Sparsity in Multiple Kernel Learning

Vladimir Koltchinskii[‡]
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160 USA
vlad@math.gatech.edu
and
Ming Yuan[§]
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205 USA
myuan@isye.gatech.edu

(April 28, 2010)

‡ The research of this author was supported in part by NSF grants MPSA-MCS-0624841, DMS-0906880 and CCF-0808863

§ The research of this author was supported in part by NSF grants MPSA-MCS-0624841 and DMS-0846234

Abstract

The problem of multiple kernel learning based on penalized empirical risk minimization is discussed. The complexity penalty is determined jointly by the empirical L_2 norms and the reproducing kernel Hilbert space (RKHS) norms induced by the kernels with a data-driven choice of regularization parameters. The main focus is on the case when the total number of kernels is large, but only a relatively small number of them is needed to represent the target function, so that the problem is sparse. The goal is to establish oracle inequalities for the excess risk of the resulting prediction rule showing that the method is adaptive both to the unknown design distribution and to the sparsity of the problem.

1 Introduction

Let $(X_i, Y_i), i = 1, \dots, n$ be independent copies of a random couple (X, Y) with values in $S \times T$, where S is a measurable space with σ-algebra $\mathcal{A}$ (typically, S is a compact subset of a finite-dimensional Euclidean space) and T is a Borel subset of $\mathbb{R}$. In what follows, P will denote the distribution of (X, Y) and Π the distribution of X. The corresponding empirical distributions, based on $(X_1, Y_1), \dots (X_n, Y_n)$ and on $(X_1, \dots, X_n)$, will be denoted by P_n and Π_n, respectively. For a measurable function $g : S \times T \mapsto \mathbb{R}$, we denote

$$Pg := \int_{S\times T} g dP = \mathbb{E}g(X, Y) \quad \text{and} \quad P_n g := \int_{S\times T} g dP_n = n^{-1} \sum_{j=1}^{n} g(X_j, Y_j).$$

Similarly, we use the notations Πf and $\Pi_n f$ for the integrals of a function $f : S \mapsto \mathbb{R}$ with respect to the measures Π and Π_n.

The goal of prediction is to learn "a reasonably good" prediction rule $f : S \to \mathbb{R}$ from the empirical data $\{(X_i, Y_i) : i = 1, 2, \dots, n\}$. To be more specific, consider a loss function $\ell : T \times \mathbb{R} \to \mathbb{R}_+$ and define the risk of a prediction rule f as

$$P(\ell \circ f) = \mathbb{E}\ell(Y, f(X)),$$

where $(\ell \circ f)(x, y) = \ell(y, f(x))$. An optimal prediction rule with respect to this loss is defined as

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

where the minimization is taken over all measurable functions and, for simplicity, it is assumed that the minimum is attained. **The excess risk** of a prediction rule f is defined as

$$\mathcal{E}(\ell \circ f) := P(\ell \circ f) - P(\ell \circ f_*).$$

Throughout the paper, the notation $a \asymp b$ means that there exists a numerical constant $c > 0$ such that $c^{-1} \leq \frac{a}{b} \leq c$. By "numerical constants" we usually mean real numbers whose precise values are not necessarily specified, or, sometimes, constants that might depend on the characteristics of the problem that are of little interest to us (for instance, some constants that depend only on the loss function).

1.1 Learning in Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}_K$ be a reproducing kernel Hilbert space (RKHS) associated with a symmetric nonnegatively definite kernel $K : S \times S \to \mathbb{R}$ such that for any $x \in S$, $K_x(\cdot) := K(\cdot, x) \in \mathcal{H}_K$ and $f(x) = \langle f, K_x \rangle_{\mathcal{H}_K}$ for all $f \in \mathcal{H}_K$ (Aronszajn (1950)). If it is known that if $f_* \in \mathcal{H}_K$ and $\|f_*\|_{\mathcal{H}_K} \leq 1$, then it is natural to estimate f_* by a solution $\hat{f}$ of the following empirical risk minimization problem:

$$\hat{f} := \operatorname{argmin}_{\|f\|_{\mathcal{H}_K} \leq 1} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)). \tag{1}$$

The size of the excess risk $\mathcal{E}(\ell \circ \hat{f})$ of such an empirical solution depends on the "smoothness" of functions in the RKHS $\mathcal{H}_K$. A natural notion of "smoothness" in this context is related to the unknown design distribution Π. Namely, let T_K be the integral operator from $L_2(\Pi)$ into $L_2(\Pi)$ with kernel K. Under a standard assumption that the kernel K is square integrable (in the theory of RKHS it is usually even assumed that S is compact and K is continuous), the operator T_K is compact and its spectrum is discrete. If $\{\lambda_k\}$ is the sequence of the eigenvalues (arranged in decreasing order) of T_K and $\{\phi_k\}$ is the corresponding $L_2(\Pi)$-orthonormal sequence of eigenfunctions, then it is well known that the RKHS-norms of functions from the linear span of $\{\phi_k\}$ can be written as

$$\|f\|_{\mathcal{H}_K}^2 = \sum_{k \geq 1} \frac{|\langle f, \phi_k \rangle_{L_2(\Pi)}|^2}{\lambda_k},$$

which means that the "smoothness" of functions in $\mathcal{H}_K$ depends on the rate of decay of eigenvalues λ_k that, in turn, depends on the design distribution Π. It is also clear that the unit balls in the RKHS $\mathcal{H}_K$ are ellipsoids in the space $L_2(\Pi)$ with "axes" $\sqrt{\lambda_k}$.

It was shown by Mendelson (2002) that the following function

$$\breve{\gamma}_n(\delta) := \left(n^{-1}\sum_{k\geq 1}(\lambda_k \wedge \delta^2)\right)^{1/2}, \qquad \delta \in [0,1]$$

provides tight upper and lower bounds (up to constants) on localized Rademacher complexities of the unit ball in $\mathcal{H}_K$ and plays an important role in the analysis of the empirical risk minimization problem (1). It is easy to see that the function $\breve{\gamma}_n^2(\sqrt{\delta})$ is concave, $\breve{\gamma}_n(0) = 0$ and, as a consequence, $\breve{\gamma}_n(\delta)/\delta$ is a decreasing function of δ and $\breve{\gamma}_n(\delta)/\delta^2$ is strictly decreasing. Hence, there exists unique positive solution of the equation $\breve{\gamma}_n(\delta) = \delta^2$. If $\bar{\delta}_n$ denotes this solution, then the results of Mendelson (2002) imply that with some constant $C > 0$ and with probability at least $1 - e^{-t}$

$$\mathcal{E}(\ell \circ \hat{f}) \leq C\left(\bar{\delta}_n^2 + \frac{t}{n}\right).$$

The size of the quantity $\bar{\delta}_n^2$ involved in this upper bound on the excess risk depends on the rate of decay of the eigenvalues λ_k as $k \to \infty$. In particular, if $\lambda_k \asymp k^{-2\beta}$ for some $\beta > 1/2$, then it is easy to see that $\breve{\gamma}_n(\delta) \asymp n^{-1/2}\delta^{1-\frac{1}{2\beta}}$ and $\bar{\delta}_n^2 \asymp n^{-2\beta/(2\beta+1)}$. Recall that unit balls in $\mathcal{H}_K$ are ellipsoids in $L_2(\Pi)$ with "axes" of the order $k^{-\beta}$ and it is well known that, in a variety of estimation problems, $n^{-2\beta/(2\beta+1)}$ represents minimax convergence rates of the squared L_2-risk for functions from such ellipsoids (for instance, from Sobolev balls of smoothness β), as in famous Pinsker's Theorem (see, e.g., Tsybakov (2009), Chapter 3).

Example. Sobolev spaces $W^{\alpha,2}(G),\ G \subset \mathbb{R}^d$ of smoothness $\alpha > d/2$ is a well known class of concrete examples of RKHS. Let $\mathbb{T}^d, d \geq 1$ denote the d-dimensional torus and let Π be the uniform distribution in $\mathbb{T}^d$. It is easy to check that, for all $\alpha > d/2$, the Sobolev space $W^{\alpha,2}(\mathbb{T}^d)$ is an RKHS generated by the kernel $K(x,y) = k(x-y),\ x,y \in \mathbb{T}$, where the function $k \in L_2(\mathbb{T}^d)$ is defined by its Fourier coefficients

$$\hat{k}_n = (|n|^2+1)^{-\alpha},\ n = (n_1,\dots,n_d) \in \mathbb{Z}^d,\ |n|^2 := n_1^2 + \cdots + n_d^2.$$

In this case, the eigenfunctions of the operator T_K are the functions of the Fourier basis and its eigenvalues are the numbers $\{(|n|^2+1)^{-\alpha} : n \in \mathbb{Z}^d\}$. For $d = 1$ and $\alpha > 1/2$, we have

$\lambda_k \asymp k^{-2\alpha}$ (recall that $\{\lambda_k\}$ are the eigenvalues arranged in decreasing order) so, $\beta = \alpha$ and $\bar{\delta}_n^2 \asymp n^{-2\alpha/(2\alpha+1)}$, which is a minimax nonparametric convergence rate for Sobolev balls in $W^{\alpha,2}(\mathbb{T})$ (see, e.g., Tsybakov (2009), Theorem 2.9). More generally, for arbitrary $d \geq 1$ and $\alpha > d/2$, we get $\beta = \alpha/d$ and $\bar{\delta}_n^2 \asymp n^{-2\alpha/(2\alpha+d)}$, which is also a minimax optimal convergence rate in this case. Suppose now that the distribution Π is uniform in a torus $\mathbb{T}^{d'} \subset \mathbb{T}^d$ of dimension $d' < d$. We will use the same kernel K, but restrict the RKHS $\mathcal{H}_K$ to the torus $\mathbb{T}^{d'}$ of smaller dimension. Let $d'' = d - d'$. For $n \in \mathbb{Z}^d$, we will write $n = (n', n'')$ with $n' \in \mathbb{Z}^{d'}, n'' \in \mathbb{Z}^{d''}$. It is easy to prove that the eigenvalues of the operator T_K become in this case

$$\sum_{n'' \in \mathbb{Z}^{d''}} (|n'|^2 + |n''|^2 + 1)^{-\alpha} \asymp (|n'|^2 + 1)^{-(\alpha - d''/2)}.$$

Due to this fact, the norm of the space $\mathcal{H}_K$ (restricted to $\mathbb{T}^{d'}$) is equivalent to the norm of the Sobolev space $W^{\alpha - d''/2,2}(\mathbb{T}^{d'})$. Since the eigenvalues of the operator T_K coincide, up to a constant, with the numbers $\{(|n'|^2 + 1)^{-(\alpha - d''/2)} : n' \in \mathbb{Z}^{d'}\}$, we get $\bar{\delta}_n^2 \asymp n^{-\frac{2\alpha - d''}{2\alpha - d'' + d'}}$ (which is again the minimax convergence rate for Sobolev balls in $W^{\alpha - d''/2,2}(\mathbb{T}^{d'})$). In the case of more general design distributions Π, the rate of decay of the eigenvalues λ_k and the corresponding size of the excess risk bound $\bar{\delta}_n^2$ depends on Π. If, for instance, Π is supported in a submanifold $S \subset \mathbb{T}^d$ of dimension $\dim(S) < d$, the rate of convergence of $\bar{\delta}_n^2$ to 0 depends on the dimension of the submanifold S rather than on the dimension of the ambient space $\mathbb{T}^d$.

Using the properties of the function $\breve{\gamma}_n$, in particular, the fact that $\breve{\gamma}_n(\delta)/\delta$ is decreasing, it is easy to observe that $\breve{\gamma}_n(\delta) \leq \bar{\delta}_n \delta + \bar{\delta}_n^2,\ \delta \in (0, 1]$. Moreover, if $\breve{\epsilon} = \breve{\epsilon}(K)$ denotes the smallest value of ϵ such that the linear function $\epsilon\delta + \epsilon^2,\ \delta \in (0, 1]$ provides an upper bound for the function $\breve{\gamma}_n(\delta),\ \delta \in (0, 1]$, then $\breve{\epsilon} \leq \bar{\delta}_n \leq 2(\sqrt{5} - 1)^{-1}\breve{\epsilon}$. Note that $\breve{\epsilon}$ also depends on n, but we do not have to emphasize this dependence in the notations since, in what follows, n is fixed. Based on the observations above, the quantity $\bar{\delta}_n$ coincides (up to a numerical constant) with the slope $\breve{\epsilon}$ of the "smallest linear majorant" of the form $\epsilon\delta + \epsilon^2$ of the function $\breve{\gamma}_n(\delta)$. This interpretation of $\bar{\delta}_n$ is of some importance in the design of complexity penalties used in this paper.

1.2 Sparse Recovery via Regularization

Instead of minimizing the empirical risk over an RKHS-ball (as in problem (1)), it is very common to define the estimator $\hat{f}$ of the target function f_* as a solution of the penalized empirical risk minimization problem of the form

$$\hat{f} := \operatorname{argmin}_{f\in\mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \epsilon \|f\|^{\alpha}_{\mathcal{H}_K} \right], \tag{2}$$

where $\epsilon > 0$ is a tuning parameter that balances the tradeoff between the empirical risk and the "smoothness" of the estimate and, most often, $\alpha = 2$ (sometimes, $\alpha = 1$). The properties of the estimator $\hat{f}$ has been studied extensively. In particular, it was possible to derive probabilistic bounds on the excess risk $\mathcal{E}(\ell \circ \hat{f})$ (oracle inequalities) with the control of the random error in terms of the rate of decay of the eigenvalues $\{\lambda_k\}$, or, equivalently, in terms of the function $\breve{\gamma}_n$ (see, e.g., Blanchard, Bousquet and Massart (2008)).

