Sparse Recovery in Convex Hulls of Infinite Dictionaries

Vladimir Koltchinskii
Georgia Institute of Technology

vlad@math.gatech.edu

Stas Minsker
Georgia Institute of Technology

sminsker@math.gatech.edu

Abstract

Let S be an arbitrary measurable space, $T \subset \mathbb{R}$ and (X, Y) be a random couple in $S \times T$ with unknown distribution P. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. copies of (X, Y). Denote by P_n the empirical distribution based on the sample $(X_i, Y_i), i = 1, \ldots, n$. Let $\mathcal{H}$ be a set of uniformly bounded functions on S. Suppose that $\mathcal{H}$ is equipped with a σ-algebra and with a finite measure μ. Let $\mathbb{D}$ be a convex set of probability densities with respect to μ. For $\lambda \in \mathbb{D}$, define the mixture $f_\lambda(\cdot) = \int\limits_{\mathcal{H}} h(\cdot)\lambda(h)d\mu(h)$. Given a loss function $\ell : T \times \mathbb{R} \mapsto \mathbb{R}$ such that, for all $y \in T$, $\ell(y, \cdot)$ is convex, denote $(\ell \bullet f)(x, y) = \ell(y; f(x))$. We study the following penalized empirical risk minimization problem

$$\hat{\lambda}_\varepsilon := \underset{\lambda \in \mathbb{D}}{\operatorname{argmin}} \left[P_n(\ell \bullet f_\lambda) + \varepsilon \int \lambda \log \lambda d\mu \right]$$

along with its distribution dependent version

$$\lambda_\varepsilon := \underset{\lambda \in \mathbb{D}}{\operatorname{argmin}} \left[P(\ell \bullet f_\lambda) + \varepsilon \int \lambda \log \lambda d\mu \right].$$

We prove that the "approximate sparsity" of λ_ε implies the "approximate sparsity" of $\hat{\lambda}_\varepsilon$ and study connections between the sparsity and the excess risk of empirical solutions $\hat{\lambda}_\varepsilon$.

1 Introduction

Sparsity phenomena in empirical risk minimization over linear spans or convex hulls of large finite dictionaries have been extensively studied in the recent years (see, e.g., [MPTJ07], [Kol09b], [BRT09] and references therein). In this paper, our goal is to extend some of these results to the case of empirical risk minimization over convex hulls of infinite dictionaries which is a standard framework in machine learning (for instance, in large margin classification, the dictionaries are often infinite families of functions such as decision stumps, decision trees or subsets of reproducing kernel Hilbert spaces).

Let S be a measurable space, $T \subset \mathbb{R}$ be a Borel set and (X, Y) be a random couple in $S \times T$ with unknown distribution P. The marginal distribution of X will be denoted by Π. Let $(X_1, Y_1), \ldots,$ (X_n, Y_n) be the training data consisting of n i.i.d. copies of (X, Y). In what follows, we will denote by P_n the empirical

[*] Partially supported by NSF grants MSPA-MCS-0624841, DMS-0906880 and CCF-0808863

distribution based on a given sample of n training examples. Similarly, Π_n will denote the empirical measure based on the sample $(X_1, \ldots, X_n)$. The integrals with respect to P and P_n are denoted by

$$Pg := \mathbb{E}g(X,Y), \quad P_n g := \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i).$$

Similar notations will be used for Π, Π_n and other measures. Let $\ell(y, \cdot)$ be the loss function such that for all $y \in T$, $\ell(y, \cdot)$ is convex. For a function $f : S \mapsto \mathbb{R}$, let $(\ell \bullet f)(x,y) := \ell(y, f(x))$. A *dictionary* is a given family $\mathcal{H}$ of measurable functions $h : S \mapsto [-1, 1]$. Assume that $\mathcal{H}$ is equipped with a σ-algebra and with a finite measure μ. In what follows, the complexity of the dictionary $\mathcal{H}$ will be characterized in terms of $L_2(\Pi)$ and $L_2(\Pi_n)$ covering numbers. Suppose Λ is a probability measure on $\mathcal{H}$ absolutely continuous with respect to μ with $\lambda = \frac{d\Lambda}{d\mu}$. The (negative) entropy $H(\lambda)$ is defined as $H(\lambda) := \int\limits_{\mathcal{H}} \lambda(h)\log\lambda(h) d\mu(h)$. In what follows, we consider only densities with finite entropies. Let f_λ denote the mixture of the functions from the dictionary $\mathcal{H}$ with respect to $\lambda : f_\lambda(\cdot) := \int\limits_{\mathcal{H}} h(\cdot)\lambda(h)\mu(dh)$. The excess risk $\mathcal{E}(f)$ of a function f is defined as

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

For simplicity, we assume throughout the paper that $\inf\limits_{g:S\mapsto R} P\left(\ell \bullet g\right)$ is attained at some uniformly bounded function f_* (the infimum is taken over all measurable functions).

Let $\mathbb{D}$ be a convex set of probability densities on $\mathcal{H}$. We will assume that, for all $\lambda \in \mathbb{D}$, $\lambda \log \lambda \in L_1(\mu)$, so, the entropy $H(\lambda)$ is finite. Consider the following penalized risk minimization problem

$$\lambda_\varepsilon := \mathrm{argmin}_{\lambda\in\mathbb{D}} F(\lambda), \quad F(\lambda) := P(\ell \bullet f_\lambda) + \varepsilon H(\lambda) \tag{1.1}$$

together with its empirical version:

$$\hat{\lambda}_\varepsilon := \mathrm{argmin}_{\lambda\in\mathbb{D}} F_n(\lambda), \quad F_n(\lambda) := P_n(\ell \bullet f_\lambda) + \varepsilon H(\lambda). \tag{1.2}$$

Note that, due to the convexity of the loss, of the negative entropy and of the set $\mathbb{D}$, both (1.1) and (1.2) are convex optimization problems. We will use the notations $\Lambda_\varepsilon, \hat{\Lambda}_\varepsilon$ for the probability measures with densities $\lambda_\varepsilon, \hat{\lambda}_\varepsilon$, respectively.

