Sparse Recovery in Large Ensembles of Kernel Machines

Vladimir Koltchinskii*
School of Mathematics, Georgia Institute of Technology
vlad@math.gatech.edu

Ming Yuan†
School of Industrial and Systems Engineering, Georgia Institute of Technology
myuan@iyse.gatech.edu

Abstract

A problem of learning a prediction rule that is approximated in a linear span of a large number of reproducing kernel Hilbert spaces is considered. The method is based on penalized empirical risk minimization with ℓ_1-type complexity penalty. Oracle inequalities on excess risk of such estimators are proved showing that the method is adaptive to unknown degree of "sparsity" of the target function.

1 Introduction

Let (X, Y) be a random couple in $S \times T$, where $(S, \mathcal{S}), (T, \mathcal{T})$ are measurable spaces. Usually, T is either a finite set, or a subset of $\mathbb{R}$ (in the first case, T can be also identified with a finite subset of $\mathbb{R}$). Most often, S is a compact domain in a finite dimensional Euclidean space, or a compact manifold. Let P denote the distribution of (X, Y) and Π denote the distribution of X. In a general framework of prediction, X is an observable instance and Y is an unobservable label which is to be predicted based on an observation of X. Let $\ell : T \times \mathbb{R} \mapsto \mathbb{R}_+$ be a loss function. It will be assumed in what follows that, for all $y \in T$, the function $\ell(y; \cdot)$ is convex. Given $f : S \mapsto \mathbb{R}$, denote

$$(\ell \bullet f)(x, y) := \ell(y, f(x))$$

and define the (true) risk of f as

$$\mathbb{E}\ell(Y; f(X)) = P(\ell \bullet f).$$

The prediction problem then can be formulated as convex risk minimization problem with the optimal prediction rule f_* defined as

$$f_* := \operatorname{argmin}_{f:S\mapsto\mathbb{R}} P(\ell \bullet f)$$

where the minimum is taken over all measurable functions $f : S \mapsto \mathbb{R}$. It will be assumed in what follows that f_* exists and it is uniformly bounded. We shall also assume the uniqueness of f_* in the following discussion.

In the case when the distribution P of (X, Y) is unknown, it has to be estimated based on the training data which (in the simplest case) consists of n independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y). Let P_n denote the empirical distribution based on the training data. Then the risk $P(\ell \bullet f)$ can be estimated by the empirical risk

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

The direct minimization of the empirical risk over a large enough family of function $f : S \mapsto \mathbb{R}$ almost inevitably leads to overfitting. To avoid it, a proper complexity regularization is needed. In this paper, we will study a problem in which the unknown target function f_* is being approximated in a linear span $\mathcal{H}$ of a large dictionary consisting of N reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_1, \dots, \mathcal{H}_N$. It will be assumed that we are given N symmetric nonnegatively definite kernels $K_j : S \times S \mapsto \mathbb{R}$, $j = 1, \dots, N$ and that $\mathcal{H}_j$ is the RKHS generated by $K_j : \mathcal{H}_j = \mathcal{H}_{K_j}$. Suppose, for simplicity, that

$$K_j(x, x) \le 1, \ x \in S, \ j = 1, \dots, N.$$

The space

$$\mathcal{H} := \text{l.s.}\Big(\bigcup_{j=1}^{N} \mathcal{H}_j\Big)$$

consists of all functions $f : S \mapsto \mathbb{R}$ that have the following (possibly, non-unique) additive representation

$$f = f_1 + \dots f_N, \ f_j \in \mathcal{H}_j, \quad f_j \in \mathcal{H}_j, \ j = 1, \dots, N$$

and it can be naturally equipped with the ℓ_1-norm:

$$\|f\|_{\ell_1} := \|f\|_{\ell_1(\mathcal{H})} := \inf\Big\{\sum_{j=1}^{N} \|f_j\|_{\mathcal{H}_j} : f = \sum_{j=1}^{N} f_j, f_j \in \mathcal{H}_j, j = 1, \dots, N\Big\}.$$

Additive models are a well-known special case of this formulation. In additive models, S is a subset of

*Partially supported by NSF grant DMS-0624841.
†Partially supported by NSF grant DMS-0624841.

$\mathbb{R}^N$, i.e., $X = (x_1, \dots, x_N)'$, and $\mathcal{H}_j$ represents a functional space defined over x_j. Several approaches have been proposed recently to exploit the sparsity in additive models (Lin and Zhang, 2006; Ravikumar et al., 2007; Yuan, 2007). In this paper, we consider an extension of ℓ_1 penalization technique to a more general class of problem.

In particular, we study the following penalized empirical risk minimization problem:

$$\hat{f}^{\varepsilon} := \operatorname{argmin}_{f \in \mathcal{H}} \Big[P_n(\ell \bullet f) + \varepsilon \|f\|_{\ell_1} \Big], \tag{1.1}$$

where $\varepsilon > 0$ is a small regularization parameter. Equivalently, this problem can be written as

$$(\hat{f}_1^{\varepsilon}, \dots, \hat{f}_N^{\varepsilon}) := \operatorname{argmin}_{f_j \in \mathcal{H}_j, j=1,\dots,N} \Big[P_n(\ell \bullet (f_1 + \cdots + f_N)) + \varepsilon \sum_{j=1}^{N} \|f_j\|_{\mathcal{H}_j} \Big]. \tag{1.2}$$

According to the representer theorem (Wahba, 1990), the components of the minimizer $\hat{f}_j^{\varepsilon}$ have the following representation:

$$\hat{f}_j^{\varepsilon}(x) = \sum_{i=1}^{n} \hat{c}_{ij} K_j(X_i, x)$$

for some real vector $\hat{\mathbf{c}}_j = (\hat{c}_{ij} : i = 1, \dots, n)$. In other words, (1.2) can be rewritten as a finite dimensional convex minimization problem over $(c_{ij} : i = 1, \dots, n; j = 1, \dots, N)$.

It is known (see, e.g., Micchelli and Pontil, 2005) that

$$\|f\|_{\ell_1(\mathcal{H})} = \inf \Big\{ \|f\|_K : K \in \operatorname{conv}\{K_j : j = 1, \dots, N\} \Big\},$$

where $\|\cdot\|_K$ denote the RKHS-norm generated by symmetric nonnegatively definite kernel K and

$$\operatorname{conv}\{K_j : j = 1, \dots, N\} := \Big\{ \sum_{j=1}^{N} c_j K_j : c_j \geq 0, \sum_{j=1}^{N} c_j = 1 \Big\}.$$

Therefore (1.2) can be also written as

$$(\hat{f}^{\varepsilon}, \hat{K}^{\varepsilon}) := \operatorname{argmin}_{K \in \operatorname{conv}(K_j, j=1,\dots,N)} \operatorname{argmin}_{f \in \mathcal{H}_K} \Big[P_n(\ell \bullet f) + \varepsilon \|f\|_K \Big], \tag{1.3}$$

leading to an interpretation of the problem as the one of learning not only the target function f_*, but also the kernel K in the convex hull of a given dictionary of kernels (which can be viewed as "aggregation" of kernel machines). Similar problems have been studied recently by Bousquet et al. (2003), Cramer et al. (2003), Lanckriet et al. (2004), Micchelli and Pontil (2005) and Srebro and Ben-David (2006) among others.