In the recent years, there has been a lot of interest in a data dependent choice of kernel K in this type of problems. In particular, given a finite (possibly large) dictionary $\{K_j : j = 1, 2, \ldots, N\}$ of symmetric nonnegatively definite kernels on S, one can try to find a "good" kernel K as a convex combination of the kernels from the dictionary:

$$K \in \mathcal{K} := \left\{ \sum_{j=1}^{N} \theta_j K_j : \theta_j \geq 0, \theta_1 + \ldots + \theta_N = 1 \right\}. \tag{3}$$

The coefficients of K need to be estimated from the training data along with the prediction rule. Using this approach for problem (2) with $\alpha = 1$ leads to the following optimization problem:

$$\hat{f} := \operatorname{argmin}_{\substack{f\in\mathcal{H}_K \\ K\in\mathcal{K}}} \left(P_n(\ell \circ f) + \epsilon \|f\|_{\mathcal{H}_K} \right). \tag{4}$$

This learning problem, often referred to as the multiple kernel learning, has been studied recently by Bousquet and Herrmann (2003), Cramer, Keshet and Singer (2003), Lanckriet, Cristianini, Bartlett, Ghaoui and Jordan (2004), Micchelli and Pontil (2005), Lin and Zhang (2006), Srebro and Ben-David (2006), Bach (2008) and Koltchinskii and Yuan (2008) among others. In particular, (see, e.g., Micchelli and Pontil (2005)), problem (4) is equivalent to the following:

$$(\hat{f}_1, \ldots, \hat{f}_N) := \operatorname{argmin}_{f_j\in\mathcal{H}_{K_j},\ j=1,\ldots,N} \left(P_n(\ell \circ (f_1 + \cdots + f_N)) + \epsilon \sum_{j=1}^{N} \|f_j\|_{\mathcal{H}_{K_j}} \right), \tag{5}$$

which is an infinite-dimensional version of LASSO-type penalization. Koltchinskii and Yuan (2008) studied this method in the case when the dictionary is large, but the target function f_* has a "sparse representation" in terms of a relatively small subset of kernels $\{K_j : j \in J\}$. It was shown that this method is adaptive to sparsity extending well known properties of LASSO to this infinite dimensional framework.

In this paper, we study a different approach to the multiple kernel learning. It is closer to the recent work on **"sparse additive models"** (see, e.g., Ravikumar, Liu, Lafferty and Wasserman (2008) and Meier, van de Geer and Bühlmann (2009)) and it is based on a "double penalization" with a combination of empirical L_2-norms (used to enforce the sparsity of the solution) and RKHS-norms (used to enforce the "smoothness" of the components). Moreover, we suggest a data-driven method of choosing the values of regularization parameters that is adaptive to unknown smoothness of the components (determined by the behavior of distribution dependent eigenvalues of the kernels).

Let $\mathcal{H}_j := \mathcal{H}_{K_j},\ j = 1, \dots, N$. Denote $\mathcal{H} := \text{l.s.}\left(\bigcup_{j=1}^N \mathcal{H}_j\right)$ ("l.s." meaning "the linear span"), and

$$\mathcal{H}^{(N)} := \left\{(h_1, \dots, h_N) : h_j \in \mathcal{H}_j,\ j = 1, \dots, N\right\}.$$

Note that $f \in \mathcal{H}$ if and only if there exists an additive representation (possibly, non-unique) $f = f_1 + \cdots + f_N$, where $f_j \in \mathcal{H}_j,\ j = 1, \dots, N$. Also, $\mathcal{H}^{(N)}$ has a natural structure of a linear space and it can be equipped with the following inner product

$$\langle (f_1, \dots, f_N), (g_1, \dots, g_N)\rangle_{\mathcal{H}^{(N)}} := \sum_{j=1}^N \langle f_j, g_j\rangle_{\mathcal{H}_j}$$

to become *the direct sum* of Hilbert spaces $\mathcal{H}_j,\ j = 1, \dots, N$.

Given a convex subset $D \subset \mathcal{H}^{(N)}$, consider the following penalized empirical risk minimization problem:

$$\left(\hat{f}_1, \dots, \hat{f}_N\right) = \underset{(f_1,\dots,f_N)\in D}{\operatorname{argmin}} \left[P_n(\ell \circ (f_1 + \cdots + f_N)) + \sum_{j=1}^N \left(\epsilon_j \|f_j\|_{L_2(\Pi_n)} + \epsilon_j^2 \|f_j\|_{\mathcal{H}_j}\right)\right]. \quad (6)$$

Note that for special choices of set D, for instance, for $D := \{(f_1, \dots, f_N) : f_j \in \mathcal{H}_j, \|f_j\|_{\mathcal{H}_j} \le R_j\}$ for some $R_j > 0,\ j = 1, \dots, N$, one can replace each component f_j involved in the optimization problem by its orthogonal projections in $\mathcal{H}_j$ onto the linear span of the functions $\{K_j(\cdot, X_i), i = 1, \dots, n\}$ and reduce the problem to a convex optimization over a finite dimensional space (of dimension nN).

The complexity penalty in the problem (6) is based on two norms of the components f_j of an additive representation: the empirical L_2-norm, $\|f_j\|_{L_2(\Pi_n)}$, with regularization parameter ϵ_j, and an RKHS-norm, $\|f_j\|_{\mathcal{H}_j}$, with regularization parameter ϵ_j^2. The empirical L_2-norm (the lighter norm) is used to enforce the sparsity of the solution whereas the RKHS norms (the heavier norms) are used to enforce the "smoothness" of the components. This is similar to the approach taken in Meier, van de Geer and Bühlmann (2009) in the context of classical additive models, i.e., in the case when $S := [0,1]^N$, $\mathcal{H}_j := W^{\alpha,2}([0,1])$ for some smoothness $\alpha > 1/2$ and the space $\mathcal{H}_j$ is a space of functions depending on the j-th variable. In this case, the regularization parameters ϵ_j are equal (up to a constant) to $n^{-\alpha/(2\alpha+1)}$. The quantity ϵ_j^2, used in the "smoothness part" of the penalty, coincides with the minimax convergence rate in a one component smooth problem. At the same time, the quantity ϵ_j, used in the "sparsity part" of the penalty, is equal to the square root of the minimax rate (which is similar to the choice of regularization parameter in standard sparse recovery methods such as LASSO). This choice of regularization parameters results in the excess risk of the order $dn^{-2\alpha/(2\alpha+1)}$, where d is the number of components of the target function (the degree of sparsity of the problem).

The framework of multiple kernel learning considered in this paper includes many generalized versions of classical additive models. For instance, one can think of the case when $S := [0,1]^{m_1} \times \cdots \times [0,1]^{m_N}$ and $\mathcal{H}_j = W^{\alpha,2}([0,1]^{m_j})$ is a space of functions depending on the j-th block of variables. In this case, a proper choice of regularization parameters (for uniform design distribution) would be $\epsilon_j = n^{-\alpha/(2\alpha+m_j)}$, $j = 1, \dots, N$ (so, these parameters and the error rates for different components of the model are different). It should be also clear from the discussion in Section 1.1 that, if the design distribution Π is unknown, the minimax convergence rates for the one component problems are also unknown. For instance, if the projections of design points on the cubes $[0,1]^{m_j}$ are distributed in lower dimensional submanifolds of these cubes, then the unknown dimensions of the submanifolds rather than the dimensions m_j would be involved in the minimax rates and in the regularization parameters ϵ_j. Because of this, data driven choice of regularization parameters ϵ_j that provides adaptation to the unknown design distribution Π and to the unknown "smoothness" of the components (related to this distribution) is a major issue in multiple kernel learning. From this point of view, even in the case of classical additive models, the choice of regularization

parameters that is based only on Sobolev type smoothness and ignores the design distribution is not adaptive. Note that, in the infinite dimensional LASSO studied in Koltchinskii and Yuan (2008), the regularization parameter ϵ is chosen the same way as in the classical LASSO ($\epsilon \asymp \sqrt{\frac{\log N}{n}}$), so, it is not related to the smoothness of the components. However, the oracle inequalities proved in Koltchinskii and Yuan (2008) give correct size of the excess risk only for special choices of kernels that depend on unknown "smoothness" of the components of the target function f_*, so, this method is not adaptive either.

1.3 Adaptive Choice of Regularization Parameters

Denote

$$\hat{K}_j := \left(\frac{K_j(X_l, X_k)}{n}\right)_{l,k=1,n}.$$

This $n \times n$ Gram matrix can be viewed as an empirical version of the integral operator T_{K_j} from $L_2(\Pi)$ into $L_2(\Pi)$ with kernel K_j. Denote $\hat{\lambda}_k^{(j)}, k = 1, 2, \dots$ the eigenvalues of $\hat{K}_j$ arranged in decreasing order. We also use the notation $\lambda_k^{(j)},\ k = 1, 2, \dots$ for the eigenvalues of the operator $T_{K_j} : L_2(\Pi) \mapsto L_2(\Pi)$ with kernel K_j arranged in decreasing order. Define functions $\breve{\gamma}_n^{(j)},\ \hat{\gamma}_n^{(j)}$,

$$\breve{\gamma}_n^{(j)}(\delta) := \left(\frac{1}{n}\sum_{k=1}^{n}(\lambda_k^{(j)} \wedge \delta^2)\right)^{1/2} \quad \text{and} \quad \hat{\gamma}_n^{(j)}(\delta) := \left(\frac{1}{n}\sum_{k=1}^{n}(\hat{\lambda}_k^{(j)} \wedge \delta^2)\right)^{1/2},$$

and, for a fixed given $A \geq 1$, let

$$\hat{\epsilon}_j := \inf\left\{\epsilon \geq \sqrt{\frac{A\log N}{n}} : \hat{\gamma}_n^{(j)}(\delta) \leq \epsilon\delta + \epsilon^2,\ \forall \delta \in (0,1]\right\}. \tag{7}$$

One can view $\hat{\epsilon}_j$ as an empirical estimate of the quantity $\breve{\epsilon}_j = \breve{\epsilon}(K_j)$ that (as we have already pointed out) plays a crucial role in the bounds on the excess risk in empirical risk minimization problems in the RKHS context. In fact, since most often $\breve{\epsilon}_j \geq \sqrt{A\log N/n}$, we will redefine this quantity as

$$\breve{\epsilon}_j := \inf\left\{\epsilon \geq \sqrt{\frac{A\log N}{n}} : \breve{\gamma}_n^{(j)}(\delta) \leq \epsilon\delta + \epsilon^2,\ \forall \delta \in (0,1]\right\}. \tag{8}$$

We will use the following values of regularization parameters in problem (6): $\epsilon_j = \tau\hat{\epsilon}_j$, where τ is a sufficiently large constant.

It should be emphasized that the structure of complexity penalty and the choice of regularization parameters in (6) are closely related to the following bound on Rademacher processes indexed by functions from an RKHS $\mathcal{H}_K$: with a high probability, for all $h \in \mathcal{H}_K$,

$$|R_n(h)| \le C\left[\breve{\epsilon}(K)\|h\|_{L_2(\Pi)} + \breve{\epsilon}^2(K)\|h\|_{\mathcal{H}_K}\right].$$

Such bounds follow from the results of Section 3 and they provide a way to prove sparsity oracle inequalities for the estimators (6). The Rademacher process is defined as

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

where $\{\varepsilon_j\}$ is a sequence of i.i.d. Rademacher random variables (taking values $+1$ and -1 with probability $1/2$ each) independent of $\{X_j\}$.

We will use several basic facts of the empirical processes theory throughout the paper. They include symmetrization inequalities and contraction (comparison) inequalities for Rademacher processes that can be found in the books of Ledoux and Talagrand (1991) and van der Vaart and Wellner (1996). We also use Talagrand's concentration inequality for empirical processes (see, Talagrand (1996), Bousquet (2002)).

The main goal of the paper is to establish oracle inequalities for the excess risk of the estimator $\hat{f} = \hat{f}_1 + \ldots + \hat{f}_N$. In these inequalities, the excess risk of $\hat{f}$ is compared with the excess risk of an oracle $f := f_1 + \ldots + f_N,\ (f_1, \ldots, f_N) \in D$ with an error term depending on the degree of sparsity of the oracle, i.e., on the number of non-zero components $f_j \in \mathcal{H}_j$ in its additive representation. The oracle inequalities will be stated in the next section. Their proof relies on probabilistic bounds for empirical L_2-norms and data dependent regularization parameters $\hat{\epsilon}_j$. The results of Section 3 show that they can be bounded by their respective population counterparts. Using these tools and some bounds on empirical processes derived in Section 5, we prove in Section 4 the oracle inequalities for the estimator $\hat{f}$.