Our first aim is to study problem (1.1) and to bound the approximation error $\mathcal{E}(f_{\lambda_\varepsilon})$ of its solution, which, for the losses of quadratic type, is equivalent to bounding the $L_2(\Pi)$-approximation error $\|f_{\lambda_\varepsilon} - f_*\|^2_{L_2(\Pi)}$. We show that the size of this error can be controlled in terms of the approximation error of oracle solutions $\lambda \in \mathbb{D}$ that are "sparse" in the sense that they are concentrated on a "small" set of functions $\mathcal{H}' \subset \mathcal{H}$ and, at the same time, possess some regularity properties. Moreover, we show that if there exist "sparse" oracles providing good approximation of the target function, then solutions λ_ε of problem (1.1) are "approximately sparse" in the sense that they "concentrate" on the support of "sparse" oracles.

Next, we study the relationship between the problems (1.1) and (1.2). We show that the "approximate sparsity" of the true penalized solution λ_ε implies that the corresponding empirical solution $\hat{\lambda}_\varepsilon$ possesses the same property with a high probability. More precisely, if there exists a measurable set $\mathcal{H}' \subset \mathcal{H}$ such that $\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$ is small and, at the same time, there exists a subspace $L \subset L_2(\Pi)$ of small dimension d that provides a good $L_2(\Pi)$-approximation of the functions from the set $\mathcal{H}'$, we will show that in this case, with a high probability, the empirical solution $\hat{\lambda}_\varepsilon$ is also approximately supported on the same set $\mathcal{H}'$ in the sense that $\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$ is small. Thus, both the empirical solution $\hat{\lambda}_\varepsilon$ and the true solution λ_ε follow the same "sparsity pattern": they are concentrated on the same set of functions $\mathcal{H}'$ which can be well approximated by a linear subspace of small dimension. We also obtain probabilistic bounds on the random error $\left|\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) - \mathcal{E}(f_{\lambda_\varepsilon})\right|$, or, equivalently, the $L_2(\Pi)$-random error $\|f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}\|^2_{L_2(\Pi)}$, in terms of characteristics of the sparsity of the problem such as the measure $\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$ and the dimension d of the approximating space L. At the same time, we derive upper bounds on the Kullback-Leibler type distance between $\hat{\lambda}_\varepsilon$ and λ_ε.

The idea of using entropy for complexity penalization is not new in machine learning (see, e.g., [Zha01]). An approach to sparse recovery based on entropy penalization has been studied in the case of finite dictionaries $\mathcal{H}$ (see [Kol09a], [Kol08]). As in these papers, the fact that the penalty is strictly convex allows us to study the random error independently of the approximation error, but geometric parameters of the dictionary needed to control these two errors are not quite the same. ℓ_1-type penalization in the case of infinite dictionaries was suggested in [RSSZ07] (however, sharp generalization error bounds were not studied in this paper).

2 Preliminaries

Assumptions on the loss. Assume that for all $y \in T$, $\ell(y, \cdot)$ is a convex twice differentiable function, ℓ''_u [here and in what follows the derivatives of the loss are taken with respect to the second variable] is uniformly bounded in $T \times [-1, 1]$ and $\sup_{y \in T} \ell(y; 0) < +\infty$, $\sup_{y \in T} |\ell'_u(y; 0)| < +\infty$. It will be also assumed that

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

In what follows, the loss functions ℓ satisfying the above assumptions will be called **the losses of quadratic type.** In particular, the assumptions imply that

$$m\|f_\lambda - f_*\|^2_{L_2(\Pi)} \leq \mathcal{E}\left(f_\lambda\right) \leq M\|f_\lambda - f_*\|^2_{L_2(\Pi)},$$

where $M := \frac{1}{2} \sup\limits_{y,u} \ell''_u(y, u)$. Moreover, the following simple proposition also holds for such losses:

Proposition 1 *There exists a constant $C > 0$ depending only on ℓ such that for all $\lambda, \bar{\lambda} \in \mathbb{D}$,*

$$|\mathcal{E}(f_{\bar{\lambda}}) - \mathcal{E}(f_\lambda)| \leq C\left[\|f_{\bar{\lambda}} - f_\lambda\|^2_{L_2(\Pi)} \bigvee \sqrt{\mathcal{E}(f_\lambda)}\|f_{\bar{\lambda}} - f_\lambda\|_{L_2(\Pi)}\right].$$

Proof: See the proof of Theorem 3 in [Kol09a]. □

Common examples of such loss functions include the usual quadratic loss $\ell(y, u) = (y - u)^2$ used in the regression setting (with T being a bounded interval of $\mathbb{R}$) as well as the exponential loss $\ell(y, u) = e^{-yu}$ and the logit loss $\ell(y, u) = \log_2\left(1 + e^{-yu}\right)$ used in large margin classification (with $T = \{-1, 1\}$).

Existence of solutions. We provide sufficient conditions of existence of solutions of problems (1.1) and (1.2). Recall that all the densities λ in question have finite entropy.

