin J} |\lambda_j| \|h_j\|_\infty \le e\left(\sum_{j \notin J} |\lambda_j|^p\right)^{1/p} \le e\Delta.$$

Now we get

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\big|R_n(f_\lambda - f_{\lambda^\varepsilon})\big|\colon \lambda \in \Lambda(\delta; \Delta)\Big\} \\
&\quad \le \mathbb{E}\sup\Big\{\big|R_n(f_{\lambda, J} - f_{\lambda^\varepsilon})\big|\colon \|f_{\lambda, J} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le \delta + e\Delta\Big\} + \mathbb{E}\sup\left\{\big|R_n(f_\lambda - f_{\lambda, J})\big|\colon \sum_{j \notin J} |\lambda_j| \le \Delta\right\}.
\end{aligned}$$

In the first expectation, the supremum is over a set in a d-dimensional linear space of functions (the linear span of $\{h_j\colon j \in J\}$) with a bound on $L_2(\Pi)$-norm. This yields (in a standard way, see, e.g., [16], Section 2, Example 1) that

$$\mathbb{E}\sup\Big\{\big|R_n(f_{\lambda, J} - f_{\lambda^\varepsilon})\big|\colon \|f_{\lambda, J} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le \delta + e\Delta\Big\} \le C(\delta + e\Delta)\sqrt{\frac{d}{n}}$$

with some constant $C > 0$. The bound on the second expectation is also rather standard:

$$\begin{aligned}
&\mathbb{E}\sup\left\{\big|R_n(f_\lambda - f_{\lambda, J})\big|\colon \sum_{j \notin J} |\lambda_j| \le \Delta\right\} \\
&\quad \le \mathbb{E}\sup\left\{\left|R_n\left(\sum_{j \notin J} \lambda_j h_j\right)\right|\colon \sum_{j \notin J} |\lambda_j| \le \Delta\right\} \le \Delta\mathbb{E}\max_{j \notin J}\big|R_n(h_j)\big| \le C\Delta\sqrt{\frac{\log N}{n}}.
\end{aligned}$$

Therefore

$$\mathbb{E}\sup\bigl\{\bigl|R_n(f_\lambda - f_{\lambda^\varepsilon})\bigr|\colon \lambda \in \Lambda(\delta;\Delta)\bigr\} \le C(\delta + e\Delta)\sqrt{\frac{d}{n}} + C\Delta\sqrt{\frac{\log N}{n}}.$$

It follows that with probability at least $1 - e^{-t}$

$$\begin{aligned}\alpha_n(\delta;\Delta) &\le \tilde{\beta}_n(\delta;\Delta;t)\\ &:= C\bigl(\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1} + 1\bigr)\biggl[\delta\sqrt{\frac{d}{n}} + \delta\sqrt{\frac{t}{n}} + \Delta\sqrt{\frac{d}{n}} + \Delta\sqrt{\frac{\log N}{n}}\biggr]\\ &\quad + C\bigl(1 + L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\bigl\|\lambda^0\bigr\|_{\ell_1}\frac{t}{n}\end{aligned}$$

with a constant $C > 0$ that depends only on ℓ.

To construct a bound that is uniform in δ, Δ satisfying (A.1), define

$$\delta_j := C_1\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}2^{-j} \quad \text{and} \quad \Delta_j := C_1\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}2^{-j}$$

and replace t by $t + 2\log(j+1) + 2\log(k+1)$. Using the union bound, we get that with probability at least

$$1 - \sum_{j,k\ge 0}\exp\bigl\{-t - 2\log(j+1) - 2\log(k+1)\bigr\} = 1 - \biggl(\sum_{j\ge 0}(j+1)^{-2}\biggr)^2\exp\{-t\} \ge 1 - 4e^{-t}$$

we have, for all δ and Δ satisfying (A.1), and for j, k such that

$$\delta \in (\delta_{j+1}, \delta_j] \quad \text{and} \quad \Delta \in (\Delta_{k+1}, \Delta_k],$$

that

$$\alpha_n(\delta;\Delta) \le \tilde{\beta}_n(\delta_j, \Delta_j, t + 2\log j + 2\log k).$$

Note that

$$2\log j \le 2\log\log_2\biggl(\frac{C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}}{\delta_j}\biggr) \le 2\log\log_2\biggl(\frac{2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}}{\delta}\biggr)$$

and

$$2\log k \le 2\log\log_2\biggl(\frac{2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}}{\Delta}\biggr).$$