2 Oracle Inequalities

Considering the problem in the case when the domain D of (6) is not bounded, say, $D = \mathcal{H}^{(N)}$, leads to additional technical complications and might require some changes in the estimation procedure. To avoid this, we assume below that D is a bounded convex subset of $\mathcal{H}^{(N)}$. It

will be also assumed that, for all $j = 1, \dots, N$, $\sup_{x \in S} K_j(x,x) \le 1$, which, by elementary properties of RKHS, implies that $\|f_j\|_{L_\infty} \le \|f_j\|_{\mathcal{H}_j}, \qquad j = 1, \dots, N$. Because of this,

$$R_D := \sup_{(f_1,\dots,f_N) \in D} \|f_1 + \cdots + f_N\|_{L_\infty} < +\infty.$$

Denote $R_D^* := R_D \vee \|f_*\|_{L_\infty}$. We will allow the constants involved in the oracle inequalities stated and proved below to depend on the value of R_D^* (so, implicitly, it is assumed that this value is not too large).

We shall also assume that N is large enough, say, so that $\log N \ge 2 \log\log n$. This assumption is not essential to our development and is in place to avoid an extra term of the order $n^{-1} \log\log n$ in our risk bounds.

2.1 Loss Functions of Quadratic Type

We will formulate the **assumptions on the loss function** ℓ. The main assumption is that, for all $y \in T$, $\ell(y, \cdot)$ is a nonnegative convex function. In addition, we will assume that $\ell(y, 0),\ y \in T$ is uniformly bounded from above by a numerical constant. Moreover, suppose that, for all $y \in T$, $\ell(y, \cdot)$ is twice continuously differentiable and its first and second derivatives are uniformly bounded in $T \times [-R_D^*, R_D^*]$. Denote

$$m(R) := \frac{1}{2} \inf_{y \in T} \inf_{|u| \le R} \frac{\partial^2 \ell(y,u)}{\partial u^2}, \quad M(R) := \frac{1}{2} \sup_{y \in T} \sup_{|u| \le R} \frac{\partial^2 \ell(y,u)}{\partial u^2} \tag{9}$$

and let $m_* := m(R_D^*),\ M_* := M(R_D^*)$. We will assume that $m_* > 0$.

Denote

$$L_* := \sup_{|u| \le R_D^*, y \in T} \left| \frac{\partial \ell}{\partial u}(y,u) \right|.$$

Clearly, for all $y \in T$, the function $\ell(y, \cdot)$ satisfies Lipschitz condition with constant L_*.

The constants m_*, M_*, L_* will appear in a number of places in what follows. Without loss of generality, we can also assume that $m_* \le 1$ and $L_* \ge 1$ (otherwise, m_* and L_* can be replaced by a lower bound and an upper bound, respectively).

The loss functions satisfying the assumptions stated above will be called **the losses of quadratic type.**

If ℓ is a loss of quadratic type and $f = f_1 + \cdots + f_N,\ (f_1, \dots, f_N) \in D$, then

$$m_* \|f - f_*\|_{L_2(\Pi)}^2 \le \mathcal{E}(\ell \circ f) \le M_* \|f - f_*\|_{L_2(\Pi)}^2. \tag{10}$$

This bound easily follows from a simple argument based on Taylor expansion and it will be used later in the paper. If $\mathcal{H}$ is dense in $L_2(\Pi)$, then (10) implies that

$$\inf_{f\in\mathcal{H}} P(\ell\circ f) = \inf_{f\in L_2(\Pi)} P(\ell\circ f) = P(\ell\circ f_*). \tag{11}$$

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, $m(R)=1$ for all $R>0$. 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 functions of the form $\ell(y,u)=\phi(yu)$ are 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.

2.2 Geometry of the Dictionary

Now we introduce several important geometric characteristics of dictionaries consisting of kernels (or, equivalently, of RKHS). These characteristics are related to the degree of "dependence" of spaces of random variables $\mathcal{H}_j\subset L_2(\Pi),\ j=1,\dots,N$ and they will be involved in the oracle inequalities for the excess risk $\mathcal{E}(\ell\circ\hat{f})$.

First, for $J\subset\{1,\dots,N\}$ and $b\in[0,+\infty]$, denote

$$C_J^{(b)} := \biggl\{(h_1,\dots,h_N)\in\mathcal{H}^{(N)} : \sum_{j\notin J}\|h_j\|_{L_2(\Pi)} \le b\sum_{j\in J}\|h_j\|_{L_2(\Pi)}\biggr\}.$$

Clearly, the set $C_J^{(b)}$ is a cone in the space $\mathcal{H}^{(N)}$ that consists of vectors $(h_1,\dots,h_N)$ whose components corresponding to $j\in J$ "dominate" the rest of the components. This family of cones increases as b increases. For $b=0$, $C_J^{(b)}$ coincides with the linear subspace of vectors for which $h_j=0, j\notin J$. For $b=+\infty$, $C_J^{(b)}$ is the whole space $\mathcal{H}^{(N)}$.

The following quantity will play the most important role:

$$\beta_{2,b}(J;\Pi) := \beta_{2,b}(J) := \inf\biggl\{\beta>0 : \Bigl(\sum_{j\in J}\|h_j\|^2_{L_2(\Pi)}\Bigr)^{1/2} \le \beta\Bigl\|\sum_{j=1}^N h_j\Bigr\|_{L_2(\Pi)},\ (h_1,\dots,h_N)\in C_J^{(b)}\biggr\}.$$

Clearly, $\beta_{2,b}(J;\Pi)$ is a nondecreasing function of b. In the case of "simple dictionary" that consists of one-dimensional spaces similar quantities have been used in the literature on sparse recovery (see, e.g., Koltchinskii (2008), (2009a,b,c)).

The quantity $\beta_{2,b}(J;\Pi)$ can be upper bounded in terms of some other geometric characteristics that describe how "dependent" the spaces of random variables $\mathcal{H}_j \subset L_2(\Pi)$ are. These characteristics will be introduced below.

Given $h_j \in \mathcal{H}_j,\ j = 1,\dots,N$, denote by $\kappa(\{h_j : j \in J\})$ the minimal eigenvalue of the Gram matrix $(\langle h_j, h_k\rangle_{L_2(\Pi)})_{j,k\in J}$. Let

$$\kappa(J) := \inf\Big\{\kappa(\{h_j : j \in J\}) : h_j \in \mathcal{H}_j, \|h_j\|_{L_2(\Pi)} = 1\Big\}. \tag{12}$$

We will also use the notation

$$\mathcal{H}_J = \text{l.s.}\left(\bigcup_{j\in J}\mathcal{H}_j\right). \tag{13}$$

The following quantity is the maximal cosine of the angle in the space $L_2(\Pi)$ between the vectors in the subspaces $\mathcal{H}_I$ and $\mathcal{H}_J$ for some $I, J \subset \{1,\dots,N\}$:

$$\rho(I,J) := \sup\left\{\frac{\langle f, g\rangle_{L_2(\Pi)}}{\|f\|_{L_2(\Pi)}\|g\|_{L_2(\Pi)}} : f \in \mathcal{H}_I, g \in \mathcal{H}_J, f \neq 0, g \neq 0\right\}. \tag{14}$$

Denote $\rho(J) := \rho(J, J^c)$. The quantities $\rho(I,J)$ and $\rho(J)$ are very similar to the notion of **canonical correlation** in the multivariate statistical analysis.

There are other important geometric characteristics, frequently used in the theory of sparse recovery, including so called **"restricted isometry constants"** by Candes and Tao (2006). Define $\delta_d(\Pi)$ to be the smallest $\delta > 0$ such that for all $(h_1,\dots,h_N) \in \mathcal{H}^{(N)}$ and all $J \subset \{1,\dots,N\}$ with $\text{card}(J) = d$,

$$(1-\delta)\left(\sum_{j\in J}\|h_j\|^2_{L_2(\Pi)}\right)^{1/2} \le \left\|\sum_{j\in J} h_j\right\|_{L_2(\Pi)} \le (1+\delta)\left(\sum_{j\in J}\|h_j\|^2_{L_2(\Pi)}\right)^{1/2}.$$

This condition with a sufficiently small value of $\delta_d(\Pi)$ means that for all choices of J with $\text{card}(J) = d$ the functions in the spaces $\mathcal{H}_j,\ j \in J$ are "almost orthogonal" in $L_2(\Pi)$.

The following simple proposition easily follows from some statements in Koltchinskii (2009a,b), (2008) (where the case of simple dictionaries consisting of one-dimensional spaces $\mathcal{H}_j$ was considered).

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

$$\beta_{2,\infty}(J;\Pi) \le \frac{1}{\sqrt{\kappa(J)(1-\rho^2(J))}}.$$

Also, if $\mathrm{card}(J) = d$ *and* $\delta_{3d}(\Pi) \le \frac{1}{8b}$, *then* $\beta_{2,b}(J;\Pi) \le 4$.

Thus, such quantities as $\beta_{2,\infty}(J;\Pi)$ or $\beta_{2,b}(J;\Pi)$, for finite values of b, are reasonably small provided that the spaces of random variables $\mathcal{H}_j,\ j = 1, \dots, N$ satisfy proper conditions of "weakness of correlations".

2.3 Excess Risk Bounds

We are now in a position to formulate our main theorems that provide oracle inequalities for the excess risk $\mathcal{E}(\ell \circ \hat{f})$. In these theorems, $\mathcal{E}(\ell \circ \hat{f})$ will be compared with the excess risk $\mathcal{E}(\ell \circ f)$ of an oracle $(f_1, \dots, f_N) \in D$. Here and in what follows, $f := f_1 + \dots + f_N \in \mathcal{H}$. This is a little abuse of notation: we are ignoring the fact that such an additive representation of a function $f \in \mathcal{H}$ is not necessarily unique. In some sense, f denotes both the vector $(f_1, \dots, f_N) \in \mathcal{H}^{(N)}$ and the function $f_1 + \dots + f_N \in \mathcal{H}$. However, this is not going to cause a confusion in what follows. We will also use the following notations:

$$J_f := \{1 \le j \le N : f_j \ne 0\} \text{ and } d(f) := \mathrm{card}(J_f).$$

The error terms of the oracle inequalities will depend on the quantities $\breve{\epsilon}_j = \breve{\epsilon}(K_j)$ related to the "smoothness" properties of the RKHS and also on the geometric characteristics of the dictionary introduced above. In the first theorem, we will use the quantity $\beta_{2,\infty}(J_f;\Pi)$ to characterize the properties of the dictionary. In this case, there will be no assumptions on the quantities $\breve{\epsilon}_j$: these quantities could be of different order for different kernel machines, so, different components of the additive representation could have different "smoothness". In the second theorem, we will use a smaller quantity $\beta_{2,b}(J;\Pi)$ for a proper choice of parameter $b < \infty$. In this case, we will have to make an additional assumption that $\breve{\epsilon}_j,\ j = 1, \dots, N$ are all of the same order (up to a constant).

In both cases, we consider penalized empirical risk minimization problem (6) with data-dependent regularization parameters $\epsilon_j = \tau\hat{\epsilon}_j$, where $\hat{\epsilon}_j,\ j = 1, \dots, N$ are defined by (7) with some $A \ge 4$ and $\tau \ge BL_*$ for a numerical constant B.

Theorem 2 *There exist numerical constants $C_1, C_2 > 0$ such that, for all all oracles $(f_1, \dots, f_N) \in D$, with probability at least $1 - 3N^{-A/2}$,*

$$\mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\tau \sum_{j=1}^{N} \breve{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \tau^2 \sum_{j=1}^{N} \breve{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right)$$
$$\leq 2\mathcal{E}(\ell \circ f) + C_2 \tau^2 \sum_{j \in J_f} \breve{\epsilon}_j^2 \left(\frac{\beta_{2,\infty}^2(J_f, \Pi)}{m_*} + \|f_j\|_{\mathcal{H}_j} \right). \tag{15}$$

This result means that if there exists an oracle $(f_1, \dots, f_N) \in D$ such that

(a) the excess risk $\mathcal{E}(\ell \circ f)$ is small;

(b) the spaces $\mathcal{H}_j, j \in J_f$ are not strongly correlated with the spaces $\mathcal{H}_j, j \notin J_f$;

(c) $\mathcal{H}_j, j \in J_f$ are "well posed" in the sense that $\kappa(J_f)$ is not too small;

(d) $\|f_j\|_{\mathcal{H}_j}, j \in J_f$ are all bounded by a reasonable constant,

then the excess risk $\mathcal{E}(\ell \circ \hat{f})$ is essentially controlled by $\sum_{j \in J_f} \breve{\epsilon}_j^2$. At the same time, the oracle inequality provides a bound on the $L_2(\Pi)$-distances between the estimated components $\hat{f}_j$ and the components of the oracle (of course, everything is under the assumption that the loss is of quadratic type and m_* is bounded away from 0).

Not also that the constant 2 in front of the excess risk of the oracle $\mathcal{E}(\ell \circ f)$ can be replaced by $1 + \delta$ for any $\delta > 0$ with minor modifications of the proof (in this case, the constant C_2 depends on δ and is of the order $1/\delta$).