Proposition 2 *Problems (1.1), (1.2) have unique solutions in every convex weakly compact subset $\mathbb{D}$ of L_p, $p \geq 1$.*

Proof: First, we show that the entropy functional $H(\lambda) := \int\limits_{\mathcal{H}} \lambda \log \lambda d\mu$ is lower semi-continuous in $L_p(\mu)$, $p \geq 1$. Indeed, the functional is lower semi-continuous iff the level sets $\mathcal{L}_t = \{\lambda : H(\lambda) \leq t\}$ are closed. Suppose $\lambda_n \in \mathcal{L}_t$, $\lambda_n \to \lambda_0$ in L_p. We can extract the subsequence λ_{n_k} converging to λ_0 pointwise. Noting that $s\log(s) + e^{-1} \geq 0$ and applying the Fatou lemma to the sequence $\{\lambda_{n_k} \log(\lambda_{n_k})\}$, we derive the result. Next, under the assumptions on the loss, $F(\lambda)$ is convex, bounded from below and lower semi-continuous(continuity of the risk $P(\ell \bullet f_\lambda)$ follows from the uniform boundedness of the dictionary and integral Minkowski inequality), so that the level sets $\mathcal{L}_t = \{\lambda : \ F(\lambda) \le t\}$ are closed and convex. Mazur's theorem implies that they are also closed in weak topology, so F is weakly lower semi-continuous. Given a minimizing sequence λ_n, we can extract a weakly convergent subsequence $\lambda_{n_k} \longrightarrow \lambda_\infty$, and conclude that $\lambda_\infty \in \mathbb{D}, \ -\infty < F(\lambda_\infty) \le \liminf_{k\to\infty} F\left(\lambda_{n_k}\right)$. Convexity of the set $\mathbb{D}$ and strict convexity of the functional F implies the uniqueness of the solution of (1.1). Replacing F by F_n, we get similar statements for (1.2). □

We will assume throughout the paper that $\mathbb{D}$ is a convex set of probability densities such that $\lambda \log \lambda \in L_1(\mu), \ \lambda \in \mathbb{D}$ and solutions of the problems (1.1), (1.2) exist in $\mathbb{D}$.

Differentiability of the risk and of the entropy. To derive necessary conditions of the minima in the optimization problems (1.1), (1.2), we have to study differentiability properties of the functions involved. For $G : \mathbb{D} \mapsto \mathbb{R}$, $\lambda \in \mathbb{D}$ and ν such that $\bar\lambda := \lambda + t_0\nu \in \mathbb{D}$ for some $t_0 > 0$, denote

$$DG(\lambda;\nu) := \lim_{t\downarrow 0} \frac{G(\lambda + t\nu) - G(\lambda)}{t},$$

provided that the limit exists. $DG(\lambda;\nu)$ is the (directional) derivative of G at point λ in the direction ν.

First note that, under our assumptions on the loss function ℓ, both the true risk $\mathbb{D} \ni \lambda \mapsto P(\ell \bullet f_\lambda) =: L(\lambda)$ and the empirical risk $\mathbb{D} \ni \lambda \mapsto P(\ell \bullet f_\lambda) := L_n(\lambda)$ have directional derivatives at any point $\lambda \in \mathbb{D}$ in the direction of any other point $\bar\lambda = \lambda + t_0\nu \in \mathbb{D}, t_0 > 0$. Moreover, the following formulas hold:

$$DL(\lambda,\nu) = P(\ell' \bullet f_\lambda)f_\nu \text{ and } DL_n(\lambda,\nu) = P_n(\ell' \bullet f_\lambda)f_\nu. \tag{2.1}$$

Let λ_1, λ_2 be two densities from $\mathbb{D}$ and Λ_1, Λ_2 the corresponding probability measures on $\mathcal{H}$. Denote by $K(\lambda_1|\lambda_2) := \int_{\mathcal{H}} \log \frac{\lambda_1}{\lambda_2}\lambda_1 d\mu$ the Kullback-Leibler divergence between λ_1 and λ_2 and let $K(\lambda_1,\lambda_2) := K(\lambda_1|\lambda_2) + K(\lambda_2|\lambda_1)$ be the symmetrized Kullback-Leibler divergence.

Proposition 3 *For all $\lambda_1, \lambda_2 \in \mathbb{D}$, $\tau \in (0,1)$ and measurable $\mathcal{H}' \subset \mathcal{H}$*

$$DH(\lambda_1 + \tau(\lambda_2 - \lambda_1); \lambda_2 - \lambda_1) = \int_{\mathcal{H}} \log(\lambda_1 + \tau(\lambda_2 - \lambda_1))(\lambda_2 - \lambda_1)d\mu, \tag{2.2}$$

$$K(\lambda_1,\lambda_2) = \lim_{t\to 0} \int_{\mathcal{H}} \log \frac{(1-t)\lambda_1 + t\lambda_2}{t\lambda_1 + (1-t)\lambda_2}(\lambda_1 - \lambda_2)d\mu, \tag{2.3}$$

$$\Lambda_1(\mathcal{H}\setminus\mathcal{H}') \le 2\Lambda_2(\mathcal{H}\setminus\mathcal{H}') + K(\lambda_1,\lambda_2). \tag{2.4}$$

The proof of (2.2) and (2.3) is based on a convexity argument and on the monotone convergence theorem. To show (2.4), note that by the well known inequality relating the Kullback-Leibler and Hellinger distances, for all $\mathcal{H}' \subset \mathcal{H}$

$$K(\lambda_1,\lambda_2) \ge 2\int\limits_{\mathcal{H}} \left(\sqrt{\lambda_1} - \sqrt{\lambda_2}\right)^2 \ge 2 \int\limits_{\mathcal{H}\setminus\mathcal{H}'} \left(\sqrt{\lambda_1} - \sqrt{\lambda_2}\right)^2 \ge$$

$$\ge 2 \int\limits_{\mathcal{H}\setminus\mathcal{H}'} \left(\lambda_1 + \lambda_2 - \frac{\lambda_1}{2} - 2\lambda_2\right) = \Lambda_1(\mathcal{H}\setminus\mathcal{H}') - 2\Lambda_2(\mathcal{H}\setminus\mathcal{H}').$$

3 Bounding approximation error

In what follows, our goal is to compare the excess risk of the estimator λ_ε with the excess risk of "oracles" $\lambda \in \mathbb{D}$. Define a cone $\mathbb{K} := \{c(\lambda_1 - \lambda_2) : \ \lambda_1, \lambda_2 \in \mathbb{D}, \ c \in \mathbb{R}\}$ and for $w \in L_2(\mu)$, define **the alignment**

coefficient $\gamma(w)$ to be

$$\gamma(w) := \sup\left\{\langle w, u\rangle_{L_2(\mu)} \ : \ \|f_u\|_{L_2(\Pi)} = 1, \ u \in \mathbb{K}\right\}.$$