Thus

$$\begin{aligned}&\tilde{\beta}_n(\delta_j, \Delta_j, t + 2\log j + 2\log k)\\ &\quad \le \tilde{\beta}_n\biggl(2\delta, 2\Delta, t + 2\log\log_2\biggl(\frac{2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}}{\delta}\biggr) + 2\log\log_2\biggl(\frac{2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}}{\Delta}\biggr)\biggr) =: \bar{\beta}_n(\delta;\Delta;t)\end{aligned}$$

and we have with probability at least $1 - 4e^{-t}$, for all δ and Δ satisfying (A.1),

$$\alpha_n(\delta;\Delta) \le \bar{\beta}_n(\delta;\Delta;t).$$

Take now $t = A\log N + \log 4$ (so that $4e^{-t} = N^{-A}$). Clearly, with some constant C that depends only on ℓ we have (note that $e\|\lambda^0\|_{\ell_1}$ is an upper bound on $\|\lambda^{\varepsilon/c}\|_{\ell_1}$)

$$\bar{\beta}_n(\delta;\Delta;t) \le C\bigl(1+\|\lambda^0\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d}{n}} \vee \delta\sqrt{\frac{A\log N}{n}} \vee \delta\sqrt{\frac{2\log\log_2(2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}/\delta)}{n}}$$
$$\vee\, \delta\sqrt{\frac{2\log\log_2(2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}/\Delta)}{n}} \vee \Delta\sqrt{\frac{d}{n}} \vee \Delta\sqrt{\frac{\log N}{n}}\biggr] \vee C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}.$$

Note that, for all δ and Δ satisfying (A.1),

$$\delta\sqrt{\frac{2\log\log_2(2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}/\delta)}{n}} \le C\delta\sqrt{\frac{\log\log n}{n}}$$

and

$$\delta\sqrt{\frac{2\log\log_2(2C_1\|\lambda^{\varepsilon/c}\|_{\ell_1}/\Delta)}{n}} \le C\delta\sqrt{\frac{\log\log n}{n}}$$

(where we used the fact that $\|\lambda^{\varepsilon/c}\|_{\ell_1} = \mathrm{O}(\varepsilon^{-1})$). Since $A\log N \ge \gamma\log n \ge \gamma\log\log n$, it follows that, for δ and Δ satisfying (A.1),

$$\bar{\beta}_n(\delta;\Delta;t) \le C\bigl(1+\|\lambda^0\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d+A\log N}{n}} \vee \Delta\sqrt{\frac{d+\log N}{n}}\biggr]$$
$$\vee C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}.$$

Then, with probability at least $1-2N^{-A}$, for all δ and Δ satisfying (A.1),

$$\alpha_n(\delta;\Delta) \le C\bigl(1+\|\lambda^0\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d+A\log N}{n}} \vee \Delta\sqrt{\frac{d+\log N}{n}}\biggr]$$
$$\vee\, C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n}$$

(the bound $2N^{-A}$ is the sum of two probability bounds: one N^{-A} comes from Theorem 1 and another one from Talagrand's inequality, which we used earlier in the proof).

Similarly, it is possible to prove that with probability at least $1-2N^{-A}$

$$\breve{\alpha}_n(\delta;\Delta) \le C\bigl(1+L\|\lambda^\varepsilon\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d+A\log N}{n}} \vee \Delta\sqrt{\frac{d+\log N}{n}}\biggr]$$
$$\vee\, C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n},$$

where C, L depend only on ℓ (the main advantage of this bound is in the case when the first derivative of the loss function is bounded and, as a consequence, $L=0$). The only difference with the previous case is that since now the Rademacher process is indexed by functions

$$\bigl(\ell'\bullet f_{\lambda^\varepsilon}\bigr)(f_\lambda - f_{\lambda^\varepsilon}),\quad \lambda\in\Lambda(\delta;\Delta),$$

there is no need to use the contraction inequality. Instead, one can introduce the functions

$$\tilde{h}_j := \bigl(\ell'\bullet f_{\lambda^\varepsilon}\bigr)h_j,\quad j=1,\dots,N,$$

that are bounded by $C(1+L\|\lambda^{\varepsilon}\|_{\ell_1})$, and use the fact that the Rademacher process is indexed by functions from the linear span of $\tilde{h}_1,\dots,\tilde{h}_N$, which simplifies the derivation and allows one to get the bound above. Hence, both inequalities hold simultaneously with probability at least $1-4N^{-A}$.