Suppose now that there exists $\breve{\epsilon} > 0$ and a constant $\Lambda > 0$ such that

$$\Lambda^{-1} \leq \frac{\breve{\epsilon}_j}{\breve{\epsilon}} \leq \Lambda, \ j = 1, \dots, N.$$

Theorem 3 *There exist numerical constants $C_1, C_2, b > 0$ such that, for all oracles $(f_1, \dots, f_N) \in D$, with probability at least $1 - 3N^{-A/2}$,*

$$\mathcal{E}(\ell \circ \hat{f}) + \frac{C_1}{\Lambda} \left(\tau \breve{\epsilon} \sum_{j=1}^{N} \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \tau^2 \breve{\epsilon}^2 \sum_{j=1}^{N} \|\hat{f}_j\|_{\mathcal{H}_j} \right)$$
$$\leq 2\mathcal{E}(\ell \circ f) + C_2 \Lambda \tau^2 \breve{\epsilon}^2 \left(\frac{\beta_{2,b\Lambda^2}^2(J_f, \Pi)}{m_*} d(f) + \sum_{j \in J_f} \|f_j\|_{\mathcal{H}_j} \right). \tag{16}$$

As before, the constant 2 in the upper bound can be replaced by $1+\delta$, but, in this case, the constants C_2 and b would be of the order $\frac{1}{\delta}$. The meaning of this result is that if there exists an oracle $(f_1, \dots, f_N) \in D$ such that

(a) the excess risk $\mathcal{E}(\ell \circ f)$ is small;

(b) the "restricted isometry" constant $\delta_{3d}(\Pi)$ is small for $d = d(f)$;

(c) $\|f_j\|_{\mathcal{H}_j}, j \in J_f$ are all bounded by a reasonable constant,

then the excess risk $\mathcal{E}(\ell \circ \hat{f})$ is essentially controlled by $d(f)\breve{\epsilon}^2$. At the same time, the distance $\sum_{j=1}^N \|\hat{f}_j - f_j\|_{L_2(\Pi)}$ between the estimator and the oracle is controlled by $d(f)\breve{\epsilon}$. In particular, this implies that the empirical solution $(\hat{f}_1, \dots, \hat{f}_N)$ is "approximately sparse" in the sense that $\sum_{j \notin J_f} \|\hat{f}\|_{L_2(\Pi)}$ is of the order $d(f)\breve{\epsilon}$.

Remarks. 1. It is easy to check that theorems 2 and 3 hold also if one replaces N in the definitions (7) of $\hat{\epsilon}_j$ and (8) of $\breve{\epsilon}_j$ by an arbitrary $\bar{N} \geq N$ such that $\log \bar{N} \geq 2 \log \log n$ (a similar condition on N introduced early in Section 2 is not needed here). In this case, the probability bounds in the theorems become $1 - 3\bar{N}^{-A/2}$. This change might be of interest if one uses the results for a dictionary consisting of just one RKHS ($N = 1$), which is not the focus of this paper.

2. If the distribution dependent quantities $\breve{\epsilon}_j$, $j = 1, \dots, N$ are known and used as regularization parameters in (6), the oracle inequalities of theorems 2 and 3 also hold (with obvious simplifications of their proofs). For instance, in the case when $S = [0,1]^N$, the design distribution Π is uniform and, for each $j = 1, \dots, N$, $\mathcal{H}_j$ is a Sobolev space of functions of smoothness $\alpha > 1/2$ depending only on the j-th variable, we have $\breve{\epsilon}_j \asymp n^{-\alpha/(2\alpha+1)}$. Taking in this case

$$\epsilon_j = \tau \left(n^{-\alpha/(2\alpha+1)} \bigvee \sqrt{\frac{A \log N}{n}} \right)$$

would lead to oracle inequalities for sparse additive models is spirit of Meier, van de Geer and Bühlmann (2009). More precisely, if $\mathcal{H}_j := \{h \in W^{\alpha,2}[0,1] : \int_0^1 h(x)dx = 0\}$, then, for uniform distribution Π, the spaces $\mathcal{H}_j$ are orthogonal in $L_2(\Pi)$ (recall that $\mathcal{H}_j$ is viewed as a space of functions depending on the j-th coordinate). Assume, for simplicity, that ℓ is the quadratic loss and that the regression function f_* can be represented as $f_* = \sum_{j \in J} f_{*,j}$, where J is a subset of $\{1, \dots, N\}$ of cardinality d and $\|f_{*,j}\|_{\mathcal{H}_j} \leq 1$. Then it easily follows

from the bound of Theorem 3 that with probability at least $1 - 3N^{-A/2}$

$$\mathcal{E}(f) = \|f - f_*\|^2_{L_2(\Pi)} \le C\tau^2 d\left(n^{-2\alpha/(2\alpha+1)} \bigvee \frac{A\log N}{n}\right).$$

Note that, up to a constant, this essentially coincides with the minimax lower bound in this type of problems obtained recently by Raskutti, Wainwright and Yu (2009). Of course, if the design distribution is not necessarily uniform, an adaptive choice of regularization parameters might be needed even in such simple examples and the approach described above leads to minimax optimal rates.

3 Preliminary Bounds

In this section, the case of a single RKHS $\mathcal{H}_K$ associated with a kernel K is considered. We assume that $K(x,x) \le 1,\ x \in S$. This implies that, for all $h \in \mathcal{H}_K$, $\|h\|_{L_2(\Pi)} \le \|h\|_{L_\infty} \le \|h\|_{\mathcal{H}_K}$.

3.1 Comparison of $\|\cdot\|_{L_2(\Pi_n)}$ and $\|\cdot\|_{L_2(\Pi)}$

First, we study the relationship between the empirical and the population L_2 norms for functions in $\mathcal{H}_K$.

Theorem 4 *Assume that $A \ge 1$ and $\log N \ge 2\log\log n$. Then there exists a numerical constant $C > 0$ such that with probability at least $1 - N^{-A}$ for all $h \in \mathcal{H}_K$*

$$\|h\|_{L_2(\Pi)} \le C\left(\|h\|_{L_2(\Pi_n)} + \bar\epsilon\|h\|_{\mathcal{H}_K}\right); \tag{17}$$

$$\|h\|_{L_2(\Pi_n)} \le C\left(\|h\|_{L_2(\Pi)} + \bar\epsilon\|h\|_{\mathcal{H}_K}\right), \tag{18}$$

where

$$\bar\epsilon = \bar\epsilon(K) := \inf\left\{\epsilon \ge \sqrt{\frac{A\log N}{n}} : \mathbb{E}\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ \|h\|_{L_2(\Pi)}\le\delta}} |R_n(h)| \le \epsilon\delta + \epsilon^2, \forall\delta \in (0,1]\right\}. \tag{19}$$

Proof. Observe that the inequalities hold trivially when $h = 0$. We shall therefore consider only the case when $h \ne 0$. By symmetrization inequality,

$$\mathbb{E}\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} \left|(\Pi_n - \Pi)h^2\right| \le 2\mathbb{E}\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} \left|R_n(h^2)\right|, \tag{20}$$

and, by contraction inequality, we further have

$$\mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} \left|(\Pi_n-\Pi)h^2\right| \le 8\mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} |R_n(h)|\,. \tag{21}$$

The definition of $\bar\epsilon$ implies that

$$\mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} \left|(\Pi_n-\Pi)h^2\right| \le 8\mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi)}\le 2^{-j+1}}} |R_n(h)| \le 8\left(\bar\epsilon 2^{-j+1}+\bar\epsilon^2\right). \tag{22}$$

An application of Talagrand's concentration inequality yields

$$\begin{aligned}
\sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} \left|(\Pi_n-\Pi)h^2\right| &\le 2\Bigg(\mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} \left|(\Pi_n-\Pi)h^2\right| \\
&\qquad +2^{-j+1}\sqrt{\frac{t+2\log j}{n}}+\frac{t+2\log j}{n}\Bigg) \\
&\le 32\left(\bar\epsilon 2^{-j}+\bar\epsilon^2+2^{-j}\sqrt{\frac{t+2\log j}{n}}+\frac{t+2\log j}{n}\right)
\end{aligned}$$

with probability at least $1-\exp(-t-2\log j)$ for any natural number j. Now, by the union bound, for all j such that $2\log j \le t$,

$$\sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}} \left|(\Pi_n-\Pi)h^2\right| \le 32\left(\bar\epsilon 2^{-j}+\bar\epsilon^2+2^{-j}\sqrt{\frac{t+2\log j}{n}}+\frac{t+2\log j}{n}\right) \tag{23}$$

with probability at least

$$1-\sum_{j:2\log j\le t}\exp(-t-2\log j) = 1-\exp(-t)\sum_{j:2\log j\le t} j^{-2} \ge 1-2\exp(-t). \tag{24}$$

Recall that $\bar\epsilon \ge (A\log N/n)^{1/2}$ and $\|h\|_{L_2(\Pi)} \le \|h\|_{\mathcal{H}_K}$. Taking $t = A\log N + \log 4$, we easily get that, for all $h\in\mathcal{H}_K$ such that $\|h\|_{\mathcal{H}_K}=1$ and $\|h\|_{L_2(\Pi)} \ge \exp\{-N^{A/2}\}$,

$$\left|(\Pi_n-\Pi)h^2\right| \le C\left(\bar\epsilon\|h\|_{L_2(\Pi)}+\bar\epsilon^2\right) \tag{25}$$

with probability at least $1-0.5N^{-A}$ and with a numerical constant $C>0$. In other words, with the same probability, for all $h\in\mathcal{H}_K$ such that $\frac{\|h\|_{L_2(\Pi)}}{\|h\|_{\mathcal{H}_K}} \ge \exp\{-N^{A/2}\}$,

$$\left|(\Pi_n-\Pi)h^2\right| \le C\left(\bar\epsilon\|h\|_{L_2(\Pi)}\|h\|_{\mathcal{H}_K}+\bar\epsilon^2\|h\|_{\mathcal{H}_K}^2\right). \tag{26}$$

Therefore, for all $h \in \mathcal{H}_K$ such that

$$\frac{\|h\|_{L_2(\Pi)}}{\|h\|_{\mathcal{H}_K}} > \exp(-N^{A/2}) \tag{27}$$

we have

$$\begin{aligned} \|h\|^2_{L_2(\Pi)} = \Pi h^2 &\leq \|h\|^2_{L_2(\Pi_n)} + C\left(\bar{\epsilon}\|h\|_{L_2(\Pi)}\|h\|_{\mathcal{H}_K} + \bar{\epsilon}^2\|h\|^2_{\mathcal{H}_K}\right), \\ \|h\|^2_{L_2(\Pi_n)} = \Pi_n h^2 &\leq \|h\|^2_{L_2(\Pi)} + C\left(\bar{\epsilon}\|h\|_{L_2(\Pi)}\|h\|_{\mathcal{H}_K} + \bar{\epsilon}^2\|h\|^2_{\mathcal{H}_K}\right). \end{aligned}$$

It can be now deduced that, for a proper value of numerical constant C,

$$\|h\|_{L_2(\Pi)} \leq C\left(\|h\|_{L_2(\Pi_n)} + \bar{\epsilon}\|h\|_{\mathcal{H}_K}\right) \text{ and } \|h\|_{L_2(\Pi_n)} \leq C\left(\|h\|_{L_2(\Pi)} + \bar{\epsilon}\|h\|_{\mathcal{H}_K}\right). \tag{28}$$

It remains to consider the case when

$$\frac{\|h\|_{L_2(\Pi)}}{\|h\|_{\mathcal{H}_K}} \leq \exp(-N^{A/2}). \tag{29}$$

Following a similar argument as before, with probability at least $1 - 0.5N^{-A}$,

$$\begin{aligned} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi)} \leq \exp(-N^{A/2})}} \left|(\Pi_n - \Pi)h^2\right| &\leq 16\Bigg(\bar{\epsilon}\exp(-N^{A/2}) + \bar{\epsilon}^2 \\ &\qquad + \exp(-N^{A/2})\sqrt{\frac{A\log N}{n}} + \frac{A\log N}{n}\Bigg). \end{aligned}$$

Under the conditions $A \geq 1, \log N \geq 2\log\log n$,

$$\bar{\epsilon} \geq \left(\frac{A\log N}{n}\right)^{1/2} \geq \exp(-N^{A/2}). \tag{30}$$

Then

$$\sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi)} \leq \exp(-N^{A/2})}} \left|(\Pi_n - \Pi)h^2\right| \leq C\bar{\epsilon}^2. \tag{31}$$

with probability at least $1 - 0.5N^{-A}$, which also implies (17) and (18), and the result follows. ∎

Theorem 4 shows that the two norms $\|h\|_{L_2(\Pi_n)}$ and $\|h\|_{L_2(\Pi)}$ are of the same order up to an error term $\bar{\epsilon}\|h\|_{\mathcal{H}_K}$.

3.2 Comparison of $\hat{\epsilon}(K)$, $\bar{\epsilon}(K)$, $\breve{\epsilon}(K)$ and $\check{\epsilon}(K)$

Recall the definitions

$$\breve{\gamma}_n(\delta) := \left(n^{-1} \sum_{k=1}^{\infty} (\lambda_k \wedge \delta^2) \right)^{1/2}, \qquad \delta \in (0,1]$$

where $\{\lambda_k\}$ are the eigenvalues of the integral operator T_K from $L_2(\Pi)$ into $L_2(\Pi)$ with kernel K, and, for some $A \geq 1$,

$$\breve{\epsilon}(K) := \inf \left\{ \epsilon \geq \sqrt{\frac{A \log N}{n}} : \breve{\gamma}_n(\delta) \leq \epsilon\delta + \epsilon^2, \forall \delta \in (0,1] \right\}.$$

It follows from Lemma 42 of Mendelson (2002) (with an additional application of Cauchy-Schwarz inequality for the upper bound and Hoffmann-Jørgensen inequality for the lower bound, see also Koltchinskii (2008)) that, for some numerical constants $C_1, C_2 > 0$,

$$C_1 \left(n^{-1} \sum_{k=1}^{n} (\lambda_k \wedge \delta^2) \right)^{1/2} - n^{-1} \leq \mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi)} \leq \delta}} |R_n(h)| \leq C_2 \left(n^{-1} \sum_{k=1}^{n} (\lambda_k \wedge \delta^2) \right)^{1/2}, \quad (32)$$

This fact and the definitions of $\breve{\epsilon}(K), \bar{\epsilon}(K)$ easily imply the following result.