It is easy to see that, for all constants $c \in \mathbb{R}$, $\gamma(w+c) = \gamma(w)$. Denote by K the Gram operator of the dictionary, i.e.,

$$(Ku)(h) = \int_{\mathcal{H}} \langle h, g\rangle_{L_2(\Pi)} u(g)\mu(dg), \ h \in L_2(\Pi),$$

which is a self-adjoint nonnegatively definite operator. Clearly,

$$\|f_u\|^2_{L_2(\Pi)} = \langle Ku, u\rangle_{L_2(\mu)} = \langle K^{\frac{1}{2}}u, K^{\frac{1}{2}}u\rangle_{L_2(\mu)}$$

and it is easy to see that for all $w \in \mathrm{Im}(K^{1/2})$, $\gamma(w) \le \|K^{-\frac{1}{2}}w\|_{L_2(\mu)}$. Roughly speaking, $\gamma(w)$ is small if the function w is "properly aligned" with eigenspaces of the Gram operator K of the dictionary (say, it belongs to the linear span of eigenspaces corresponding to large enough eigenvalues of K).

The following theorem shows that the approximation error $\mathcal{E}(f_{\lambda_\varepsilon})$ of the true solution λ_ε can be controlled by the approximation error $\mathcal{E}(f_\lambda)$ of "oracles" $\lambda \in \mathbb{D}$ up to an error term of the size $\gamma^2(\log\lambda)\varepsilon^2$. Moreover, it also shows that, for any oracle $\lambda \in \mathbb{D}$, f_{λ_ε} belongs to an $L_2(\Pi)$ ball around f_λ whose radius is, up to a constant, $\|f_\lambda - f_*\|_{L_2(\Pi)} \vee \gamma(\log\lambda)\varepsilon$. At the same time, λ_ε belongs to a Kullback-Leibler "ball" around λ whose radius is $\frac{1}{\varepsilon}\|f_\lambda - f_*\|^2_{L_2(\Pi)} \vee \gamma^2(\log\lambda)\varepsilon$. Thus, the existence of an oracle λ that approximates the target function well (i.e., $\|f_\lambda - f_*\|_{L_2(\Pi)}$ is small) and that is "well aligned" with the dictionary (i.e., $\gamma(\log\lambda)$ is small) would imply that f_{λ_ε} is $L_2(\Pi)$-close to f_λ and, at the same time, λ_ε is close to λ in the Kullback-Leibler distance. It would also imply that the approximation error $\mathcal{E}(f_{\lambda_\varepsilon})$ is small and that the measures Λ_ε and Λ (with densities λ_ε and λ) have similar "concentration pattern" (as the last two inequalities of the theorem show).

Theorem 1 *There exists a constant $C > 0$ depending only on the loss such that, for all oracles $\lambda \in \mathbb{D}$,*

$$\|f_{\lambda_\varepsilon} - f_\lambda\|^2_{L_2(\Pi)} + \varepsilon K(\lambda_\varepsilon, \lambda) \le C\left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} \bigvee \gamma^2(\log\lambda)\varepsilon^2\right].$$

Moreover, the following bound on the excess risk of λ_ε holds

$$\mathcal{E}(f_{\lambda_\varepsilon}) \le \inf_{\lambda\in\mathbb{D}}\left[\mathcal{E}(f_\lambda) + C\sqrt{\mathcal{E}(f_\lambda)}\gamma(\log\lambda)\varepsilon + C\gamma^2(\log\lambda)\varepsilon^2\right]$$

and, for all $\mathcal{H}' \subset \mathcal{H}$,

$$\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}') \le 2\Lambda(\mathcal{H}\setminus\mathcal{H}') + \frac{C}{\varepsilon}\left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} \bigvee \gamma^2(\log\lambda)\varepsilon^2\right],$$

$$\Lambda(\mathcal{H}\setminus\mathcal{H}') \le 2\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}') + \frac{C}{\varepsilon}\left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} \bigvee \gamma^2(\log\lambda)\varepsilon^2\right].$$

In concrete examples below, the dictionary is of the form $\mathcal{H} = \{h_t : t \in I\}$, where $I \subset \mathbb{R}^d$ is a bounded domain in $\mathbb{R}^d$. In such cases, one can assume that μ is a measure on I and the mixing densities λ are functions on I. Often, it happens that $K^{-1/2}$ can be defined in terms of certain differential operators and the alignment coefficient $\gamma(w)$ is bounded by a Sobolev type norm of the function w : for some $A > 0$ and $\alpha > 0$,

$$\gamma(w) \le A\|w\|_{\mathbb{W}^{2,\alpha}(I)}. \tag{3.1}$$

We are interested in those oracles $\lambda \in \Lambda$ for which $\gamma(\log\lambda)$ is not too large and it is controlled by "smoothness" and "sparsity" of λ. Assume that μ is the Lebesgue measure on I and that condition (3.1) holds. Let

$\lambda := \sum_{j=1}^d \lambda_j + \delta$, where $\delta \in (0,1)$ and λ_j are nonnegative functions, $\lambda_j \in C^\infty(\mathbb{R}^d)$, $\mathrm{supp}(\lambda_j) \subset U_j$, $U_j \subset T$ being disjoint balls. Finally, we assume that $\sum_{j=1}^d \int_{\mathbb{R}^d} \lambda_j(t)dt = 1-\delta$. Then it is easy to see that

$$\log\lambda = \sum_{j=1}^d (\log(\lambda_j+\delta) - \log\delta) + \log\delta$$

and, for each $j = 1,\dots,d$, the function $w_j := \log(\lambda_j+\delta) - \log\delta \in C^\infty(\mathbb{R}^d)$ and it is supported in U_j. Therefore, since the functions w_j have disjoint supports,

$$\gamma(\log\lambda) \le A_1 \Big\|\sum_{j=1}^d w_j\Big\|_{\mathbb{W}^{2,\alpha}(I)} \le A\Big(\sum_{j=1}^d \|w_j\|^2_{\mathbb{W}^{2,\alpha}(I)}\Big)^{1/2}.$$

In this model, δ plays the role of a small "background density" (needed to make λ bounded away from 0) and densities $\lambda_j, j=1,\dots,d$ are "spikes". The resulting oracle density λ is "approximately" sparse in the sense that most of the mass is concentrated in a small part of the space (in the union of balls U_j). For smooth enough "spikes", $\gamma(\log\lambda)$ becomes of the order $\sqrt{d}$, so, it depends on the "sparsity" of the problem.