A relatively simple modification of the previous derivations also yields the following bounds

$$\alpha_n(\delta;\Delta)\le C\bigl(1+\|\lambda^0\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d+A\log N}{n}}\vee\Delta\sqrt{\frac{\log N}{n}}\vee\Delta U(J)\frac{\log N}{n}\biggr]$$
$$\vee C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n} \tag{A.4}$$

and

$$\check{\alpha}_n(\delta;\Delta)\le C\bigl(1+\|\lambda^{\varepsilon}\|_{\ell_1}\bigr)\biggl[\delta\sqrt{\frac{d+A\log N}{n}}\vee\Delta\sqrt{\frac{\log N}{n}}\vee\Delta U(J)\frac{\log N}{n}\biggr]$$
$$\vee C\bigl(1+L\|\lambda^0\|_{\ell_1}\bigr)\|\lambda^0\|_{\ell_1}\frac{A\log N}{n} \tag{A.5}$$

that again hold with probability at least $1-4N^{-A}$. Indeed,

$$f_\lambda - f_{\lambda^{\varepsilon}} = P_J(f_\lambda - f_{\lambda^{\varepsilon}}) + \sum_{j\notin J}\lambda_j\bigl(h_j-h'_j\bigr)$$

and, for all $\lambda\in\Lambda(\delta,\Delta)$,

$$\bigl\|P_J(f_\lambda - f_{\lambda^{\varepsilon}})\bigr\|_{L_2(\Pi)}\le\|f_\lambda - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}\le\delta.$$

Therefore,

$$\mathbb{E}\sup\Bigl\{\bigl|R_n\bigl(P_J(f_\lambda - f_{\lambda^{\varepsilon}})\bigr)\bigr|\colon\lambda\in\Lambda(\delta;\Delta)\Bigr\}\le C\delta\sqrt{\frac{d}{n}}.$$

On the other hand,

$$\mathbb{E}\sup\biggl\{\biggl|R_n\biggl(\sum_{j\notin J}\lambda_j\bigl(h_j-h'_j\bigr)\biggr)\biggr|\colon\lambda\in\Lambda(\delta;\Delta)\biggr\}\le\Delta\mathbb{E}\max_{j\notin J}\bigl|R_n\bigl(h_j-h'_j\bigr)\bigr|.$$

Standard bounding of the sup-norm of Rademacher sums yields

$$\mathbb{E}\max_{j\notin J}\bigl|R_n\bigl(h_j-h'_j\bigr)\bigr|\le C\mathbb{E}\max_{j\notin J}\|h_j-h'_j\|_{L_2(\Pi_n)}\sqrt{\frac{\log N}{n}}$$
$$\le C\max_{j\notin J}\|h_j-h'_j\|_{L_2(\Pi)}\sqrt{\frac{\log N}{n}}$$
$$+\sqrt{\mathbb{E}\max_{j\notin J}\bigl|\|h_j-h'_j\|^2_{L_2(\Pi_n)}-\|h_j-h'_j\|^2_{L_2(\Pi)}\bigr|}\sqrt{\frac{\log N}{n}},$$

which by using the symmetrization inequality in combination with the Rademacher contraction inequality gives the following bound (note that $\|h_j-h'_j\|_{L_2(\Pi)}\le 1$)

$$\mathbb{E}\max_{j\notin J}\bigl|R_n\bigl(h_j-h'_j\bigr)\bigr|\le C\biggl[\sqrt{\frac{\log N}{n}}+\sqrt{\max_{j\notin J}\|h_j-h'_j\|_{\infty}\mathbb{E}\max_{j\notin J}\bigl|R_n\bigl(h_j-h'_j\bigr)\bigr|}\sqrt{\frac{\log N}{n}}\biggr].$$

Solving this inequality for

$$\mathbb{E}\max_{j\notin J}\bigl|R_n\bigl(h_j-h'_j\bigr)\bigr|$$

results in the bound

$$\mathbb{E}\max_{j\notin J}\bigl|R_n\bigl(h_j-h'_j\bigr)\bigr|\le C\biggl[\sqrt{\frac{\log N}{n}}+U(J)\frac{\log N}{n}\biggr],$$

which, by a repetition of the previous argument, implies (A.4) and (A.5).

Combining all these bounds, we show that (A.2) and (A.3) hold with probability at least $1-8N^{-A}$. The factor 8 can be easily removed by changing the value of the constant C, implying the statement of the lemma. □

In Lemma 5, the conditions and notations of Theorem 6 are used.

Lemma 5. *Let $C_1>0$. There exist constants $L,C,c>0$ that depend only on ℓ such that with probability at least $1-N^{-A}$, for all*

$$n^{-1/2}\le\delta\le C_1\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1}\quad\text{and}\quad n^{-1/2}\le\Delta\le C_1\bigl\|\lambda^{\varepsilon/c}\bigr\|_{\ell_1},$$

the following bounds hold:

$$\begin{aligned}\alpha_n(\delta;\Delta)\le\beta_n(\delta;\Delta):=C\bigl(1+\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\biggl[&\delta\sqrt{\frac{d+A\log N}{n}}\vee\Delta\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)\\&\vee\gamma_d\bigl(\lambda^{\varepsilon}\bigr)\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)\biggr]\\&\vee C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\bigl\|\lambda^0\bigr\|_{\ell_1}\frac{A\log N}{n}\end{aligned}\tag{A.6}$$