The choice of ℓ_1-norm for complexity penalization is related to our interest in the case when the total number N of spaces $\mathcal{H}_j$ in the dictionary is very large, but the target function f_* can be approximated reasonably well by functions from relatively small number d of such spaces. The ℓ_1-penalization technique has been commonly used to recover sparse solutions in the case of simple dictionaries that consist of one-dimensional spaces $\mathcal{H}_j$ (see, e.g, Koltchinskii (2007) and references therein). The goal is to extend this methodology to more general class of problems that include aggregation of large ensembles of kernel machines and sparse additive models. In the case of additive models with the quadratic loss, (1.1) becomes the so-called COSSO estimate recently introduced by Lin and Zhang (2006).

For $f \in \mathcal{H}$, define the excess risk of f as

$$\mathcal{E}(f) = P(\ell \bullet f) - P(\ell \bullet f_*) = P(\ell \bullet f) - \inf_{g: S \mapsto \mathbb{R}} P(\ell \bullet g).$$

Our main goal is to control the excess risk of $\hat{f}^{\varepsilon}$, $\mathcal{E}(\hat{f}^{\varepsilon})$.

Throughout the paper, we shall also make the following assumption

$$n^{\gamma} \leq N \leq e^n$$

for some $\gamma > 0$.

It will also be assumed that the loss function ℓ satisfies the following properties: for all $y \in T$, $\ell(y, \cdot)$ is twice differentiable, ℓ''_u is a uniformly bounded function in $T \times \mathbb{R}$,

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

and

$$\tau(R) := \frac{1}{2} \inf_{y \in T} \inf_{|u| \leq R} \ell''_u(y, u) > 0, \ R > 0. \tag{1.4}$$

We also assume without loss of generality that, for all R, $\tau(R) \leq 1$. These assumptions imply that

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

with some constants $L_1, L \geq 0$ (if ℓ'_u is uniformly bounded, one can take $L = 0$).

The following bound on the excess risk holds under the assumptions on the loss:

$$\begin{aligned} & \tau(\|f\|_{\infty} \vee \|f_*\|_{\infty}) \|f - f_*\|^2_{L_2(\Pi)} \\ \leq \ & \mathcal{E}(f) \leq C \|f - f_*\|^2_{L_2(\Pi)} \end{aligned} \tag{1.5}$$

with a constant $C > 0$ depending only on ℓ. This bound easily follows from a simple argument based on Taylor expansion and it will be used later in the paper.

The quadratic loss $\ell(y, u) := (y - u)^2$ in the case when $T \subset \mathbb{R}$ is a bounded set is one of the main examples of such loss functions. In this case, $\tau(R) = 1$ for all R. In regression problems with a bounded response variable, more general loss functions of the form $\ell(y, u) := \phi(y - u)$ can be also used, 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 problems, the loss function of the form $\ell(y,u) = \phi(yu)$ is commonly used, 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 $\phi(u) = \log_2(1+e^{-u})$ (often referred to as the logit loss) is a specific example.

We will assume in what follows that $\mathcal{H}$ is dense in $L_2(\Pi)$, which, together with (1.5), implies that

$$\inf_{f\in\mathcal{H}} P(\ell \bullet f) = \inf_{f\in L_2(\Pi)} P(\ell \bullet f) = P(\ell \bullet f_*).$$

We also need several basic facts about RKHS which can be found in, for example, Wahba (1990). Let K be a symmetric nonnegatively definite kernel on $S\times S$ with

$$\sup_{x\in S} K(x,x) \le 1$$

and $\mathcal{H}_K$ be the corresponding RKHS. Given a probability measure Π on S, let $\phi_k, k \ge 1$ be the orthonormal system of fuctions in $L_2(\Pi)$ such that the following spectral representation (as in Mercer's theorem) holds:

$$K(x,y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(y),\ x,y \in S,$$

which is true under mild regularity conditions. Without loss of generality we can and do assume that $\{\lambda_k\}$ is a decreasing sequence, $\lambda_k \to 0$. It is well known that for $f,g \in \mathcal{H}_K$,

$$\langle f,g\rangle_{\mathcal{H}_K} = \sum_{k=1}^{\infty} \frac{\langle f,\phi_k\rangle_{L_2(\Pi)}\langle g,\phi_k\rangle_{L_2(\Pi)}}{\lambda_k}.$$

Denote $H_D \subset \mathcal{H}_K$ the linear span of functions $f \in \mathcal{H}_K$ such that

$$\sum_{k=1}^{\infty} \frac{\langle f,\phi_k\rangle^2_{L_2(\Pi)}}{\lambda_k^2} < \infty$$

and let $D : H_D \mapsto L_2(\Pi)$ be a linear operator defined as follows:

$$Df := \sum_{k=1}^{\infty} \frac{\langle f,\phi_k\rangle_{L_2(\Pi)}}{\lambda_k}\phi_k,\ f \in H_D.$$

Then we obviously have

$$\langle f,g\rangle_{\mathcal{H}_K} = \langle Df, g\rangle_{L_2(\Pi)}, f \in H_D, g \in \mathcal{H}_K.$$

Given a dictionary $\{\mathcal{H}_1,\dots,\mathcal{H}_N\}$ of RKHS, one can quite similarly define spectral representations of kernels K_j with nonincreasing sequences of eignevalues $\{\lambda_k^{(j)} : k \ge 1\}$ and orthonormal in $L_2(\Pi)$ eigenfunctions $\{\phi_k^{(j)} : k \ge 1\}$. This also defines spaces H_{D_j} and linear operators $D_j : H_{D_j} \mapsto L_2(\Pi)$ such that

$$\langle f,g\rangle_{\mathcal{H}_j} = \langle D_j f, g\rangle_{L_2(\Pi)}, f \in H_{D_j}, g \in \mathcal{H}_{K_j}.$$

2 Bounding the ℓ_1-norm

Our first goal is to derive upper bounds on $\|\hat f^{\varepsilon}\|_{\ell_1}$ that hold with a high probability. In what follows we use the notation

$$(\ell' \bullet f)(x,y) := \ell'_u(y, f(x)),$$

where $\ell'_u(y,u)$ is the derivative of ℓ with respect to the second variable.