We now consider three more specific examples of the dictionaries.

Fourier dictionary. Suppose that $S := \mathbb{R}^d$ and let $\mathcal{H} = \{\cos\langle t,\cdot\rangle,\ t \in I\}$, where $I \subset \mathbb{R}^d$ is a bounded open set symmetric about the origin, i.e., $I = -I$. It can be assumed now that the measure μ and the densities λ are defined on the set I. Suppose that measures μ, Π are absolutely continuous with respect to the Lebesgue measure with densities m and p, respectively. It will be assumed that $m(t) = m(-t), t \in I$. We will also assume that for $\lambda \in \mathbb{D}$, $\lambda(t) = \lambda(-t), t \in I$. When it is needed, it will be assumed that functions λ, m are defined on the whole space $\mathbb{R}^d$ and are equal to 0 on $\mathbb{R}^d \setminus I$. Clearly, the function f_λ is then the Fourier transform of λm :

$$f_\lambda(\cdot) = \int_{\mathbb{R}^d} e^{i\langle t,\cdot\rangle}\lambda(t)m(t)dt := \widehat{\lambda m}(\cdot).$$

Therefore, assuming that the density p is positive, we get, for all $w \in C^\infty(\mathbb{R}^d),\ u \in \mathbb{K}$

$$\langle w,u\rangle_{L_2(\mu)} = \langle w, u\cdot m\rangle_{L_2(\mathbb{R}^d)} = \langle \widehat{w}, \widehat{um}\rangle_{L_2(\mathbb{R}^d)} = \langle \widehat{w}, f_u\rangle_{L_2(\mathbb{R}^d)} = \Big\langle \frac{\widehat{w}}{p^{1/2}}, f_u p^{1/2}\Big\rangle_{L_2(\mathbb{R}^d)},$$

which easily implies that $\gamma(w) \le \Big\|\frac{\widehat{w}}{\sqrt{p}}\Big\|_{L_2(\mathbb{R}^d)}$. Under an additional assumption that for some $L > 0, \alpha > 0$, $p(x) \ge L(1+|x|^2)^{-\alpha}, x \in \mathbb{R}^d$, we get the following bound: $\gamma(w) \le A_1\|(I+\Delta)^{\alpha/2}w\|_{L_2(\mathbb{R}^d)} \le A\|w\|_{\mathbb{W}^{2,\alpha}(\mathbb{R}^d)}$ (where Δ stands for the Laplace operator).

Location dictionary. Suppose now that $S := \mathbb{T}^d$ is the d-dimensional torus and let $\mathcal{H} = \{h(\cdot-\theta),\ \theta \in \mathbb{T}^d\}$ for some bounded function $h : \mathbb{T}^d \to \mathbb{R}$ and let μ be the Haar measure on $\mathbb{T}^d$. Assume that Π is a probability measure on $\mathbb{T}^d$ with density p (with respect to the Haar measure) that is bounded away from 0 by a constant $L > 0$. Then, a simple Fourier analysis argument shows that

$$\gamma(w) \le A\Big(\sum_{n\in\mathbb{Z}^d}\Big|\frac{\hat{w}_n}{\hat{h}_n}\Big|^2\Big)^{1/2},$$

where $\hat{w}_n, \hat{h}_n$ denote the Fourier coefficients of functions w, h. Under the assumption that $|\hat{h}_n| \ge L(1+|n|^2)^{-\alpha/2}$, it easily follows that

$$\gamma(w) \le A\|w\|_{\mathbb{W}^{2,\alpha}(\mathbb{T}^d)}.$$

Monotone functions dictionary. Assuming that $S = [0,1]$, let $\mathcal{H} := \{I_{[0,s]} : s \in [0,1]\}$ and let μ be the Lebesgue measure in $[0,1]$. The mixtures of functions from $\mathcal{H}$ are decreasing absolutely continuous functions $f : [0,1] \mapsto [0,1]$ such that $f(0) = 1$ and $f(1) = 0$. Suppose that Π is the Lebesgue measure in $[0,1]$. The Gram operator K is given by the kernel $K(s,t) = \langle I_{[0,s]}, I_{[0,t]}\rangle_{L_2(\Pi)} = \min(s,t)$. Clearly, K is a compact self-adjoint operator. It is well known that its eigenvalues are $\left(\frac{1}{\pi(k+1/2)}\right)^2$ and the corresponding eigenfunctions are $\phi_k(t) = \sqrt{2}\sin((k+1/2)\pi t), k = 0,1,2,\dots$. For a function $w \in \mathbb{W}^{2,1}[0,1],\ w(0) = 0$, $w = \sum\limits_{k=0}^{\infty} w_k\phi_k$, we have

$$\left(K^{-1/2}w\right)(t) = \sum_{k=0}^{\infty}\pi(k+1/2)w_k\phi_k(t) = w'(t).$$

Hence

$$\gamma(w) \le \|K^{-1/2}w\|_{L_2[0,1]} = \pi\left(\sum_{k=0}^{\infty}(k+1/2)^2 w_k^2\right)^{1/2} \le A\|w\|_{\mathbb{W}^{2,1}[0,1]}.$$

Assume again that $\mathcal{H}$ is an arbitrary dictionary.