Proposition 5 *Under the condition $K(x,x) \leq 1, x \in S$, there exist numerical constants $C_1, C_2 > 0$ such that*

$$C_1 \breve{\epsilon}(K) \leq \bar{\epsilon}(K) \leq C_2 \breve{\epsilon}(K). \quad (33)$$

If K is the kernel of the projection operator onto a finite-dimensional subspace $\mathcal{H}_K$ of $L_2(\Pi)$, it is easy to check that $\breve{\epsilon}(K) \asymp \sqrt{\frac{\dim(\mathcal{H}_K)}{n}}$ (recall the notation $a \asymp b$, which means that there exists a numerical constant $c > 0$ such that $c^{-1} \leq a/b \leq c$). If the eigenvalues λ_k decay at a polynomial rate, i.e., $\lambda_k \asymp k^{-2\beta}$ for some $\beta > 1/2$, then $\breve{\epsilon}(K) \asymp n^{-\beta/(2\beta+1)}$.

Recall the notation

$$\hat{\epsilon}(K) := \inf \left\{ \epsilon \geq \sqrt{\frac{A \log N}{n}} : \left(\frac{1}{n} \sum_{k=1}^{n} \left(\hat{\lambda}_k \wedge \delta^2 \right) \right)^{1/2} \leq \epsilon\delta + \epsilon^2, \forall \delta \in (0,1] \right\}, \quad (34)$$

where $\{\hat{\lambda}_k\}$ denote the eigenvalues of the Gram matrix $\hat{K} := \Big(K(X_i, X_j) \Big)_{i,j=1,\ldots,n}$. It follows again from the results of Mendelson (2002) [namely, one can follow the proof of Lemma 42

in the case when the RKHS $\mathcal{H}_K$ is restricted to the sample $X_1, \ldots, X_n$ and the expectations are conditional on the sample; then one uses Cauchy-Schwarz and Hoffmann-Jørgensen inequalities as in the proof of (32)] that for some numerical constants $C_1, C_2 > 0$

$$C_1 \left(n^{-1} \sum_{k=1}^{n} (\hat{\lambda}_k \wedge \delta^2) \right)^{1/2} - n^{-1} \leq \mathbb{E}_\varepsilon \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi_n)} \leq \delta}} |R_n(h)| \leq C_2 \left(n^{-1} \sum_{k=1}^{n} (\hat{\lambda}_k \wedge \delta^2) \right)^{1/2}, \quad (35)$$

where $\mathbb{E}_\varepsilon$ indicates that the expectation is taken over the Rademacher random variables only (conditionally on $X_1, \ldots, X_n$). Therefore, if we denote by

$$\tilde{\epsilon}(K) := \inf \left\{ \epsilon \geq \sqrt{\frac{A \log N}{n}} : \mathbb{E}_\varepsilon \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi_n)} \leq \delta}} |R_n(h)| \leq \epsilon\delta + \epsilon^2, \forall \delta \in (0,1] \right\} \quad (36)$$

the empirical version of $\bar{\epsilon}(K)$, then $\hat{\epsilon}(K) \asymp \tilde{\epsilon}(K)$. We will now show that $\tilde{\epsilon}(K) \asymp \bar{\epsilon}(K)$ with a high probability.

Theorem 6 *Suppose that $A \geq 1$ and $\log N \geq 2 \log\log n$. There exist numerical constants $C_1, C_2 > 0$ such that*

$$C_1 \bar{\epsilon}(K) \leq \tilde{\epsilon}(K) \leq C_2 \bar{\epsilon}(K), \quad (37)$$

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

Proof. Let $t := A \log N + \log 14$. It follows from Talagrand concentration inequality that

$$\begin{aligned} & \mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j} < \|h\|_{L_2(\Pi)} \leq 2^{-j+1}}} |R_n(h)| \\ \leq \; & 2 \left(\sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j} < \|h\|_{L_2(\Pi)} \leq 2^{-j+1}}} |R_n(h)| + 2^{-j+1} \sqrt{\frac{t + 2\log j}{n}} + \frac{t + 2\log j}{n} \right). \end{aligned}$$

with probability at least $1 - \exp(-t - 2\log j)$. On the other hand, as derived in the proof of Theorem 4 (see (23))

$$\sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ 2^{-j} < \|h\|_{L_2(\Pi)} \leq 2^{-j+1}}} \left| (\Pi_n - \Pi) h^2 \right| \leq 32 \left(\bar{\epsilon} 2^{-j} + \bar{\epsilon}^2 + 2^{-j} \sqrt{\frac{t + 2\log j}{n}} + \frac{t + 2\log j}{n} \right) \quad (38)$$

with probability at least $1-\exp(-t-2\log j)$. We will use these bounds only for j such that $2\log j \le t$. In this case, the second bound implies that, for some numerical constant $c>0$ and all h satisfying the conditions $\|h\|_{\mathcal{H}_K}=1, 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}$, we have $\|h\|_{L_2(\Pi_n)}\le c(2^{-j}+\bar{\epsilon})$ (again, see the proof of Theorem 4). Combining these bounds, we get that with probability at least $1-2\exp(-t-2\log j)$,

$$\mathbb{E}\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}}|R_n(h)| \le 2\Bigg(\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ \|h\|_{L_2(\Pi_n)}\le c\delta_j}}|R_n(h)| + 2^{-j+1}\sqrt{\frac{t+2\log j}{n}}+\frac{t+2\log j}{n}\Bigg).$$

where $\delta_j=\bar{\epsilon}+2^{-j}$.

Applying now Talagrand concentration inequality to the Rademacher process conditionally on the observed data $X_1,\dots,X_n$ yields

$$\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ \|h\|_{L_2(\Pi_n)}\le c\delta_j}}|R_n(h)| \le 2\Bigg(\mathbb{E}_\varepsilon\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ \|h\|_{L_2(\Pi_n)}\le c\delta_j}}|R_n(h)| + C\delta_j\sqrt{\frac{t+2\log j}{n}}+\frac{t+2\log j}{n}\Bigg),$$

with conditional probability at least $1-\exp(-t-2\log j)$. From this and from the previous bound it is not hard to deduce that, for some numerical constants C, C' and for all j such that $2\log j\le t$,

$$\begin{aligned}
&\mathbb{E}\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ 2^{-j}<\|h\|_{L_2(\Pi)}\le 2^{-j+1}}}|R_n(h)|\\
\le\ & C'\Bigg(\mathbb{E}_\varepsilon\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ \|h\|_{L_2(\Pi_n)}\le c\delta_j}}|R_n(h)| + \delta_j\sqrt{\frac{t+2\log j}{n}}+\frac{t+2\log j}{n}\Bigg)\\
\le\ & C(\tilde{\epsilon}\delta_j+\tilde{\epsilon}^2)\le C(\tilde{\epsilon}2^{-j}+\tilde{\epsilon}\bar{\epsilon}+\tilde{\epsilon}^2)
\end{aligned}$$

with probability at least $1-3\exp(-t-2\log j)$. In obtaining the second inequality, we used the definition of $\tilde{\epsilon}$ and the fact that, for $t=A\log N+\log 14, 2\log j\le t$, $c_1\tilde{\epsilon}\ge(t+2\log j/n)^{1/2}$, where c_1 is a numerical constant. Now, by the union bound, the above inequality holds with probability at least

$$1-3\sum_{j:2\log j\le t}\exp(-t-2\log j)\ge 1-6\exp(-t) \tag{39}$$

for all j such that $2\log j\le t$ simultaneously. Similarly, it can be shown that

$$\mathbb{E}\sup_{\substack{\|h\|_{\mathcal{H}_K}=1\\ \|h\|_{L_2(\Pi)}\le\exp(-N^{A/2})}}|R_n(h)|\le C\Big(\tilde{\epsilon}\exp(-N^{A/2})+\tilde{\epsilon}\bar{\epsilon}+\tilde{\epsilon}^2\Big)$$

with probability at least $1 - \exp(-t)$.

For $t = A \log N + \log 14$, we get

$$\mathbb{E} \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi)} \le \delta}} |R_n(h)| \le C \left(\tilde{\epsilon}\delta + \tilde{\epsilon}\bar{\epsilon} + \tilde{\epsilon}^2 \right), \tag{40}$$

for all $0 < \delta \le 1$, with probability at least $1 - 7\exp(-t) = 1 - N^{-A}/2$. Now by the definition of $\bar{\epsilon}$, we obtain

$$\bar{\epsilon} \le C \max\{\tilde{\epsilon}, (\tilde{\epsilon}\bar{\epsilon} + \tilde{\epsilon}^2)^{1/2}\}, \tag{41}$$

which implies that $\bar{\epsilon} \le C\tilde{\epsilon}$ with probability at least $1 - N^{-A}/2$.

Similarly one can show that

$$\mathbb{E}_\epsilon \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi)} \le \delta}} |R_n(h)| \le C \left(\bar{\epsilon}\delta + \tilde{\epsilon}\bar{\epsilon} + \bar{\epsilon}^2 \right), \tag{42}$$

for all $0 < \delta \le 1$, with probability at least $1 - N^{-A}/2$, which implies that $\tilde{\epsilon} \le C\bar{\epsilon}$ with probability at least $1 - N^{-A}/2$. The proof can then be completed by the union bound. ∎

Define

$$\check{\epsilon} := \check{\epsilon}(K) := \inf \left\{ \epsilon \ge \sqrt{\frac{A \log N}{n}} : \sup_{\substack{\|h\|_{\mathcal{H}_K}=1 \\ \|h\|_{L_2(\Pi)} \le \delta}} |R_n(h)| \le \epsilon\delta + \epsilon^2, \forall \delta \in (0, 1] \right\}. \tag{43}$$

The next statement can be proved similarly to Theorem 6.

Theorem 7 *There exist numerical constants $C_1, C_2 > 0$ such that*

$$C_1 \bar{\epsilon}(K) \le \check{\epsilon}(K) \le C_2 \bar{\epsilon}(K), \tag{44}$$

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

Suppose now that $\{K_1, \ldots, K_N\}$ is a dictionary of kernels. Recall that $\bar{\epsilon}_j = \bar{\epsilon}(K_j)$, $\hat{\epsilon}_j = \hat{\epsilon}(K_j)$ and $\check{\epsilon}_j = \check{\epsilon}(K_j)$.

It follows from theorems 4, 6, 7 and the union bound that with probability at least $1 - 3N^{-A+1}$ for all $j = 1, \dots, N$

$$\|h\|_{L_2(\Pi)} \le C\left(\|h\|_{L_2(\Pi_n)} + \bar{\epsilon}_j\|h\|_{\mathcal{H}_K}\right), \|h\|_{L_2(\Pi_n)} \le C\left(\|h\|_{L_2(\Pi)} + \bar{\epsilon}_j\|h\|_{\mathcal{H}_K}\right),\ h \in \mathcal{H}_j, \quad (45)$$

$$C_1\bar{\epsilon}_j \le \hat{\epsilon}_j \le C_2\bar{\epsilon}_j \text{ and } C_1\bar{\epsilon}_j \le \check{\epsilon}_j \le C_2\bar{\epsilon}_j. \quad (46)$$

Note also that

$$3N^{-A+1} = \exp\{-(A-1)\log N + \log 3\} \le \exp\{-(A/2)\log N\} = N^{-A/2},$$

provided that $A \ge 4$ and $N \ge 3$. Thus, under these additional constraints, (45) and (46) hold for all $j = 1, \dots, N$ with probability at least $1 - N^{-A/2}$.

4 Proofs of the Oracle Inequalities

For an arbitrary set $J \subseteq \{1, \dots, N\}$ and $b \in (0, +\infty)$, denote

$$\mathcal{K}_J^{(b)} := \left\{ (f_1, \dots, f_N) \in \mathcal{H}^{(N)} : \sum_{j \notin J} \bar{\epsilon}_j\|f_j\|_{L_2(\Pi)} \le b \sum_{j \in J} \bar{\epsilon}_j\|f_j\|_{L_2(\Pi)} \right\} \quad (47)$$

and let

$$\beta_b(J) = \inf\left\{ \beta \ge 0 : \sum_{j \in J} \bar{\epsilon}_j\|f_j\|_{L_2(\Pi)} \le \beta\|f_1 + \cdots + f_N\|_{L_2(\Pi)}, (f_1, \dots, f_N) \in \mathcal{K}_J^{(b)} \right\}. \quad (48)$$

It is easy to see that, for all nonempty sets J, $\beta_b(J) \ge \max_{j \in J} \bar{\epsilon}_j \ge \sqrt{\frac{A \log N}{n}}$.

Theorems 2 and 3 will be easily deduced from the following technical result.