Theorem 1 *There exists a constant $D > 0$ depending only on ℓ such that for all $A \ge 1$ and for all $\varepsilon > 0$ and $f \in \mathcal{H}$ satisfying the condition*

$$\varepsilon \ge D\|\ell' \bullet f\|_{\infty}\sqrt{\frac{A\log N}{n}} \bigvee 4 \max_{1\le k\le N} \sup_{\|h_k\|_{\mathcal{H}_k}\le 1} |P\left((\ell' \bullet f)h_k\right)|,$$

the following bound holds

$$\mathbb{P}\left\{\|\hat f^{\varepsilon}\|_{\ell_1} \ge 3\|f\|_{\ell_1}\right\} \le N^{-A}. \tag{2.1}$$

In particular, if $\varepsilon \ge D\|\ell' \bullet f^{\varepsilon/4}\|_{\infty}\sqrt{\frac{A\log N}{n}}$, then

$$\mathbb{P}\left\{\|\hat f^{\varepsilon}\|_{\ell_1} \ge 3\|f^{\varepsilon/4}\|_{\ell_1}\right\} \le N^{-A}. \tag{2.2}$$

Proof. By the definition of $\hat f^{\varepsilon}$, for all $f \in \mathcal{H}$,

$$P_n(\ell \bullet \hat f^{\varepsilon}) + \varepsilon\|\hat f^{\varepsilon}\|_{\ell_1} \le P_n(\ell \bullet f) + \varepsilon\|f\|_{\ell_1}.$$

The convexity of the functional $f \mapsto P_n(\ell \bullet f)$ implies that

$$P_n(\ell \bullet \hat f^{\varepsilon}) - P_n(\ell \bullet f) \ge P_n\left((\ell' \bullet f)(\hat f^{\varepsilon} - f)\right).$$

As a result,

$$\begin{aligned}\varepsilon\|\hat f^{\varepsilon}\|_{\ell_1} &\le \varepsilon\|f\|_{\ell_1} + P_n\left((\ell' \bullet f)(f - \hat f^{\varepsilon})\right)\\ &\le \varepsilon\|f\|_{\ell_1} + \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1} |P_n\left((\ell' \bullet f)h_k\right)| \times\\ &\qquad \times\|\hat f^{\varepsilon} - f\|_{\ell_1}.\end{aligned}$$

It follows that

$$\begin{aligned}&\left(\varepsilon - \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1} |P_n\left((\ell' \bullet f)h_k\right)|\right)\|\hat f^{\varepsilon}\|_{\ell_1}\\ &\le \left(\varepsilon + \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1} |P_n\left((\ell' \bullet f)h_k\right)|\right)\|f\|_{\ell_1}.\end{aligned}$$

Under the assumption

$$\varepsilon > \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1} |P_n\left((\ell' \bullet f)h_k\right)|,$$

this yields

$$\|\hat f^{\varepsilon}\|_{\ell_1} \le \frac{\varepsilon + \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P_n\left((\ell' \bullet f)h_k\right)|}{\varepsilon - \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P_n\left((\ell' \bullet f)h_k\right)|}\|f\|_{\ell_1}. \tag{2.3}$$

Note that

$$\max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P_n\left((\ell'\bullet f)h_k\right)| \le \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|(P_n-P)(\ell'\bullet f)h_k| + \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P\left((\ell'\bullet f)h_k\right)|.$$

Also, for any $i=1,\dots,N$

$$\sup_{\|h_i\|_{\mathcal{H}_i}\le 1}|(P_n-P)(\ell'\bullet f)h_i| = \sup_{\|h_i\|_{\mathcal{H}_i}\le 1}\Big|n^{-1}\sum_{j=1}^n\Big((\ell'\bullet f)(X_j,Y_j)\langle h_i,K_i(X_j,\cdot)\rangle_{\mathcal{H}_i} - \mathbb{E}(\ell'\bullet f)(X_j,Y_j)\langle h_i,K_i(X_j,\cdot)\rangle_{\mathcal{H}_i}\Big)\Big| = \Big\|n^{-1}\sum_{j=1}^n\Big((\ell'\bullet f)(X_j,Y_j)K_i(X_j,\cdot) - \mathbb{E}(\ell'\bullet f)(X_j,Y_j)K_i(X_j,\cdot)\Big)\Big\|_{\mathcal{H}_i}.$$

Using Bernstein's type inequality in Hilbert spaces, we are easily getting the bound

$$\max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|(P_n-P)(\ell'\bullet f)h_k| \le C\|\ell'\bullet f\|_\infty\left(\sqrt{\frac{A\log N}{n}}\bigvee\frac{A\log N}{n}\right)$$

with probability at least $1-N^{-A}$. As soon as

$$\varepsilon \ge 4C\|\ell'\bullet f\|_\infty\left(\sqrt{\frac{A\log N}{n}}\bigvee\frac{A\log N}{n}\right)$$

and

$$\varepsilon \ge 4\max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}}|P\left((\ell'\bullet f)h_k\right)|,$$

we get

$$\max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P_n\left((\ell'\bullet f)h_k\right)| \le \varepsilon/2,$$

and it follows from (2.3) that with probability at least $1-N^{-A}$

$$\|\hat f^\varepsilon\|_{\ell_1} \le \frac{\varepsilon+\varepsilon/2}{\varepsilon-\varepsilon/2}\|f\|_{\ell_1} = 3\|f\|_{\ell_1},$$

implying (2.1).

In particular, we can use in (2.1) $f := f^{\varepsilon/4}$. Then, by the necessary conditions of extremum in the definition of $f^{\varepsilon/4}$,

$$\max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P\left((\ell'\bullet f^{\varepsilon/4})h_k\right)| \le \frac{\varepsilon}{4},$$

and the second bound follows. ∎

We now provide an alternative set of conditions on ϵ so that (2.1) holds. By the conditions on the loss,

$$\|\ell'\bullet f\|_\infty \le C(1+L\|f\|_\infty) \le C(1+L\|f\|_{\ell_1})$$

with constants C, L depending only on ℓ (if ℓ' is uniformly bounded, $L=0$).

Since, by the necessary conditions of minimum at f_*,

$$P\left((\ell'\bullet f_*)h_k\right)=0,\ h_k\in\mathcal{H}_k, k=1,\dots,N,$$

we also have

$$\max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P\left((\ell'\bullet f)h_k\right)| = \max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P\Big((\ell'\bullet f)-(\ell'\bullet f_*)\Big)h_k| \le C\|f-f_*\|_{L_2(\Pi)}$$

where we used the fact that $\ell'_u(y,u)$ is Lipschitz with respect to u. Therefore, the condition on ε in (2.1) is satisfied if

$$\varepsilon \ge D(1+\|f\|_{\ell_1})\sqrt{\frac{A\log N}{n}}$$

and

$$\|f-f_*\|_{L_2(\Pi)} \le \varepsilon/D$$

with a properly chosen D (depending only on ℓ).

3 Oracle inequalities

In what follows we will assume that $R>0$ is such that

$$\|\hat f^\varepsilon\|_{\ell_1} \le R$$

with probability at least $1-N^{-A}$. In particular, if $\bar f\in\mathcal{H}$ satisfies the assumption of Theorem 1, i.e.,

$$\varepsilon \ge D\|\ell'\bullet\bar f\|_\infty\sqrt{\frac{A\log N}{n}}\bigvee 4\max_{1\le k\le N}\sup_{\|h_k\|_{\mathcal{H}_k}\le 1}|P\left((\ell'\bullet\bar f)h_k\right)|,$$

then one can take $R=3\|\bar f\|_{\ell_1}$.