Weakly correlated partitions. Let $\mathcal{H}_j, j = 1,\dots,N$ be a measurable partition of $\mathcal{H}$. As a concrete example of such a partition, one can consider the case when $S = [0,1]^N$ and, for each $j = 1,\dots,N$, $\mathcal{H}_j$ is a class of functions depending on the j-th variable. We are interested in the situation when the number N of function classes $\mathcal{H}_j$ is large and they are "weakly correlated“. This might be viewed as an extension to the case of infinite dictionaries of usual notions of "almost orthogonality“ (such as, for instance, restricted isometry property of Candes and Tao) frequently used in the literature on sparse recovery. It is also close to "sparse additive models“ and "sparse multiple kernel learning“ (see [KY08], [MvdGB09]). Suppose there exist oracles $\lambda \in \mathbb{D}$ such that f_λ provides a good approximation of the target f_* and, at the same time, λ is "sparse“ in the sense that it is concentrated mostly on a small number of sets $\mathcal{H}_j$. For each set $\mathcal{H}_j$, let $K_j : L_2(\mathcal{H}_j;\mu) \mapsto L_2(\mathcal{H}_j,\mu)$ be the integral operator (self-adjoint and nonnegatively definite) defined by

$$(K_j u)(h) := \int_{\mathcal{H}_j}\mathrm{cov}_\Pi(h,g)u(g)\mu(dg),\ h \in \mathcal{H}_j,$$

where $\mathrm{cov}_\Pi(h,g) := \Pi(hg) - \Pi(h)\Pi(g)$. We will also denote

$$\sigma_\Pi(g) := \sqrt{\mathrm{cov}_\Pi(g,g)} \text{ and } \rho_\Pi(h,g) := \frac{\mathrm{cov}_\Pi(h,g)}{\sigma_\Pi(h)\sigma_\Pi(g)}.$$

Let $\mathcal{L}_j$ be the subspace of $L_2(\Pi)$ spanned by $\mathcal{H}_j$ and, for $J \subset \{1,\dots,N\}$, let

$$\beta_2(J) := \inf\left\{\beta > 0 : \forall f_j \in \mathcal{L}_j, j = 1,\dots,N\ \sum_{j\in J}\sigma_\Pi^2(f_j) \le \beta^2\sigma_\Pi^2\left(\sum_{j=1}^{N} f_j\right)\right\}.$$

Note that if the spaces $\mathcal{L}_j, j = 1,\dots,N$ are uncorrelated, i.e., $\mathrm{cov}_\Pi(h,g) = 0, h \in \mathcal{L}_i, g \in \mathcal{L}_j, i \ne j$, then $\beta_2(J) = 1$. More generally, given $h_j \in \mathcal{L}_j,\ j = 1,\dots,N$, denote by $\kappa(\{h_j : j \in J\})$ the minimal eigenvalue of the covariance matrix $(\mathrm{cov}_\Pi(h_i,h_j))_{i,j\in J}$. Let

$$\kappa(J) := \inf\left\{\kappa(\{h_j : j \in J\}) : h_j \in \mathcal{L}_j, \sigma_\Pi(h_j) = 1\right\}.$$

Denote $\mathcal{L}_J = \mathrm{l.s.}\left(\bigcup_{j\in J}\mathcal{L}_j\right)$ (here l.s. means linear span) and let $\rho(J) := \sup\left\{\rho_\Pi(f,g) : f \in \mathcal{L}_J, g \in \mathcal{L}_{J^c}\right\}$. The quantity $\rho(J)$ should be compared with the notion of **canonical correlation** often used in the multivariate statistical analysis. It is easy to check (see [Kol08], proposition 7.1) that

$$\beta_2(J) \le \frac{1}{\sqrt{\kappa(J)(1-\rho^2(J))}}.$$

The next proposition easily follows from the definitions of $\gamma(w), \beta_2(J)$ and the operators K_j.

Proposition 4 *For all $J \subset \{1,\dots,N\}$ and all $w = \sum_{j\in J} w_j$ with $w_j \in \mathrm{Im}(K_j^{1/2})$,*

$$\gamma(w) \le \beta_2(J)\biggl(\sum_{j\in J} \|K_j^{-1/2} w_j\|^2_{L_2(\mathcal{H}_j,\mu)}\biggr)^{1/2}. \tag{3.2}$$

If now $\lambda := \sum_{j\in J} \lambda_j + \delta$, where $\delta \in (0,1)$, λ_j are nonnegative functions defined on $\mathcal{H}_j$ and

$$\sum_{j=1}^{d} \int_{\mathcal{H}_j} \lambda_j(h)dh = 1 - \delta,$$

then $\log\lambda = \sum_{j\in J} w_j I_{\mathcal{H}_j} + \log\delta$, where $w_j := \log(\lambda_j+\delta) - \log\delta$. Therefore, (3.2) implies

$$\gamma(\log\lambda) \le \beta_2(J)\biggl(\sum_{j\in J} \|K_j^{-1/2} w_j\|^2_{L_2(\mathcal{H}_j,\mu)}\biggr)^{1/2}.$$

4 Bounding Random Error

The purpose of this section is to develop exponential bounds on the random error $\Bigl|\mathcal{E}(f_{\hat\lambda_\varepsilon}) - \mathcal{E}(f_{\lambda_\varepsilon})\Bigr|$ that depend on the "approximate sparsity" of the true penalized solution λ_ε. Since we are dealing with a loss ℓ of quadratic type, bounding the random error is essentially equivalent to bounding the norm $\|f_{\hat\lambda_\varepsilon} - f_{\lambda_\varepsilon}\|_{L_2(\Pi)}$ (see Proposition 1). At the same time, we provide upper bounds on the symmetrized Kullback-Leibler distance between $\hat\lambda_\varepsilon$ and λ_ε and show that the "approximate sparsity" properties of these two functions are closely related.