and

$$\begin{aligned}\breve{\alpha}_n(\delta;\Delta)\le\breve{\beta}_n(\delta;\Delta):=C\bigl(1+L\bigl\|\lambda^{\varepsilon}\bigr\|_{\ell_1}\bigr)\biggl[&\delta\sqrt{\frac{d+A\log N}{n}}\vee\Delta\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)\\&\vee\gamma_d\bigl(\lambda^{\varepsilon}\bigr)\biggl(\sqrt{\frac{A\log N}{n}}\vee\biggl(U(J)\frac{\log N}{n}\wedge\sqrt{\frac{d}{n}}\biggr)\biggr)\biggr]\\&\vee C\bigl(1+L\bigl\|\lambda^0\bigr\|_{\ell_1}\bigr)\bigl\|\lambda^0\bigr\|_{\ell_1}\frac{A\log N}{n}.\end{aligned}\tag{A.7}$$

Acknowledgment

The author is thankful to the anonymous referee for numerous helpful suggestions and corrections.

References

[1] A. Barron, L. Birgé and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** (1999) 301–413. MR1679028

[2] P. Bartlett, O. Bousquet and S. Mendelson. Local Rademacher complexities. *Ann. Statist.* **33** (2005) 1497–1537. MR2166554

[3] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization. Analysis, Algorithms and Engineering Applications.* MPS/SIAM, *Series on Optimization,* Philadelphia, 2001. MR1857264

[4] F. Bunea, A. Tsybakov and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.* **35** (2007) 1674–1697. MR2351101

[5] F. Bunea, A. Tsybakov and M. Wegkamp. Sparsity oracle inequalities for the LASSO. *Electron. J. Statist.* **1** (2007) 169–194. MR2312149

[6] E. Candes and T. Tao. The Dantzig selector statistical estimation when p is much larger than n. *Ann. Statist.* **35** (2007) 2313–2351.

[7] E. Candes, M. Rudelson, T. Tao and R. Vershynin. Error correction via linear programming. In *Proc. 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS05)* 295–308. IEEE, Pittsburgh, PA, 2005.

[8] E. Candes, J. Romberg and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** (2006) 1207–1223. MR2230846

[9] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Springer, New York, 2004. MR2163920

[10] D. L. Donoho. For most large underdetermined systems of equations the minimal ℓ^1-norm near-solution approximates the sparsest near-solution. Preprint, 2004.

[11] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* **59** (2006) 797–829. MR2217606

[12] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory* **52** (2006) 1289–1306. MR2241189

[13] D. L. Donoho, M. Elad and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** (2006) 6–18. MR2237332

[14] van de S. Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** (2008) 614–645.

[15] V. Koltchinskii. Model selection and aggregation in sparse classification problems. Oberwolfach Reports Meeting on Statistical and Probabilistic Methods of Model Selection, October, 2005.

[16] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk mnimization. *Ann. Statist.* **34** (2006) 2593–2656. MR2329442

[17] V. Koltchinskii and D. Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *Ann. Statist.* **33** (2005) 1455–1496. MR2166553

[18] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, New York, 1991. MR1102015

[19] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Tolouse (IX)* **9** (2000) 245–303. MR1813803

[20] P. Massart. *Concentration Inequalities and Model Selection*. Springer, Berlin, 2007. MR2319879

[21] S. Mendelson, A. Pajor and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in Asymptotic Geometric Analysis. *Geom. Funct. Anal.* **17** (2007) 1248–1282.

[22] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the LASSO. *Ann. Statist.* **34** (2006) 1436–1462. MR2278363

[23] A. Nemirovski. Topics in non-parametric statistics. In *Ecole d'Et'e de Probabilités de Saint-Flour XXVIII*, *1998* 85–277. P. Bernard (Ed). Springer, New York, 2000. MR1775640

[24] M. Rudelson and R. Vershynin. Geometric approach to error correcting codes and reconstruction of signals. *Int. Math. Res. Not.* **64** (2005) 4019–4041. MR2206919

[25] R. Tibshirani. Regression shrinkage and selection via Lasso. *J. Royal Statist. Soc. Ser. B* **58** (1996) 267–288. MR1379242

[26] A. Tsybakov. Optimal rates of aggregation. In *Proc. 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines*, 303–313. *Lecture Notes in Artificial Intelligence* **2777**. Springer, New York, 2003.

[27] van der A. Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996. MR1385671

[28] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.* **28** (2000) 75–87. MR1762904

[29] Y. Yang. Aggregating regression procedures for a better performance. *Bernoulli* **10** (2004) 25–47. MR2044592

[30] P. Zhao and B. Yu. On model selection consistency of LASSO. *J. Mach. Learn. Res.* **7** (2006) 2541–2563. MR2274449