We need some measures of dependence (in a probabilistic sense) between the spaces $\mathcal{H}_j, j=1,\dots,N$. In the case of a simple dictionary $\{h_1,\dots,h_N\}$ consisting of N functions (equivalently, N one-dimensional spaces) the error of sparse recovery depends on the Gram matrix of the dictionary in the space $L_2(\Pi)$ (see, e.g., Koltchinskii (2007)). A similar approach is taken here. Given $h_j\in\mathcal{H}_j,\ j=1,\dots,N$ and $J\subset\{1,\dots,N\}$, denote by $\kappa(\{h_j: j\in J\})$ the minimal eigenvalue of the Gram matrix $\Big(\langle h_i,h_j\rangle_{L_2(\Pi)}\Big)_{i,j\in J}$ and $\bar\kappa(\{h_j: j\in J\})$ its maximal eigenvalue. Let

$$\kappa(J) := \inf\Big\{\kappa(\{h_j: j\in J\}) : h_j\in\mathcal{H}_j, \|h_j\|_{L_2(\Pi)}=1\Big\}$$

and

$$\bar\kappa(J) := \sup\Big\{\kappa(\{h_j: j\in J\}) : h_j\in\mathcal{H}_j, \|h_j\|_{L_2(\Pi)}=1\Big\}$$

Also, denote L_J the linear span of subspaces $\mathcal{H}_j, j\in J$. Let

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

In what follows, we will consider a set $\mathcal{O} = \mathcal{O}(M_1, M_2)$ of functions (more precisely, their additive representations) $f = f_1 + \cdots + f_N \in \mathcal{H},\ f_j \in \mathcal{H}_j,\ j = 1, \ldots, N$ that will be called "admissible oracles". Let $J_f := \{j : f_j \neq 0\}$ and suppose the following assumptions hold:

O1. The "relevant" part J_f of the dictionary satisfies the condition

$$\frac{\bar{\kappa}(J_f)}{\kappa(J_f)(1-\rho^2(J_f))} \leq M_1.$$

O2. For some $\beta > 1/2$ and for all $j \in J_f$

$$\lambda_k^{(j)} \leq M_2 k^{-2\beta},\ k = 1, 2, \ldots$$

Recall that D_j is the linear operator defined in the first section. Denote

$$\zeta(f) := \frac{1}{\mathrm{card}(J_f)} \sum_{j \in J_f} \frac{\|D_j f_j\|^2_{L_2(\Pi)}}{\|f_j\|^2_{\mathcal{H}_j}}.$$

We are now in the position to state the main result of this paper.

Theorem 2 *There exist constants D, L depending only on ℓ ($L = 0$ if ℓ'_u is uniformly bounded) such that for all $A \geq 1$, for all $f \in \mathcal{O}$ with $\mathrm{card}(J_f) = d$ and for all*

$$\varepsilon \geq D(1 + LR)\sqrt{\frac{\log N}{n}}$$

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

$$\begin{aligned} &\mathcal{E}(\hat{f}^\varepsilon) + 2\varepsilon \sum_{j \notin J_f} \|\hat{f}^\varepsilon_j\|_{\mathcal{H}_j} \\ \leq\ & 7\mathcal{E}(f) + K\left[\frac{d^{(2\beta-1)/(2\beta+1)}}{n^{2\beta/(2\beta+1)}} + \zeta(f) d\varepsilon^2 + \frac{A\log N}{n}\right], \end{aligned}$$

where K is a constant depending on $\ell, R, M_1, M_2, \|f\|_\infty$ and $\|f_*\|_\infty$.

The meaning of this result can be described as follows. Suppose there exists an oracle f such that the excess risk of f is small (i.e., f provides a good approximation of f_*); the set J_f is small (i.e., f has a sparse representation in the dictionary); the condition (O1) is satisfied, i.e. the relevant part of the dictionary is "well posed" in the sense that the spaces $\mathcal{H}_j, j \in J_j$ are not "too dependent" among themselves and with the rest of the spaces in the dictionary; the condition (O2) is satisfied, which means "sufficient smoothness" of functions in the spaces $\mathcal{H}_j, j \in J_f$; finally, the components $f_j, j \in J_f$ of the oracle f are even smoother in the sense that the quantities $\frac{\|D_j f_j\|_{L_2(\Pi)}}{\|f_j\|_{\mathcal{H}_j}}, j \in J_f$ are properly bounded. Then the excess risk of the empirical solution $\hat{f}^\varepsilon$ is controlled by the excess risk of the oracle as well as by its degree of sparsity d and, at the same time, $\hat{f}^\varepsilon$ is approximately sparse in the sense that $\sum_{j \notin J_f} \|f^\varepsilon_j\|_{\mathcal{H}_j}$ is small. In other words, the solution obtained via ℓ_1-penalized empirical risk minimization is adaptive to sparsity (at least, subject to constraints described above).

Proof. Throughout the proof we fix representations $f = f_1 + \cdots + f_N$ and $\hat{f}^\varepsilon = \hat{f}^\varepsilon_1 + \cdots + \hat{f}^\varepsilon_N$ (and we use (1.2) to define $\hat{f}^\varepsilon_j$). The definition of $\hat{f}^\varepsilon_j$ implies that for all $f \in \mathcal{H}$,

$$P_n(\ell \bullet \hat{f}^\varepsilon) + \varepsilon \|\hat{f}^\varepsilon\|_{\ell_1} \leq P_n(\ell \bullet f) + \varepsilon \|f\|_{\ell_1}.$$

Therefore,

$$\begin{aligned} &\mathcal{E}(\hat{f}^\varepsilon) + \varepsilon \sum_{j \notin J_f} \|\hat{f}^\varepsilon_j\|_{\mathcal{H}_j} \\ \leq\ & \mathcal{E}(f) + \varepsilon \sum_{j \in J_f} (\|f_j\|_{\mathcal{H}_j} - \|\hat{f}^\varepsilon_j\|_{\mathcal{H}_j}) \\ & + (P - P_n)(\ell \bullet f - \ell \bullet \hat{f}^\varepsilon). \end{aligned}$$

We first show that

$$\begin{aligned} &\mathcal{E}(\hat{f}^\varepsilon) + 2\varepsilon \sum_{j \notin J_f} \|\hat{f}^\varepsilon_j\|_{\mathcal{H}_j} \\ \leq\ & 3\mathcal{E}(f) + \frac{2\zeta(f) d}{\tau \kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2 \\ & + 2(P - P_n)(\ell \bullet f - \ell \bullet \hat{f}^\varepsilon), \end{aligned}$$

where

$$\tau = \tau(\|f\|_\infty \vee \|\hat{f}^\varepsilon\|_\infty \vee \|f_*\|_\infty).$$

Let $s_j(f_j)$ be a subgradient of $f_j \mapsto \|f_j\|_{\mathcal{H}_j}$ at $f_j \in \mathcal{H}_j$, i.e. $s_j(f_j) = \frac{f_j}{\|f_j\|_{\mathcal{H}_j}}$ if $f_j \neq 0$ and $s_j(f_j)$ is an arbitrary vector with $\|s_j(f_j)\|_{\mathcal{H}_j} \leq 1$ otherwise. Then we have

$$\begin{aligned} \|f_j\|_{\mathcal{H}_j} - \|\hat{f}^\varepsilon_j\|_{\mathcal{H}_j} &\leq \langle s_j(f_j), f_j - \hat{f}^\varepsilon_j \rangle_{\mathcal{H}_j} \\ &= \langle D_j s_j(f_j), f_j - \hat{f}^\varepsilon_j \rangle_{L_2(\Pi)} \\ &\leq \|D_j s_j(f_j)\|_{L_2(\Pi)} \|f_j - \hat{f}^\varepsilon_j\|_{L_2(\Pi)}. \end{aligned}$$