Let $\mathcal{H}'$ be a measurable subset of $\mathcal{H}$. In the theorem below, it will be a subset of the dictionary $\mathcal{H}$ on which both $\hat\lambda_\varepsilon$ and λ_ε are approximately concentrated. Let L be a finite dimensional subspace of $L_2(\Pi)$ that will be used to approximate the functions from $\mathcal{H}'$. Let $d := \dim(L)$ and denote $U_L(x) := \sup_{h\in L, \|h\|_{L_2(\Pi)}\le 1} |h(x)|$. It is easy to check (using the Cauchy-Schwarz inequality) that $\|U_L\|_{L_2(\Pi)} = \sqrt{d}$. Denote $U(L) := \|U_L\|_{L_\infty} + 1$. Note that $U(L)$ is of the order $\sqrt{d}$ if there exists an orthonormal basis $\phi_1,\dots,\phi_d$ of L such that the functions ϕ_j are uniformly bounded by a constant. Finally, let $\rho(\mathcal{H}';L) := \sup_{h\in\mathcal{H}'} \|P_{L^\perp} h\|_{L_2(\Pi)}$, where $P_{L^\perp}$ stands for the orthogonal projection on $L^\perp$. We are interested in those subspaces L for which d and $U(L)$ are not very large and $\rho(\mathcal{H}';L)$ is small enough, i.e., the space L provides a reasonably good $L_2(\Pi)$-approximation of the functions from $\mathcal{H}'$. A natural choice of L might be a subspace spanned on the centers of the $L_2(\Pi)$-balls of small enough radius δ covering $\mathcal{H}'$; in this case $\rho(\mathcal{H}';L) \le \delta$ and d is equal to the cardinality of such a δ-covering.

For a function class $\mathcal{G}$ and a probability measure Q on S, let $N(\mathcal{G};L_2(Q);\varepsilon)$ denote the minimal number of $L_2(Q)$-balls of radius ε covering $\mathcal{G}$. We will need the following **complexity assumption** on the base class $\mathcal{H}$: there exists a nonnegative nonincreasing function Ω such that $\Omega(u) \to \infty$ as $u \to 0$, Ω is a regularly varying function of exponent $\alpha \in [0,2)$ and, with probability 1,

$$\log N(\mathcal{H};L_2(\Pi_n);u/2) \le \Omega(u),\ u > 0,\ n \in \mathbb{N}. \tag{4.1}$$

In particular, for VC-type classes of VC-dimension V such a bound holds with $\Omega(u)$ of the order $V\log(1/u)$.

Theorem 2 *Suppose that the complexity assumption (4.1) holds. There exist constants $C, D > 0$ depending only on ℓ such that for all measurable subsets $\mathcal{H}' \subset \mathcal{H}$, for all finite dimensional subspaces $L \subset L_2(\Pi)$ with $d := \dim(L)$ and $\rho := \rho(\mathcal{H}';L)$, for all*

$$\varepsilon \ge D\sqrt{\frac{\Omega(1/\sqrt{d})}{n}}$$

and for all $t > 0$, the following bounds hold with probability at least $1 - e^{-t}$:

$$\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \le C \left[\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \bigvee \frac{d + t_n}{n\varepsilon} \bigvee \frac{\rho}{\varepsilon} \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d})}{n\varepsilon} \right], \tag{4.2}$$

$$\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \le C \left[\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \bigvee \frac{d + t_n}{n\varepsilon} \bigvee \frac{\rho}{\varepsilon} \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d})}{n\varepsilon} \right] \tag{4.3}$$

and

$$\|f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon K(\hat{\lambda}_\varepsilon, \lambda_\varepsilon) \le C \biggl[\frac{d + t_n}{n} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee$$
$$\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d})}{n} \biggr], \tag{4.4}$$

where $t_n := t + 4 \log \log_2 n + 2 \log 2$.

Theorems 1, 2 and Proposition 1 yield the following **sparsity oracle inequality** for the excess risk of $f_{\hat{\lambda}_\varepsilon}$: for all oracles $\lambda \in \mathbb{D}$, with probability at least $1 - e^{-t}$,

$$\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) \le 2\mathcal{E}(f_\lambda) + C \Biggl[\frac{d + t_n}{n} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \Lambda(\mathcal{H} \setminus \mathcal{H}') \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee$$
$$\bigvee \frac{U(L)\Omega(\rho/\sqrt{d})}{n} \bigvee \varepsilon^2 \gamma^2 (\log \lambda) \Biggr].$$

Proof: Let λ_ε be the solution of (1.1) and $\hat{\lambda}_\varepsilon$ be the solution of (1.2). Denote

$$\Lambda_\varepsilon(A) := \int_A \lambda_\varepsilon(h)\mu(dh), \ \ \hat{\Lambda}_\varepsilon(A) := \int_A \hat{\lambda}_\varepsilon(h)\mu(dh).$$

Using (2.1) and (2.2), for all $\tau \in (0, 1)$, the directional derivative of F exists at the point $\lambda_\varepsilon + \tau\hat{\lambda}_\varepsilon$ in the direction $\hat{\lambda}_\varepsilon - \lambda_\varepsilon$ and

$$DF(\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon); \hat{\lambda}_\varepsilon - \lambda_\varepsilon) = \tag{4.5}$$
$$P(\ell' \bullet f_{\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon)})(f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}) + \varepsilon \int_{\mathcal{H}} (\hat{\lambda}_\varepsilon - \lambda_\varepsilon) \log(\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon)) d\mu \ge 0.$$