Theorem 8 *There exist numerical constants $C_1, C_2, B > 0$ and $b > 0$ such that, for all $\tau \ge BL_*$ in the definition of $\epsilon_j = \tau\hat{\epsilon}_j$, $j = 1, \dots, N$ and for all oracles $(f_1, \dots, f_N) \in D$,*

$$\mathcal{E}(\ell \circ \hat{f}) + C_1\left(\sum_{j=1}^N \tau\bar{\epsilon}_j\|\hat{f}_j - f_j\|_{L_2(\Pi)} + \sum_{j=1}^N \tau^2\bar{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j}\right) \quad (49)$$

$$\le\ 2\mathcal{E}(\ell \circ f) + C_2\tau^2\left(\sum_{j \in J_f} \bar{\epsilon}_j^2\|f_j\|_{\mathcal{H}_j} + \frac{\beta_b^2(J_f)}{m_*}\right) \quad (50)$$

with probability at least $1 - 3N^{-A/2}$. Here $A \ge 4$ is a constant involved in the definitions of $\bar{\epsilon}_j, \hat{\epsilon}_j, j = 1, \dots, N$.

Proof. Recall that

$$\left(\hat{f}_1, \ldots, \hat{f}_N\right) := \underset{(f_1,\ldots,f_N)\in D}{\operatorname{argmin}} \left[P_n(\ell \circ (f_1 + \cdots + f_N)) + \sum_{j=1}^{N} \left(\tau \hat{\epsilon}_j \|f_j\|_{L_2(\Pi_n)} + \tau^2 \hat{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) \right],$$

and that we write $f := f_1 + \cdots + f_N, \qquad \hat{f} := \hat{f}_1 + \cdots + \hat{f}_N$. Hence, for all $(f_1, \ldots, f_N) \in D$,

$$\begin{aligned} & P_n(\ell \circ \hat{f}) + \sum_{j=1}^{N} \left(\tau \hat{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi_n)} + \tau^2 \hat{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\ \leq \; & P_n(\ell \circ f) + \sum_{j=1}^{N} \left(\tau \hat{\epsilon}_j \|f_j\|_{L_2(\Pi_n)} + \tau^2 \hat{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right). \end{aligned}$$

By a simple algebra,

$$\begin{aligned} & \mathcal{E}(\ell \circ \hat{f}) + \sum_{j=1}^{N} \left(\tau \hat{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi_n)} + \tau^2 \hat{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\ \leq \; & \mathcal{E}(\ell \circ f) + \sum_{j=1}^{N} \left(\tau \hat{\epsilon}_j \|f_j\|_{L_2(\Pi_n)} + \tau^2 \hat{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) + \left| (P_n - P) \left(\ell \circ \hat{f} - \ell \circ f \right) \right| \end{aligned}$$

and, by the triangle inequality,

$$\begin{aligned} & \mathcal{E}(\ell \circ \hat{f}) + \sum_{j \notin J_f} \tau \hat{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi_n)} + \sum_{j=1}^{N} \tau^2 \hat{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \\ \leq \; & \mathcal{E}(\ell \circ f) + \sum_{j \in J_f} \tau \hat{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi_n)} + \sum_{j \in J_f} \tau^2 \hat{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} + \left| (P_n - P) \left(\ell \circ \hat{f} - \ell \circ f \right) \right|. \end{aligned}$$

We now take advantage of (45) and (46) to replace $\hat{\epsilon}_j$s by $\bar{\epsilon}_j$s and $\|\cdot\|_{L_2(\Pi_n)}$ by $\|\cdot\|_{L_2(\Pi)}$. Specifically, there exists a numerical constant $C > 1$ and an event E of probability at least $1 - N^{-A/2}$ such that

$$\frac{1}{C} \leq \min \left\{ \frac{\hat{\epsilon}_j}{\bar{\epsilon}_j} : j = 1, \ldots, N \right\} \leq \max \left\{ \frac{\hat{\epsilon}_j}{\bar{\epsilon}_j} : j = 1, \ldots, N \right\} \leq C \tag{51}$$

and, for all $j = 1, \ldots, N$,

$$\frac{1}{C} \|\hat{f}_j\|_{L_2(\Pi)} - \bar{\epsilon}_j \|\hat{f}_j\|_{\mathcal{H}_j} \leq \|\hat{f}_j\|_{L_2(\Pi_n)} \leq C \left(\|\hat{f}_j\|_{L_2(\Pi)} + \bar{\epsilon}_j \|\hat{f}_j\|_{\mathcal{H}_j} \right). \tag{52}$$

Taking $\tau \geq C/(C-1)$, we have that, on the event E,

$$\begin{aligned}
& \mathcal{E}(\ell \circ \hat{f}) + \sum_{j \notin J_f} \tau \hat{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi_n)} + \sum_{j=1}^N \tau^2 \hat{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \\
\geq \; & \mathcal{E}(\ell \circ \hat{f}) + \frac{1}{C^2} \left(\sum_{j \notin J_f} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi_n)} + \sum_{j=1}^N \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\
\geq \; & \mathcal{E}(\ell \circ \hat{f}) + \frac{1}{C^2} \left(\sum_{j \notin J_f} \tau \bar{\epsilon}_j \left(\frac{1}{C} \|\hat{f}_j\|_{L_2(\Pi)} - \bar{\epsilon}_j \|\hat{f}_j\|_{\mathcal{H}_j} \right) + \sum_{j=1}^N \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\
\geq \; & \mathcal{E}(\ell \circ \hat{f}) + \frac{1}{C^3} \left(\sum_{j \notin J_f} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \sum_{j=1}^N \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \mathcal{E}(\ell \circ f) + \sum_{j \in J_f} \left(\tau \hat{\epsilon}_j \|f_j - \hat{f}_j\|_{L_2(\Pi_n)} + \tau^2 \hat{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) \\
\leq \; & \mathcal{E}(\ell \circ f) + C^2 \sum_{j \in J_f} \left(\tau \bar{\epsilon}_j \|f_j - \hat{f}_j\|_{L_2(\Pi_n)} + \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) \\
\leq \; & \mathcal{E}(\ell \circ f) + C^3 \sum_{j \in J_f} \tau \bar{\epsilon}_j \left(\|f_j - \hat{f}_j\|_{L_2(\Pi)} + \bar{\epsilon}_j \|f_j - \hat{f}_j\|_{\mathcal{H}_j} \right) + C^2 \sum_{j \in J_f} \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \\
\leq \; & \mathcal{E}(\ell \circ f) + C^3 \sum_{j \in J_f} \tau \bar{\epsilon}_j \left(\|f_j - \hat{f}_j\|_{L_2(\Pi)} + \bar{\epsilon}_j \|f_j\|_{\mathcal{H}_j} + \bar{\epsilon}_j \|\hat{f}_j\|_{\mathcal{H}_j} \right) + C^2 \sum_{j \in J_f} \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \\
\leq \; & \mathcal{E}(\ell \circ f) + 2C^3 \sum_{j \in J_f} \left(\tau \bar{\epsilon}_j \|f_j - \hat{f}_j\|_{L_2(\Pi)} + \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) + C^3 \sum_{j \in J_f} \tau \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j}.
\end{aligned}$$

Therefore, by taking τ large enough, namely $\tau \geq \frac{C}{C-1} \vee (2C^6)$, we can find numerical constants $0 < C_1 < 1 < C_2$ such that, on the event E,

$$\begin{aligned}
& \mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\sum_{j \notin J_f} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \sum_{j=1}^N \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\
\leq \; & \mathcal{E}(\ell \circ f) + C_2 \sum_{j \in J_f} \left(\tau \bar{\epsilon}_j \|f_j - \hat{f}_j\|_{L_2(\Pi)} + \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) \\
& + \left| (P_n - P) \left(\ell \circ \hat{f} - \ell \circ f \right) \right|.
\end{aligned}$$

We now bound the empirical process $\left| (P_n - P) \left(\ell \circ \hat{f} - \ell \circ f \right) \right|$, where we use the following result that will be proved in the next section. Suppose that $f = \sum_{j=1}^N f_j$, $f_j \in \mathcal{H}_j$

and $\|f\|_{L_\infty} \le R$ (we will need it with $R = R_D^*$). Denote

$$\mathcal{G}(\Delta_-, \Delta_+, R) = \Big\{ g : \sum_{j=1}^N \bar{\epsilon}_j \|g_j - f_j\|_{L_2(\Pi)} \le \Delta_-, \\ \sum_{j=1}^N \bar{\epsilon}_j^2 \|g_j - f_j\|_{\mathcal{H}_j} \le \Delta_+, \|\sum_{j=1}^N g_j\|_{L_\infty} \le R \Big\}.$$

Lemma 9 *There exists a numerical constant $C > 0$ such that for an arbitrary $A \ge 1$ involved in the definition of $\bar{\epsilon}_j$, $j = 1, \dots, N$ with probability at least $1 - 2N^{-A/2}$, for all*

$$\Delta_- \le e^N, \Delta_+ \le e^N, \tag{53}$$

the following bound holds

$$\sup_{g \in \mathcal{G}(\Delta_-, \Delta_+, R_D^*)} |(P_n - P)(\ell \circ g - \ell \circ f)| \le C L_* \left(\Delta_- + \Delta_+ + e^{-N} \right). \tag{54}$$

Assuming that

$$\sum_{j=1}^N \bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} \le e^N, \quad \sum_{j=1}^N \bar{\epsilon}_j^2 \|\hat{f}_j - f_j\|_{\mathcal{H}_j} \le e^N \tag{55}$$

and using the lemma, we get

$$\begin{aligned} & \mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\sum_{j \notin J_f} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \sum_{j=1}^N \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\ \le \; & \mathcal{E}(\ell \circ f) + C_2 \sum_{j \in J_f} \left(\tau \bar{\epsilon}_j \|f_j - \hat{f}_j\|_{L_2(\Pi)} + \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) \\ & + C_3 L_* \sum_{j=1}^N \left(\bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \bar{\epsilon}_j^2 \|\hat{f}_j - f_j\|_{\mathcal{H}_j} \right) + C_3 L_* e^{-N} \\ \le \; & \mathcal{E}(\ell \circ f) + C_2 \sum_{j \in J_f} \left(\tau \bar{\epsilon}_j \|f_j - \hat{f}_j\|_{L_2(\Pi)} + \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) \\ & + C_3 L_* \sum_{j=1}^N \left(\bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} + \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) + C_3 L_* e^{-N} \end{aligned}$$

for some numerical constant $C_3 > 0$. By choosing a numerical constant B properly, τ can be made large enough so that $2C_3 L_* \le \tau C_1 \le \tau C_2$. Then, we have

$$\begin{aligned} & \mathcal{E}(\ell \circ \hat{f}) + \frac{1}{2} C_1 \left(\sum_{j \notin J_f} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \sum_{j=1}^N \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\ \le \; & \mathcal{E}(\ell \circ f) + 2C_2 \sum_{j \in J_f} \left(\tau \bar{\epsilon}_j \|f_j - \hat{f}_j\|_{L_2(\Pi)} + \tau^2 \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \right) + (C_2/2) \tau e^{-N}, \end{aligned} \tag{56}$$

which also implies

$$\begin{aligned}
&\mathcal{E}(\ell\circ\hat{f})+\frac{1}{2}C_1\left(\sum_{j=1}^{N}\tau\bar{\epsilon}_j\|\hat{f}_j-f_j\|_{L_2(\Pi)}+\sum_{j=1}^{N}\tau^2\bar{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j}\right)\\
\leq\;&\mathcal{E}(\ell\circ f)+\Big(2C_2+\frac{C_1}{2}\Big)\sum_{j\in J_f}\tau\bar{\epsilon}_j\|f_j-\hat{f}_j\|_{L_2(\Pi)}\\
+\;&2C_2\tau^2\sum_{j\in J_f}\bar{\epsilon}_j^2\|f_j\|_{\mathcal{H}_j}+(C_2/2)\tau e^{-N}. \qquad (57)
\end{aligned}$$

We first consider the case when

$$4C_2\sum_{j\in J_f}\tau\bar{\epsilon}_j\|f_j-\hat{f}_j\|_{L_2(\Pi)}\geq\mathcal{E}(\ell\circ f)+2C_2\sum_{j\in J_f}\tau^2\bar{\epsilon}_j^2\|f_j\|_{\mathcal{H}_j}+(C_2/2)\tau e^{-N}. \qquad (58)$$

Then (56) implies that

$$\mathcal{E}(\ell\circ\hat{f})+\frac{1}{2}C_1\left(\sum_{j\notin J_f}\tau\bar{\epsilon}_j\|\hat{f}_j\|_{L_2(\Pi)}+\sum_{j=1}^{N}\tau^2\bar{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j}\right)\leq 6C_2\sum_{j\in J_f}\tau\bar{\epsilon}_j\|f_j-\hat{f}_j\|_{L_2(\Pi)}, \qquad (59)$$

which yields

$$\sum_{j\notin J_f}\tau\bar{\epsilon}_j\|\hat{f}_j\|_{L_2(\Pi)}\leq\frac{12C_2}{C_1}\sum_{j\in J_f}\tau\bar{\epsilon}_j\|f_j-\hat{f}_j\|_{L_2(\Pi)}. \qquad (60)$$