It follows that

$$\begin{aligned} &\mathcal{E}(\hat{f}^\varepsilon) + \varepsilon \sum_{j \notin J_f} \|\hat{f}^\varepsilon_j\|_{\mathcal{H}_j} \\ \leq\ & \mathcal{E}(f) + \varepsilon \left(\sum_{j \in J_f} \|D_j s_j(f_j)\|^2_{L_2(\Pi)}\right)^{1/2} \times \\ & \times \left(\sum_{j \in J_f} \|f_j - \hat{f}^\varepsilon_j\|^2_{L_2(\Pi)}\right)^{1/2} \\ & + (P - P_n)(\ell \bullet f - \ell \bullet \hat{f}^\varepsilon). \end{aligned}$$

It can also be shown that (see Koltchinskii, 2007, Proposition 1)

$$\begin{aligned} &\left(\sum_{j \in J_f} \|f_j - \hat{f}^\varepsilon_j\|^2_{L_2(\Pi)}\right)^{1/2} \\ \leq\ & \sqrt{\frac{1}{\kappa(J_f)(1-\rho^2(J_f))}} \|f - \hat{f}^\varepsilon\|_{L_2(\Pi)}. \end{aligned}$$

This allows us to write

$$\begin{aligned}&\mathcal{E}(\hat f^\varepsilon)+\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\ \le\;&\mathcal{E}(f)+\varepsilon\sqrt{\frac{\zeta(f)d}{\kappa(J_f)(1-\rho^2(J_f))}}\|f-\hat f^\varepsilon\|_{L_2(\Pi)}\\ &+(P-P_n)(\ell\bullet f-\ell\bullet\hat f^\varepsilon).\end{aligned}$$

Then, using the bounds

$$\|f-\hat f^\varepsilon\|_{L_2(\Pi)}\le\|f-f_*\|_{L_2(\Pi)}+\|\hat f^\varepsilon-f_*\|_{L_2(\Pi)}$$

and

$$\mathcal{E}(f)\ge\tau\|f-f_*\|^2_{L_2(\Pi)},\;\;\mathcal{E}(\hat f^\varepsilon)\ge\tau\|\hat f^\varepsilon-f_*\|^2_{L_2(\Pi)},$$

we get

$$\begin{aligned}&\mathcal{E}(\hat f^\varepsilon)+\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\ \le\;&\mathcal{E}(f)+\varepsilon\sqrt{\frac{\zeta(f)d}{\kappa(J_f)(1-\rho^2(J_f))}}\times\\ &\times\Big(\sqrt{\frac{\mathcal{E}(f)}{\tau}}+\sqrt{\frac{\mathcal{E}(\hat f^\varepsilon)}{\tau}}\Big)\\ &+(P-P_n)(\ell\bullet f-\ell\bullet\hat f^\varepsilon).\end{aligned}$$

Applying the inequality $ab\le a^2/2+b^2/2$, we show that

$$\begin{aligned}&\varepsilon\sqrt{\frac{\zeta(f)d}{\kappa(J_f)(1-\rho^2(J_f))}}\sqrt{\frac{\mathcal{E}(f)}{\tau}}\\ \le\;&\frac{\mathcal{E}(f)}{2}+\frac{\zeta(f)d}{2\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2.\end{aligned}$$

Similarly,

$$\begin{aligned}&\varepsilon\sqrt{\frac{\zeta(f)d}{\kappa(J_f)(1-\rho^2(J_f))}}\sqrt{\frac{\mathcal{E}(\hat f^\varepsilon)}{\tau}}\\ \le\;&\frac{\mathcal{E}(\hat f^\varepsilon)}{2}+\frac{\zeta(f)d}{2\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2.\end{aligned}$$

This leads to the following bound

$$\begin{aligned}&\mathcal{E}(\hat f^\varepsilon)+\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\ \le\;&\mathcal{E}(f)+\frac{\mathcal{E}(\hat f^\varepsilon)}{2}+\frac{\zeta(f)d}{2\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2\\ &+\frac{\mathcal{E}(f)}{2}+\frac{\zeta(f)d}{2\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2\\ &+(P-P_n)(\ell\bullet f-\ell\bullet\hat f^\varepsilon).\end{aligned}$$

It easily follows that

$$\begin{aligned}&\mathcal{E}(\hat f^\varepsilon)+2\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\ \le\;&3\mathcal{E}(f)+\frac{2\zeta(f)d}{\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2\\ &+2(P-P_n)(\ell\bullet f-\ell\bullet\hat f^\varepsilon).\end{aligned}$$

Denote

$$\begin{aligned}\alpha_n(\delta,\Delta,R):=\sup\Big\{&|(P_n-P)(\ell\bullet g-\ell\bullet f)|:\\ &\|g-f\|_{L_2(\Pi)}\le\delta,\sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta,\|g\|_{\ell_1}\le R\Big\}.\end{aligned}$$

If $\|\hat f^\varepsilon\|_{\ell_1}\le R$ (which holds with probability at least $1-N^{-A}$), then we have

$$\begin{aligned}&\mathcal{E}(\hat f^\varepsilon)+2\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\ \le\;&3\mathcal{E}(f)+\frac{2\zeta(f)d}{\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2\\ &+2\alpha_n\Big(\|\hat f^\varepsilon-f\|_{L_2(\Pi)},\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j},R\Big)\end{aligned}$$

with $\tau=\tau(R\vee\|f\|_\infty\vee\|f_*\|_\infty)$. We use Lemma 8 to get

$$\begin{aligned}&\mathcal{E}(\hat f^\varepsilon)+2\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\ \le\;&3\mathcal{E}(f)+\frac{2\zeta(f)d}{\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2\\ &+C(1+LR)\Big[\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\times\\ &\times\|\hat f^\varepsilon-f\|_{L_2(\Pi)}\sqrt{\frac{dm}{n}}+R\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}}+\\ &+R\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}}+\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\times\\ &\times\sqrt{\frac{\log(N-d)+1}{n}}\Big]+C(1+LR)\times\\ &\times\|\hat f^\varepsilon-f\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}}\\ &+CR(1+LR)\frac{A\log N}{n}\end{aligned}\tag{3.1}$$

(Lemma 8 can be used only under the assumption $R\le e^N$; however, for very large $R>e^N$, the proof of the inequality of the theorem is very simple). Recall that

$$\|\hat f^\varepsilon-f\|_{L_2(\Pi)}\le\sqrt{\frac{\mathcal{E}(\hat f^\varepsilon)}{\tau}}+\sqrt{\frac{\mathcal{E}(f)}{\tau}}.$$

Under the assumption

$$\varepsilon\ge C(1+LR)\sqrt{\frac{\log N}{n}},$$

we get

$$\begin{aligned}
&\mathcal{E}(\hat f^\varepsilon)+\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\
\le\;& 3\mathcal{E}(f)+2\frac{2\zeta(f)d}{\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2+\\
&C(1+LR)\Bigg[\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\times\\
&\times\Bigg(\sqrt{\frac{\mathcal{E}(\hat f^\varepsilon)}{\tau}}+\sqrt{\frac{\mathcal{E}(f)}{\tau}}\Bigg)\sqrt{\frac{dm}{n}}+R\times\\
&\times\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}}+R\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}}\Bigg]\\
&+C(1+LR)\Bigg(\sqrt{\frac{\mathcal{E}(\hat f^\varepsilon)}{\tau}}+\sqrt{\frac{\mathcal{E}(f)}{\tau}}\Bigg)\sqrt{\frac{A\log N}{n}}\\
&+CR(1+LR)\frac{A\log N}{n}. \qquad (3.2)
\end{aligned}$$

Then we have

$$\begin{aligned}
&C(1+LR)\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\sqrt{\frac{\mathcal{E}(\hat f^\varepsilon)}{\tau}}\sqrt{\frac{dm}{n}}\\
\le\;& \frac14\mathcal{E}(\hat f^\varepsilon)+2C^2(1+LR)^2\frac{\bar\kappa(J_f)}{\tau\kappa(J_f)(1-\rho^2(J_f))}\frac{dm}{n}
\end{aligned}$$

and

$$\begin{aligned}
&C(1+LR)\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\sqrt{\frac{\mathcal{E}(f)}{\tau}}\sqrt{\frac{dm}{n}}\\
\le\;& \frac14\mathcal{E}(f)+2C^2(1+LR)^2\frac{\bar\kappa(J_f)}{\tau\kappa(J_f)(1-\rho^2(J_f))}\frac{dm}{n}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&C(1+LR)\sqrt{\frac{\mathcal{E}(\hat f^\varepsilon)}{\tau}}\sqrt{\frac{A\log N}{n}}\\
\le\;& \frac14\mathcal{E}(\hat f^\varepsilon)+2C^2(1+LR)^2\frac{A\log N}{n}
\end{aligned}$$

and

$$\begin{aligned}
&C(1+LR)\sqrt{\frac{\mathcal{E}(f)}{\tau}}\sqrt{\frac{A\log N}{n}}\\
\le\;& \frac14\mathcal{E}(f)+2C^2(1+LR)^2\frac{A\log N}{n}.
\end{aligned}$$

This yields the following bound

$$\begin{aligned}
&\frac12\mathcal{E}(\hat f^\varepsilon)+\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\
\le\;& \frac72\mathcal{E}(f)+2\frac{2\zeta(f)d}{\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2+\\
&4C^2(1+LR)^2\Bigg[\frac{\bar\kappa(J_f)}{\tau\kappa(J_f)(1-\rho^2(J_f))}\frac{dm}{n}\\
&+R\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}}+\\
&R\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}}\Bigg]+4C^2(1+LR)^2\times\\
&\times\frac{A\log N}{n}+CR(1+LR)\frac{A\log N}{n}. \qquad (3.3)
\end{aligned}$$

It remains to take

$$\begin{aligned}
m :=\;& \frac{n^{1/(2\beta+1)}}{d^{2/(2\beta+1)}}\Bigg(\frac{(1+LR)^2}{R\tau}\Bigg)^{-2/(2\beta+1)}\times\\
&\times\Bigg(\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}\Bigg)^{-2/(2\beta+1)}
\end{aligned}$$

to get the following bound (with some constant $C>0$)

$$\begin{aligned}
&\mathcal{E}(\hat f^\varepsilon)+2\varepsilon\sum_{j\notin J_f}\|\hat f_j^\varepsilon\|_{\mathcal{H}_j}\\
\le\;& 7\mathcal{E}(f)+8\frac{\zeta(f)d}{\tau\kappa(J_f)(1-\rho^2(J_f))}\varepsilon^2+\\
&+C\Bigg(\frac{(1+LR)^2}{\tau}\Bigg)^{(2\beta-1)/(2\beta+1)}\times\\
&\times\Bigg(\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}\Bigg)^{(2\beta-1)/(2\beta+1)}\times\\
&\times\frac{d^{(2\beta-1)/(2\beta+1)}}{n^{2\beta/2\beta+1}}+\Bigg(4C^2(1+LR)^2+\\
&+CR(1+LR)\Bigg)\frac{A\log N}{n}, \qquad (3.4)
\end{aligned}$$

which implies the result. ∎

4 Appendix

The Rademacher process is defined as

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

where $\{\varepsilon_j\}$ are i.i.d. Rademacher random variables independent of $\{X_j\}$.

We will need several bounds for Rademacher processes indexed by functions from RKHS (some of them are well known; see, e.g., Mendelson (2002) and Blanchard, Bousquet and Massart (2007)). We state them without proofs for brevity.

First we consider a single RKHS $\mathcal{H}_K$ where K is a kernel with eingenvalues λ_k and eigenfunctions ϕ_k (in $L_2(\Pi)$).

Lemma 3 *The following bound holds:*

$$\mathbb{E}\sup_{\|h\|_{\mathcal{H}_K}\le 1}|R_n(h)|\le\sqrt{\frac{\sum_{k=1}^{\infty}\lambda_k}{n}}.$$

Let $m\ge 1$. Denote by L the linear span of the functions $\{\phi_k : k=1,\dots,m\}$ and by $L^{\perp}$ the closed linear span (in $L_2(\Pi)$) of the functions $\{\phi_k : k\ge m+1\}$. $P_L, P_{L^{\perp}}$ will denote orthogonal projectors in $L_2(\Pi)$ on the corresponding subspaces.

Lemma 4 *For all $m\ge 1$,*

$$\mathbb{E}\sup_{\|h\|_{\mathcal{H}_K}\le 1}|R_n(P_{L^{\perp}}h)|\le\sqrt{\frac{\sum_{k=m+1}^{\infty}\lambda_k}{n}}.$$

We now turn to the case of a dictionary $\{\mathcal{H}_j : j=1,\dots,N\}$ of RKHS with kernels $\{K_j : j=1,\dots N\}$. As before, denote $\{\lambda_k^{(j)} : k\ge 1\}$ the eigenvalues (arranged in decreasing order) and $\{\phi_k^{(j)} : k\ge 1\}$ the $L_2(\Pi)$-orthonormal eigenfunctions of K_j. The following bounds will be needed in this case.