Therefore, $(\hat{f}_1-f_1,\ldots,\hat{f}_N-f_N)\in\mathcal{K}_{J_f}^{(b)}$ with $b:=12C_2/C_1$. Using the definition of $\beta_b(J_f)$, it follows from (57), (58) and the assumption $C_1<1<C_2$ that

$$\begin{aligned}
&\mathcal{E}(\ell\circ\hat{f})+\frac{1}{2}C_1\left(\sum_{j=1}^{N}\tau\bar{\epsilon}_j\|\hat{f}_j-f_j\|_{L_2(\Pi)}+\sum_{j=1}^{N}\tau^2\bar{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j}\right)\\
\leq\;&\Big(6C_2+\frac{C_1}{2}\Big)\tau\beta_b(J_f)\|f-\hat{f}\|_{L_2(\Pi)}\\
\leq\;&7C_2\tau\beta_b(J_f)\left(\|f-f_*\|_{L_2(\Pi)}+\|f_*-\hat{f}\|_{L_2(\Pi)}\right).
\end{aligned}$$

Recall that for losses of quadratic type

$$\mathcal{E}(\ell\circ f)\geq m_*\|f-f_*\|_{L_2(\Pi)}^2\qquad\text{and}\qquad\mathcal{E}(\ell\circ\hat{f})\geq m_*\|\hat{f}-f_*\|_{L_2(\Pi)}^2. \qquad (61)$$

Then

$$\begin{aligned}
&\mathcal{E}(\ell\circ\hat{f})+\frac{1}{2}C_1\left(\sum_{j=1}^{N}\tau\bar{\epsilon}_j\|\hat{f}_j-f_j\|_{L_2(\Pi)}+\sum_{j=1}^{N}\tau^2\bar{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j}\right)\\
\leq\;&7\tau C_2 m_*^{-1/2}\beta_b(J_f)\left(\mathcal{E}^{1/2}(\ell\circ f)+\mathcal{E}^{1/2}(\ell\circ\hat{f})\right).
\end{aligned}$$

Using the fact that $ab \le (a^2+b^2)/2$, we get

$$7\tau C_2 m_*^{-1/2}\beta_b(J_f)\mathcal{E}^{1/2}(\ell\circ f) \le (49/2)\tau^2 C_2^2 m_*^{-1}\beta_b^2(J_f) + \frac{1}{2}\mathcal{E}(\ell\circ f), \tag{62}$$

and

$$7\tau C_2 m_*^{-1/2}\beta_b(J_f)\mathcal{E}^{1/2}(\ell\circ \hat{f}) \le (49/2)\tau^2 C_2^2 m_*^{-1}\beta_b^2(J_f) + \frac{1}{2}\mathcal{E}(\ell\circ \hat{f}). \tag{63}$$

Therefore,

$$\mathcal{E}(\ell\circ\hat{f}) + C_1\sum_{j=1}^N \tau\bar{\epsilon}_j\|\hat{f}_j\|_{L_2(\Pi)} + C_1\sum_{j=1}^N \tau^2\bar{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j} \le \mathcal{E}(\ell\circ f) + 100\tau^2 C_2^2 m_*^{-1}\beta_b^2(J_f). \tag{64}$$

We now consider the case when

$$4C_2\sum_{j\in J_f}\tau\bar{\epsilon}_j\|f_j-\hat{f}_j\|_{L_2(\Pi)} < \mathcal{E}(\ell\circ f) + 2C_2\sum_{j\in J_f}\tau^2\bar{\epsilon}_j^2\|f_j\|_{\mathcal{H}_j} + (C_2/2)\tau e^{-N}. \tag{65}$$

It is easy to derive from (57) that in this case

$$\begin{aligned}
&\mathcal{E}(\ell\circ\hat{f}) + \frac{1}{2}C_1\left(\sum_{j=1}^N \tau\bar{\epsilon}_j\|\hat{f}_j-f_j\|_{L_2(\Pi)} + \sum_{j=1}^N \tau^2\bar{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j}\right)\\
&\le \left(\frac{3}{2}+\frac{C_1}{8C_2}\right)\left(\mathcal{E}(\ell\circ f) + 2C_2\sum_{j\in J_f}\tau^2\bar{\epsilon}_j^2\|f_j\|_{\mathcal{H}_j} + (C_2/2)\tau e^{-N}\right).
\end{aligned} \tag{66}$$

Since $\beta_b(J_f) \ge \sqrt{\frac{A\log N}{n}}$ (see the comment after the definition of $\beta_b(J_f)$), we have

$$\tau e^{-N} \le \tau^2\sqrt{\frac{A\log N}{n}} \le \tau^2\beta_b^2(J_f),$$

where we also used the assumptions that $\log N \ge 2\log\log n$ and $A \ge 4$. Substituting this in (66) and then combining the resulting bound with (64) concludes the proof of (49) in the case when conditions (55) hold.

It remains to consider the case when (55) does not hold. The main idea is to show that in this case the right hand side of the oracle inequality is rather large while we still can control the left hand side, so, the inequality becomes trivial. To this end, note that, by the definition of $\hat{f}$, for some numerical constant c_1,

$$P_n(\ell\circ\hat{f}) + \sum_{j=1}^N\left(\tau\hat{\epsilon}_j\|\hat{f}_j\|_{L_2(\Pi_n)} + \tau^2\hat{\epsilon}_j^2\|\hat{f}_j\|_{\mathcal{H}_j}\right) \le n^{-1}\sum_{j=1}^n \ell(Y_j;0) \le c_1$$

(since the value of the penalized empirical risk at $\hat{f}$ is not larger than its value at $f = 0$ and, by the assumptions on the loss, $\ell(y, 0)$ is uniformly bounded by a numerical constant). The last equation implies that, on the event E defined earlier in the proof (see (51), (52)), the following bound holds:

$$\sum_{j=1}^{N} \frac{\tau}{C} \bar{\epsilon}_j \left(\frac{1}{C} \|\hat{f}_j\|_{L_2(\Pi)} - \bar{\epsilon}_j \|\hat{f}_j\|_{\mathcal{H}_j} \right) + \sum_{j=1}^{N} \frac{\tau^2}{C^2} \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \le c_1.$$

Equivalently,

$$\frac{\tau}{C^2} \sum_{j=1}^{N} \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \left(\frac{\tau^2}{C^2} - \frac{\tau}{C} \right) \sum_{j=1}^{N} \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \le c_1.$$

As soon as $\tau \ge 2C$, so that $\tau^2/C^2 - \tau/C \ge \tau^2/(2C^2)$, we have

$$\tau \sum_{j=1}^{N} \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \tau^2 \sum_{j=1}^{N} \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \le 2c_1 C^2. \tag{67}$$

Note also that, by the assumptions on the loss function,

$$\begin{aligned} &\mathcal{E}(\ell \circ \hat{f}) \le P(\ell \circ \hat{f}) \le \mathbb{E}\ell(Y; 0) + |P(\ell \circ \hat{f}) - P(\ell \circ 0)| \le c_1 + L_* \|\hat{f}\|_{L_2(\Pi)} \le \\ &c_1 + L_* \sum_{j=1}^{N} \|\hat{f}\|_{L_2(\Pi)} \le c_1 + 2c_1 C^2 L_* \frac{1}{\tau} \sqrt{\frac{n}{A \log N}}, \end{aligned} \tag{68}$$

where we used the Lipschitz condition on ℓ, and also bound (67) and the fact that $\bar{\epsilon}_j \ge \sqrt{A \log N / n}$ (by its definition).

Recall that we are considering the case when (55) does not hold. We will consider two cases: (a) when $e^N \le c_3$, where $c_3 \ge c_1$ is a numerical constant, and (b) when $e^N > c_3$. The first case is very simple since N and n are both upper bounded by a numerical constant (recall the assumption $\log N \ge 2 \log \log n$). In this case, $\beta_b(J_f) \ge \sqrt{\frac{A \log N}{n}}$ is bounded from below by a numerical constant. As a consequence of these observations, bounds (67) and (68) imply that

$$\mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\sum_{j=1}^{N} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \sum_{j=1}^{N} \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \le C_2 \tau^2 \beta_b^2(J_f)$$

for some numerical constant $C_2 > 0$. In the case (b), we have

$$\sum_{j=1}^{N} \bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \sum_{j=1}^{N} \bar{\epsilon}_j^2 \|\hat{f}_j - f_j\|_{\mathcal{H}_j} \ge e^N$$

and, in view of (67), this implies

$$\sum_{j=1}^{N} \bar{\epsilon}_j \|f_j\|_{L_2(\Pi)} + \sum_{j=1}^{N} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \geq e^N - c_1/2 \geq e^N/2.$$

So, either we have

$$\sum_{j=1}^{N} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \geq e^N/4, \quad \text{or} \quad \sum_{j=1}^{N} \bar{\epsilon}_j \|f_j\|_{L_2(\Pi)} \geq e^N/4.$$

Moreover, in the second case, we also have

$$\sum_{j=1}^{N} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \geq \sqrt{\frac{A \log N}{n}} \sum_{j=1}^{N} \bar{\epsilon}_j \|f_j\|_{L_2(\Pi)} \geq (e^N/4) \sqrt{\frac{A \log N}{n}}.$$

In both cases we can conclude that, under the assumption that $\log N \geq 2 \log \log n$ and $e^N > c_3$ for a sufficiently large numerical constant c_3,

$$\mathcal{E}(\ell \circ \hat{f}) + \sum_{j=1}^{N} \left(\tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \leq$$

$$c_1 + 2c_1 C^2 L_* \frac{1}{\tau} \sqrt{\frac{n}{A \log N}} + 2c_1 C^2 \leq \frac{\tau^2 e^N}{4} \sqrt{\frac{A \log N}{n}} \leq \tau^2 \sum_{j \in J_f} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j}.$$

Thus, in both cases (a) and (b), the following bound holds:

$$\mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\sum_{j=1}^{N} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \sum_{j=1}^{N} \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \leq C_2 \tau^2 \left(\sum_{j \in J_f} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} + \beta_b^2(J_f) \right). \quad (69)$$

To complete the proof, observe that

$$\begin{aligned}
&\mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\sum_{j=1}^{N} \tau \bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \sum_{j=1}^{N} \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\
&\leq \mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\sum_{j=1}^{N} \tau \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} + \sum_{j=1}^{N} \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) + C_1 \sum_{j \in J_f} \tau \bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} \\
&\leq C_2 \tau^2 \left(\sum_{j \in J_f} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} + \beta_b^2(J_f) \right) + C_2 \sum_{j \in J_f} \tau \bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)}. \quad (70)
\end{aligned}$$

Note also that, by the definition of $\beta_b(J_f)$, for all $b > 0$,

$$\begin{aligned}
&\sum_{j \in J_f} \tau \bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} \leq \tau \beta_b(J_f) \left\| \sum_{j \in J_f} (\hat{f}_j - f_j) \right\|_{L_2(\Pi)} \leq \\
&\tau \beta_b(J_f) \|\hat{f} - f\|_{L_2(\Pi)} + \tau \beta_b(J_f) \sqrt{\frac{n}{A \log N}} \sum_{j \notin J_f} \bar{\epsilon}_j \|\hat{f}_j\|_{L_2(\Pi)} \leq \\
&\tau \beta_b(J_f) \|\hat{f} - f\|_{L_2(\Pi)} + \tau \beta_b(J_f) \frac{2c_1 C^2}{\tau} \sqrt{\frac{n}{A \log N}}. \quad (71)
\end{aligned}$$

where we used the fact that, for all j, $\bar{\varepsilon}_j \geq \sqrt{\frac{A \log N}{n}}$ and also bound (67). By an argument similar to (61)-(64), it is easy to deduce from the last bound that

$$\begin{aligned} & C_2 \sum_{j \in J_f} \tau \bar{\varepsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} \\ & \leq \frac{3}{2} \frac{C_2^2 \tau^2}{m_*} \beta_b^2(J_f) + \frac{1}{2} \mathcal{E}(\ell \circ \hat{f}) + \frac{1}{2} \mathcal{E}(\ell \circ f) + \frac{2c_1^2 C^4}{\tau^2} \frac{n}{A \log N}. \end{aligned} \tag{72}$$

Substituting this in bound (70), we get

$$\begin{aligned} & \frac{1}{2} \mathcal{E}(\ell \circ \hat{f}) + C_1 \left(\sum_{j=1}^N \tau \bar{\epsilon}_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \sum_{j=1}^N \tau^2 \bar{\epsilon}_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \\ & \leq C_2 \tau^2 \left(\sum_{j \in J_f} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} + \beta_b^2(J_f) \right) \\ & + \frac{3}{2} \frac{C_2^2 \tau^2}{m_*} \beta_b^2(J_f) + \frac{1}{2} \mathcal{E}(\ell \circ f) + \frac{2c_1^2 C^4}{\tau^2} \frac{n}{A \log N} \\ & \leq \frac{1}{2} \mathcal{E}(\ell \circ f) + C_2' \tau^2 \left(\sum_{j \in J_f} \bar{\epsilon}_j^2 \|f_j\|_{\mathcal{H}_j} + \frac{\beta_b^2(J_f)}{m_*} \right) + \frac{2c_1^2 C^2}{\tau^2} \frac{n}{A \log N}, \end{aligned} \tag{73}$$

with some numerical constant C_2'. It is enough now to observe (considering again the cases (a) and (b), as it was done before), that either the last term is upper bounded by $\sum_{j \in J_f} \bar{\varepsilon}_j \|f_j\|_{\mathcal{H}_j}$, or it is upper bounded by $\beta_b^2(J_f)$, to complete the proof. ∎

Now, to derive Theorem 2, it is enough to check that, for a numerical constant $c > 0$,

$$\beta_b(J_f) \leq \left(\sum_{j \in J_f} \bar{\epsilon}_j^2 \right)^{1/2} \beta_{2,\infty}(J_f) \leq c \left(\sum_{j \in J_f} \breve{\epsilon}_j^2 \right)^{1/2} \beta_{2,\infty}(J_f)$$

which easily follows from the definitions of β_b and $\beta_{2,\infty}$. Similarly, the proof of Theorem 3 follows from the fact that, under the assumption that $\Lambda^{-1} \leq \frac{\breve{\epsilon}_j}{\bar{\epsilon}} \leq \Lambda$, we have $\mathcal{K}_J^{(b)} \subset K_J^{(b')}$, where $b' = c\Lambda^2 b$, c being a numerical constant. This easily implies the bound $\beta_b(J_f) \leq c_1 \Lambda \beta_{2,b'}(J_f) \sqrt{d(f)} \breve{\epsilon}$, where c_1 is a numerical constant.