Lemma 5 *With some numerical constant C,*

$$\mathbb{E}\max_{1\le j\le N}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|\le\sqrt{\frac{\max_{1\le j\le N}\sum_{k=1}^{\infty}\lambda_k^{(j)}}{n}}+C\sqrt{\frac{\log N}{n}}.$$

Proof. We use bounded difference inequality to get for each $j=1,\dots,N$ with probability at least $1-e^{-t-\log N}$

$$\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|\le\mathbb{E}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|+\frac{C\sqrt{t+\log N}}{\sqrt{n}}.$$

By the union bound, this yields with probability at least $1-Ne^{-t-\log N}=1-e^{-t}$

$$\max_{1\le j\le N}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|\le\max_{1\le j\le N}\mathbb{E}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|+\frac{C\sqrt{t}}{\sqrt{n}}+\frac{C\sqrt{\log N}}{\sqrt{n}},$$

which holds for all $t>0$ and implies that

$$\mathbb{E}\max_{1\le j\le N}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|\le\max_{1\le j\le N}\mathbb{E}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|+\frac{C\sqrt{\log N}}{\sqrt{n}}$$

with a properly chosen constant $C>0$. Note that, by Lemma 3,

$$\mathbb{E}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)|\le\sqrt{\frac{\sum_{k=1}^{\infty}\lambda_k^{(j)}}{n}},$$

which implies the result. ∎

As before, denote $L_j, L_j^{\perp}$ the subspaces of $L_2(\Pi)$ spanned on $\{\phi_k^{(j)} : k\le m\}$ and $\{\phi_k^{(j)} : k>m\}$, respectively, $P_{L_j}, P_{L_j}^{\perp}$ being the corresponding orthogonal projections. Recall that sequence $\{\lambda_k^j\}$ is nonincreasing. The following statement is a uniform version of Lemma 4.

Lemma 6 *With some numerical constant C,*

$$\begin{aligned}&\mathbb{E}\max_{1\le j\le N}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(P_{L_j^{\perp}}h_j)|\\ \le\;& 2\sqrt{\frac{\max_{1\le j\le N}\sum_{k=m+1}^{\infty}\lambda_k^{(j)}}{n}}\\ &+2\sqrt{\max_{1\le j\le N}\lambda_m^{(j)}}\sqrt{\frac{\log N+C}{n}}+2\frac{\log N+C}{n}.\end{aligned}$$

Lemma 7 *The following bound holds:*

$$\begin{aligned}&\mathbb{E}\sup\Big\{|R_n(g-f)| : \|g-f\|_{L_2(\Pi)}\le\delta, \|g\|_{\ell_1}\le R,\\ &\sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta\Big\}\le C\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\delta\sqrt{\frac{dm}{n}}\\ &+2R\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}}+CR\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}}\\ &+C\Delta\sqrt{\frac{\log(N-d)+1}{n}}.\end{aligned}$$

Proof. First note that

$$\begin{aligned}&\mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j=1}^{N}\Big(g_j-f_j)\Big)\Big| : \|g-f\|_{L_2(\Pi)}\le\delta, \|g\|_{\ell_1}\le R,\\ &\sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta\Big\}\le\mathbb{E}\sup\Big\{\Big|R_n\Big(\Big|\sum_{j\in J_f}(g_j-f_j)\Big)\Big| :\\ &\|g-f\|_{L_2(\Pi)}\le\delta, \|g\|_{\ell_1}\le R, \sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta\Big\}+\\ &\mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\notin J_f}(g_j-f_j)\Big)\Big| : \|g-f\|_{L_2(\Pi)}\le\delta, \|g\|_{\ell_1}\le R,\\ &\sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta\Big\}.\end{aligned}$$

The second term can be bounded as follows:

$$\mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\notin J_f} g_j\Big)\Big| : \|g-f\|_{L_2(\Pi)}\le\delta, \|g\|_{\ell_1}\le R, \sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta\Big\} \le \mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j} h_j\Big)\Big| : \sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta, \|h_j\|_{\mathcal{H}_j}\le 1\Big\} \le \Delta\mathbb{E}\max_{j\notin J_f}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(h_j)| \le C\Delta\sqrt{\frac{\log(N-d)+1}{n}},$$

where we used Lemma 5. As to the first term, we use the bound

$$\mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\in J_f}(g_j-f_j)\Big)\Big| : \|g-f\|_{L_2(\Pi)}\le\delta, \|g\|_{\ell_1}\le R, \sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta\Big\} \le \mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\in J_f} P_{L_j}(g_j-f_j)\Big)\Big| : \|g-f\|_{L_2(\Pi)}\le\delta\Big\} + \mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\in J_f} P_{L_j^\perp}(g_j-f_j)\Big)\Big| : \|g\|_{\ell_1}\le R\Big\}.$$

Note that

$$\Big\|\sum_{j\in J_f} P_{L_j}(g_j-f_j)\Big\|^2_{L_2(\Pi)} \le \bar\kappa(J_f)\sum_{j\in J_f}\Big\|P_{L_j}(g_j-f_j)\Big\|^2_{L_2(\Pi)} \le \bar\kappa(J_f)\sum_{j\in J_f}\|g_j-f_j\|^2_{L_2(\Pi)} \le \frac{\bar\kappa(J_f)}{\kappa(J_f)}\Big\|\sum_{j\in J_f}(g_j-f_j)\Big\|^2_{L_2(\Pi)} \le \frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}\Big\|\sum_{j=1}^n(g_j-f_j)\Big\|^2_{L_2(\Pi)} \le \frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}\delta^2.$$

Also, $\sum_{j\in J_j} P_{L_j}(g_j-f_j)$ takes values in the linear span of $\bigcup_{j\in J_f} L_j$ whose dimension $\le dm$. This yields the following bound

$$\mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\in J_f} P_{L_j}(g_j-f_j)\Big)\Big| : \|g-f\|_{L_2(\Pi)}\le\delta\Big\} \le C\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\delta\sqrt{\frac{dm}{n}}.$$

Finally, we use Lemma 6 to get

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\in J_f} P_{L_j^\perp}(g_j-f_j)\Big)\Big| : \|g\|_{\ell_1}\le R\Big\}\\
\le\ &\mathbb{E}\sup\Big\{\Big|R_n\Big(\sum_{j\in J_f}\|g_j-f_j\|_{\mathcal{H}_j} P_{L_j^\perp} h_j\Big)\Big| : \|g\|_{\ell_1}\le R, \|h_j\|_{\mathcal{H}_j}\le 1, j=1,\dots,N\Big\}\\
\le\ &2R\mathbb{E}\max_{j\in J_f}\sup_{\|h_j\|_{\mathcal{H}_j}\le 1}|R_n(P_{L_j^\perp}h_j)|\\
\le\ &2R\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}} + C\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}}.
\end{aligned}$$

Combining the above bounds we get

$$\mathbb{E}\sup\Big\{|R_n(g-f)| : \|g-f\|_{L_2(\Pi)}\le\delta, \|g\|_{\ell_1}\le R, \sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta\Big\} \le C\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\delta\sqrt{\frac{dm}{n}} + 2R\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}} + CR\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}} + C\Delta\sqrt{\frac{\log(N-d)+1}{n}}. \ \blacksquare$$

Recall that

$$\alpha_n(\delta,\Delta,R) := \sup\Big\{|(P_n-P)(\ell\bullet g-\ell\bullet f)| : g\in\mathcal{G}(\delta,\Delta,R)\Big\},$$

where

$$\mathcal{G}(\delta,\Delta,R) := \Big\{g : \|g-f\|_{L_2(\Pi)}\le\delta, \sum_{j\notin J_f}\|g_j\|_{\mathcal{H}_j}\le\Delta, \|g\|_{\ell_1}\le R\Big\}.$$

We will assume that $R\le e^N$ (recall also the assumption $N\ge n^\gamma$).

Lemma 8 *There exist constants C, L depending only on the loss ℓ ($L=0$ if ℓ' is bounded) such that for all*

$$n^{-1/2}\le\delta\le 2R, \quad n^{-1/2}\le\Delta\le R \tag{4.1}$$

and for all $A\ge 1$ the following bound holds with probability at least $1-N^{-A}$:

$$\alpha_n(\delta,\Delta,R) \le C(1+LR)\Big[\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\times \delta\sqrt{\frac{dm}{n}} + R\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}} + R\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}} + \Delta\sqrt{\frac{\log(N-d)+1}{n}}\Big] + C(1+LR)\delta\sqrt{\frac{A\log N}{n}} + CR(1+LR)\frac{A\log N}{n}. \tag{4.2}$$

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

$$\alpha_n(\delta;\Delta;R) \leq$$
$$2\left[\mathbb{E}\alpha_n(\delta;\Delta,R) + C(1+LR)\delta\sqrt{\frac{t}{n}} + \frac{CR(1+LR)t}{n}\right].$$

To apply Talagrand's inequality we used the assumptions on the loss function. It follows from these assumptions that for all $g \in \mathcal{G}(\delta,\Delta,R)$

$$\|\ell\bullet g - \ell\bullet f\|_{L_2(\Pi)} \leq C(1+LR)\|g-f\|_{L_2(\Pi)} \leq C(1+LR)\delta$$

and also

$$\|\ell \bullet g - \ell \bullet f\|_\infty \leq CR(1+LR).$$

Next, by symmetrization inequality,

$$\mathbb{E}\alpha_n(\delta;\Delta,R) \leq 2\mathbb{E}\sup\Big\{|R_n((\ell\bullet g - \ell\bullet f : g \in \mathcal{G}(\delta,\Delta,R).\Big\}.$$

We write $u = g - f$ and

$$\ell \bullet g - \ell \bullet f = \ell \bullet (f+u) - \ell \bullet f$$

and observe that the function

$$[-R,R] \ni u \mapsto \ell \bullet (f+u) - \ell \bullet f$$

is Lipschitz with constant $C(1+LR)$. This allows us to use Rademacher contraction inequality (Ledoux and Talagrand, 1991) to get

$$\mathbb{E}\alpha_n(\delta;\Delta,R) \leq C(1+LR) \times$$
$$\times \mathbb{E}\sup\Big\{\Big|R_n(g-f)\Big| : g \in \mathcal{G}(\delta,\Delta,R)\Big\}.$$

The last expectation can be further bounded by Lemma 7. As a result, we get the following bound that holds with probability at least $1 - e^{-t}$:

$$\alpha_n(\delta;\Delta,R) \leq C(1+LR)\Bigg[\sqrt{\frac{\bar\kappa(J_f)}{\kappa(J_f)(1-\rho^2(J_f))}}\delta\sqrt{\frac{dm}{n}}$$
$$+R\sqrt{\frac{\max_{j\in J_f}\sum_{k>m}\lambda_k^{(j)}}{n}} + R\sqrt{\max_{j\in J_f}\lambda_m^{(j)}}\sqrt{\frac{\log d}{n}} +$$
$$\Delta\sqrt{\frac{\log(N-d)+1}{n}}\Bigg] + C(1+LR)\delta\sqrt{\frac{t}{n}}$$
$$+\frac{CR(1+LR)t}{n} =: \tilde\beta_n(\delta,\Delta,R;t). \quad (4.3)$$

The next goal is to make the bound uniform in

$$n^{-1/2} \leq \delta \leq 2R \ \text{ and } \ n^{-1/2} \leq \Delta \leq R. \qquad (4.4)$$

To this end, consider

$$\delta_j := 2R2^{-j}, \quad \Delta_j := R2^{-j}.$$

We will replace t by $t + 2\log\log(2R\sqrt{n})$ and use bound (4.3) for all $\delta = \delta_j$ and $\Delta = \Delta_k$ satisfying the conditions (4.4). By the union bound, with probability at least

$$1 - \log(R\sqrt{n})\log(2R\sqrt{n})\exp\Big\{-t - 2\log\log(2R\sqrt{n})\Big\}$$
$$\geq 1 - e^{-t},$$

the following bound holds for all δ_j, Δ_k satisfying (4.4):

$$\alpha_n(\delta_j,\Delta_k,R) \leq \tilde\beta_n\left(\delta_j,\Delta_k,R;t+2\log\log\left(\frac{2R}{\sqrt{n}}\right)\right).$$

It is enough now to substitute in the above bound $t := A\log N$ and to use the fact that the functions $\alpha_n(\delta,\Delta,R)$ and $\tilde\beta_n(\delta,\Delta,R;t)$ are nondecreasing with respect to δ and Δ. Together with the conditions $R \leq e^N$ and $N \geq n^\gamma$, this implies the claim. ∎

References

[1] Bousquet, O. and Herrmann, D. (2003), On the complexity of learning the kernel matrix, In: *Advances in Neural Information Processing Systems 15*.

[2] Blanchard, G., Bousquet, O. and Massart, P. (2008), Statistical performance of support vector machines, *Annals of Statistics*, **36**, 489-531.

[3] Crammer, K., Keshet, J. and Singer, Y. (2003), Kernel design using boosting, In: *Advances in Neural Information Processing Systems 15*.

[4] Koltchinskii, V. (2008), Sparsity in penalized empirical risk minimization, *Ann. Inst. H. Poincaré*, to appear.

[5] Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L. and Jordan, M. (2004), Learning the kernel matrix with semidefinite programming, *Journal of Machine Learning Research*, **5**, 27-72.

[6] Ledoux, M. and Talagrand, M. (1991), *Probability in Banach Spaces*, Springer, New York.

[7] Lin, Y. and Zhang, H. (2006), Component selecion and smoothing in multivariate nonparametric regression, *Annals of Statistics*, **34**, 2272-2297.

[8] Micchelli, C. and Pontil, M. (2005), Learning the kernel function via regularization, *Journal of Machine Learning Research*, **6**, 1099-1125.

[9] Mendelson, S. (2002) Geometric parameters of kernel machines, In: *COLT 2002*, Lecture Notes in Artificial Intelligence, 2375, Springer, 29–43.

[10] Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2007), SpAM: sparse additive models, to appear in: *Advances in Neural Information Processing Systems* (NIPS 07).

[11] Srebro, N. and Ben-David, S. (2006), Learning bounds for support vector machines with learned kernels, In: *Proceedings of 19th Annual Conference on Learning Theory* (COLT 2006), 169-183.

[12] Wahba, G (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

[13] Yuan, M. (2007), Nonnegative garrote component selection in functional ANOVA models, in *Proceedings of AI and Statistics* (AISTAT 07), 656-662.