5 Bounding the Empirical Process

We now proceed to prove Lemma 9 that was used to bound $\left| (P_n - P) \left(\ell \circ \hat{f} - \ell \circ f \right) \right|$. To this end, we begin with a fixed pair (Δ_-, Δ_+). Throughout the proof, we write $R := R_D^*$. By

Talagrand's concentration inequality, with probability at least $1 - e^{-t}$

$$\begin{aligned}\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)} |(P_n - P)(\ell\circ g - \ell\circ f)| &\leq 2\Bigg(\mathbb{E}\left[\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)} |(P_n - P)(\ell\circ g - \ell\circ f)|\right] \\ &\quad + \|\ell\circ g - \ell\circ f\|_{L_2(P)}\sqrt{\frac{t}{n}} + \|\ell\circ g - \ell\circ f\|_{L_\infty}\frac{t}{n}\Bigg).\end{aligned}$$

Now note that

$$\begin{aligned}&\|\ell\circ g - \ell\circ f\|_{L_2(P)} \leq L_*\|g - f\|_{L_2(\Pi)} \\ &\leq L_*\sum_{j=1}^N \|g_j - f_j\|_{L_2(\Pi)} \leq L_*\left(\min_j \bar{\epsilon}_j\right)^{-1}\sum_{j=1}^N \bar{\epsilon}_j\|g_j - f_j\|_{L_2(\Pi)},\end{aligned}$$

where we used the fact that the Lipschitz constant of the loss ℓ on the range of functions from $\mathcal{G}(\Delta_-, \Delta_+, R)$ is bounded by L_*. Together with the fact that $\bar{\epsilon}_j \geq (A\log N/n)^{1/2}$ for all j, this yields

$$\|\ell\circ g - \ell\circ f\|_{L_2(P)} \leq L_*\sqrt{\frac{n}{A\log N}}\Delta_-. \tag{74}$$

Furthermore,

$$\begin{aligned}\|\ell\circ g - \ell\circ f\|_{L_\infty} &\leq L_*\|g - f\|_{L_\infty} \\ &\leq L_*\sum_{j=1}^N \|g_j - f_j\|_{\mathcal{H}_j} \\ &\leq L_*\frac{n}{A\log N}\Delta_+.\end{aligned}$$

In summary, we have

$$\begin{aligned}&\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)} |(P_n - P)(\ell\circ g - \ell\circ f)| \\ \leq\ & 2\Bigg(\mathbb{E}\left[\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)} |(P_n - P)(\ell\circ g - \ell\circ f)|\right] + \\ &\quad L_*\Delta_-\sqrt{\frac{t}{A\log N}} + L_*\Delta_+\frac{t}{n}\frac{n}{A\log N}\Bigg).\end{aligned}$$

Now, by symmetrization inequality,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)} |(P_n - P)(\ell\circ g - \ell\circ f)|\right] \leq 2\mathbb{E}\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)} |R_n(\ell\circ g - \ell\circ f)|. \tag{75}$$

An application of Rademacher contraction inequality further yields

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}|(P_n-P)(\ell\circ g-\ell\circ f)|\right]\le CL_*\mathbb{E}\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}|R_n(g-f)| \tag{76}$$

where $C>0$ is a numerical constant (again, it was used here that the Lipschitz constant of the loss ℓ on the range of functions from $\mathcal{G}(\Delta_-,\Delta_+,R)$ is bounded by L_*). Applying Talagrand's concentration inequality another time, we get that with probability at least $1-e^{-t}$

$$\begin{aligned}&\mathbb{E}\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}|R_n(g-f)|\\ \le\ & C\left(\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}|R_n(g-f)|+\Delta_-\sqrt{\frac{t}{A\log N}}+\Delta_+\frac{t}{n}\frac{n}{A\log N}\right)\end{aligned}$$

for some numerical constant $C>0$.

Recalling the definition of $\check{\epsilon}_j:=\check{\epsilon}(K_j)$, we get

$$|R_n(h_j)|\le\check{\epsilon}_j\|h_j\|_{L_2(\Pi)}+\check{\epsilon}_j^2\|h_j\|_{\mathcal{H}_j},\ h_j\in\mathcal{H}_j \tag{77}$$

Hence, with probability at least $1-2e^{-t}$ and with some numerical constant $C>0$

$$\begin{aligned}&\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}|(P_n-P)(\ell\circ g-\ell\circ f)|\\ \le\ & CL_*\left(\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}|R_n(g-f)|+\Delta_-\sqrt{\frac{t}{A\log N}}+\Delta_+\frac{t}{n}\frac{n}{A\log N}\right)\\ \le\ & CL_*\left(\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}\sum_{j=1}^N|R_n(g_j-f_j)|+\Delta_-\sqrt{\frac{t}{A\log N}}+\Delta_+\frac{t}{n}\frac{n}{A\log N}\right)\\ \le\ & CL_*\Bigg(\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}\sum_{j=1}^N\left(\check{\epsilon}_j\|g_j-f_j\|_{L_2(\Pi)}+\check{\epsilon}_j^2\|g_j-f_j\|_{\mathcal{H}_j}\right)\\ &\qquad+\Delta_-\sqrt{\frac{t}{A\log N}}+\Delta_+\frac{t}{n}\frac{n}{A\log N}\Bigg).\end{aligned}$$

Using (46), $\check{\epsilon}_j$ can be upper bounded by $c\bar{\epsilon}_j$ with some numerical constant $c>0$ on an event E of probability at least $1-N^{-A/2}$. Therefore, the following bound is obtained:

$$\begin{aligned}&\sup_{g\in\mathcal{G}(\Delta_-,\Delta_+,R)}|(P_n-P)(\ell\circ g-\ell\circ f)|\\ \le\ & CL_*\left(\Delta_-+\Delta_++\Delta_-\sqrt{\frac{t}{A\log N}}+\Delta_+\frac{t}{n}\frac{n}{A\log N}\right).\end{aligned}$$

It holds on the event $E \cap F(\Delta_-, \Delta_+, t)$, where $\mathbb{P}(F(\Delta_-, \Delta_+, t)) \geq 1 - 2e^{-t}$.

We will now choose $t = A \log N + 4 \log N + 4 \log(2/\log 2)$ and obtain a bound that holds uniformly over

$$e^{-N} \leq \Delta_- \leq e^N \qquad \text{and} \qquad e^{-N} \leq \Delta_+ \leq e^N. \tag{78}$$

To this end, consider

$$\Delta_j^- = \Delta_j^+ := 2^{-j}. \tag{79}$$

For any Δ_j^- and Δ_k^+ satisfying (78), we have

$$\begin{aligned} &\sup_{g \in \mathcal{G}(\Delta_j^-, \Delta_k^+, R)} |(P_n - P)(\ell \circ g - \ell \circ f)| \\ \leq \;& CL_* \left(\Delta_j^- + \Delta_k^+ + \Delta_j^- \sqrt{\frac{t}{A \log N}} + \Delta_k^+ \frac{t}{n} \frac{n}{A \log N} \right) \end{aligned}$$

on the event $E \cap F(\Delta_j^-, \Delta_k^+, t)$. Therefore, simultaneously for all Δ_j^- and Δ_k^+ satisfying (78), we have

$$\begin{aligned} &\sup_{g \in \mathcal{G}(\Delta_j^-, \Delta_k^+, R)} |(P_n - P)(\ell \circ g - \ell \circ f)| \\ \leq \;& CL_* \left(\Delta_j^- + \Delta_k^+ + \Delta_j^- \sqrt{\frac{A \log N + 4 \log N + 4 \log(2/\log 2)}{A \log N}} \right. \\ & \qquad \left. + \Delta_+ \frac{A \log N + 4 \log N + 4 \log(2/\log 2)}{n} \frac{n}{A \log N} \right) \end{aligned}$$

on the event $E' := E \bigcap \left(\bigcap_{j,k} F(\Delta_j^-, \Delta_k^+, t) \right)$. The last intersection is over all j, k such that conditions (78) hold for Δ_j^-, Δ_k^+. The number of the events in this intersection is bounded by $(2/\log 2)^2 N^2$. Therefore,

$$\mathbb{P}(E') \geq 1 - (2/\log 2)^2 N^2 \exp\left(-A \log N - 4 \log N - 4 \log(2/\log 2)\right) - \mathbb{P}(E) \geq 1 - 2N^{-A/2}. \tag{80}$$

Using monotonicity of the functions of Δ_-, Δ_+ involved in the inequalities, the bounds can be extended to the whole range of values of Δ_-, Δ_+ satisfying (78), so, with probability at least $1 - 2N^{-A/2}$ we have for all such Δ_-, Δ_+

$$\sup_{g \in \mathcal{G}(\Delta_-, \Delta_+, R)} |(P_n - P)(\ell \circ g - \ell \circ f)| \leq CL_* (\Delta_- + \Delta_+). \tag{81}$$

If $\Delta_- \leq e^{-N}$, or $\Delta_+ \leq e^{-N}$, it follows by monotonicity of the left hand side that with the same probability

$$\sup_{g \in \mathcal{G}(\Delta_-, \Delta_+, R)} |(P_n - P)(\ell \circ g - \ell \circ f)| \leq C L_* \left(\Delta_- + \Delta_+ + e^{-N} \right), \tag{82}$$

which completes the proof. ∎

Acknowledgment. The authors are thankful to the referees for a number of helpful suggestions. The first author is thankful to Evarist Giné for useful conversations about the paper.

References

[1] Aronszajn, N. (1950), Theory of reproducing kernels, *Trans. Am. Math. Soc.*, **68**, 337-404.

[2] Bach, F. (2008), Consistency of the group Lasso and multiple kernel learning, *Journal of Machine Learning Research*, 9, 1179-1225.

[3] Bickel, P., Ritov, Y. and Tsybakov, A. (2009), Simultaneous analysis of Lasso and Dantzig selector, *Annals of Statistics*, 37, 4, 1705-1732.

[4] Bousquet, O. and Herrmann, D. (2003), On the complexity of learning the kernel matrix, In: *Advances in Neural Information Processing Systems 15*, 415-422.

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

[6] Bousquet, O. (2002), A Bennett concentration inequality and its applications to suprema of empirical processes, *C.R. Acad. Sci. Paris*, 334, 495-500.

[7] Crammer, K., Keshet, J. and Singer, Y. (2003), Kernel design using boosting, In: *Advances in Neural Information Processing Systems 15*, 553-560.

[8] Koltchinskii, V. (2008), Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems, *Lecture Notes for Ecole d'Eté de Probabilités de Saint-Flour.*

[9] Koltchinskii, V. (2009a), Sparsity in penalized empirical risk minimization, *Ann. Inst. H. Poincaré (B), Probabilités et Statistiques,* 45, 1, 7-57.

[10] Koltchinskii, V. (2009b), The Dantzig selector and sparsity oracle inequalities, *Bernoulli*, 15(3), 799-828.

[11] Koltchinskii, V. (2009c), Sparse recovery in convex hulls via entropy penalization, *Annals of Statistics*, 37(3), 1332-1359.

[12] Koltchinskii, V. and Yuan, M. (2008), Sparse recovery in large ensembles of kernel machines, In: *Proceedings of 19th Annual Conference on Learning Theory* (COLT 2008), 229-238.

[13] 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.

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

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

[16] Meier, L., van de Geer, S. and Bühlmann, P. (2009), High-dimensional additive modeling, *Annals of Statistics*, 37, 3779-3821.

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

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

[19] Raskutti, G., Wainwright, M. and Yu, Bin (2009), Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness, *Advances in Neural Information Processing Systems* (NIPS 22).

[20] Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2008), SpAM: sparse additive models, *Advances in Neural Information Processing Systems* (NIPS 20), 1201-1208.

[21] 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.

[22] Talagrand, M. (1996), New concentration inequalities for product measures. *Invent. Math.*, 126, 505-563.

[23] Tsybakov, A.B. (2009), *Introduction to Nonparametric Estimation*, Springer-Verlag, New York.

[24] van der Vaart, A.W. and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes. With Applications to Statistics*, Springer-Verlag, New York.