[Note that the directional derivative of entropy H in the direction of $\hat{\lambda}_\varepsilon - \lambda_\varepsilon$ does not necessarily exist at the point λ_ε itself which explains the need in a somewhat more complicated argument given here]. Moreover, since the function $[0, 1] \ni \tau \mapsto F(\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon))$ is convex, its right derivative, which coincides with $DF(\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon); \hat{\lambda}_\varepsilon - \lambda_\varepsilon)$, is nondecreasing in $\tau \in [0, 1]$. Since λ_ε is the minimal point of F, this implies that, for $\tau \in (0, 1)$,

$$DF(\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon); \hat{\lambda}_\varepsilon - \lambda_\varepsilon) =$$
$$= P(\ell' \bullet f_{\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon)})(f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}) + \varepsilon \int_{\mathcal{H}} (\hat{\lambda}_\varepsilon - \lambda_\varepsilon) \log(\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon)) d\mu \ge 0. \tag{4.6}$$

A similar argument shows that for all $\tau \in (0, 1)$

$$DF_n(\hat{\lambda}_\varepsilon + \tau(\lambda_\varepsilon - \hat{\lambda}_\varepsilon); \hat{\lambda}_\varepsilon - \lambda_\varepsilon) = \tag{4.7}$$
$$P_n(\ell' \bullet f_{\hat{\lambda}_\varepsilon + \tau(\lambda_\varepsilon - \hat{\lambda}_\varepsilon)})(f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}) + \varepsilon \int_{\mathcal{H}} (\hat{\lambda}_\varepsilon - \lambda_\varepsilon) \log(\hat{\lambda}_\varepsilon + \tau(\lambda_\varepsilon - \hat{\lambda}_\varepsilon)) d\mu \le 0.$$

Subtracting (4.6) from (4.7) and rearranging the terms, we get

$$P\left(\ell'\bullet f_{\hat{\lambda}_\varepsilon+\tau(\lambda_\varepsilon-\hat{\lambda}_\varepsilon)} - \ell'\bullet f_{\lambda_\varepsilon+\tau(\hat{\lambda}_\varepsilon-\lambda_\varepsilon)}\right)(f_{\hat{\lambda}_\varepsilon}-f_{\lambda_\varepsilon})+ \tag{4.8}$$
$$+\,\varepsilon\int\left(\hat{\lambda}_\varepsilon-\lambda_\varepsilon\right)\log\frac{(1-\tau)\hat{\lambda}_\varepsilon+\tau\lambda_\varepsilon}{(1-\tau)\lambda_\varepsilon+\tau\hat{\lambda}_\varepsilon}d\mu \le \left|(P-P_n)(\ell'\bullet f_{\hat{\lambda}_\varepsilon+\tau(\lambda_\varepsilon-\hat{\lambda}_\varepsilon)})(f_{\hat{\lambda}_\varepsilon}-f_{\lambda_\varepsilon})\right|.$$

Under the assumptions on the loss (in particular, continuity of ℓ'), passing to the limit as $\tau\to 0$, using the dominated convergence, equation (2.3) of Proposition 3 and the bound

$$P\left(\ell'\bullet f_{\hat{\lambda}_\varepsilon}-\ell'\bullet f_{\lambda_\varepsilon}\right)(f_{\hat{\lambda}_\varepsilon}-f_{\lambda_\varepsilon}) \ge c\|f_{\hat{\lambda}_\varepsilon}-f_{\lambda_\varepsilon}\|^2_{L_2(\Pi)}$$

that holds for losses of quadratic type, we get

$$c\|f_{\hat{\lambda}_\varepsilon}-f_{\lambda_\varepsilon}\|^2_{L_2(\Pi)}+\varepsilon K(\hat{\lambda}_\varepsilon,\lambda_\varepsilon) \le \left|(P-P_n)(\ell'\bullet f_{\hat{\lambda}_\varepsilon})(f_{\hat{\lambda}_\varepsilon}-f_{\lambda_\varepsilon})\right|. \tag{4.9}$$

To complete the proof of the theorem, it remains to bound $\left|(P-P_n)(\ell'\bullet f_{\hat{\lambda}_\varepsilon})(f_{\hat{\lambda}_\varepsilon}-f_{\lambda_\varepsilon})\right|$. Let

$$\Lambda(\delta,\Delta):=\left\{\lambda\in\mathbb{D}:\ \|f_\lambda-f_{\lambda_\varepsilon}\|_{L_2(\Pi)}\le\delta,\ \int\limits_{\mathcal{H}\setminus\mathcal{H}'}\lambda(h)\mu(dh)\le\Delta\right\}$$

and

$$\alpha_n(\delta,\Delta):=\sup\left\{|(P_n-P)(\ell'\bullet f_\lambda)(f_\lambda-f_{\lambda_\varepsilon})|,\ \lambda\in\Lambda(\delta,\Delta)\right\}.$$

In what follows we will use Rademacher processes

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

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

Lemma 3 *Let $\mathcal{H}$ be a class of functions on S uniformly bounded by 1 and let $L\subset L_2(\Pi)$ be a finite dimensional subspace with $d:=\dim(L)$ and $\rho:=\rho(\mathcal{H};L)$. Suppose that assumption (4.1) holds for some function Ω. Then with some constant $C>0$*

$$\mathbb{E}\sup_{h\in\mathcal{H}}|R_n(P_{L^\perp}h)|\le C\left[\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}}\bigvee\frac{U(L)\Omega(\rho/\sqrt{d})}{n}\right].$$

Proof: The following is true for all $h_1,h_2\in\mathcal{H}$

$$|P_L(h_1)(x)-P_L(h_2)(x)|\le U_L(x)\|P_L(h_1)-P_L(h_2)\|_{L_2(\Pi)}\le U_L(x)\|h_1-h_2\|_{L_2(\Pi)}$$

and it implies that

$$\|P_L(h_1)-P_L(h_2)\|_{L_2(\Pi_n)}\le\|U_L\|_{L_2(\Pi_n)}\|h_1-h_2\|_{L_2(\Pi)}.$$

Therefore,

$$\log N(P_L(\mathcal{H});L_2(\Pi_n);u)\le\log N\left(\mathcal{H};L_2(\Pi);\frac{u}{\|U_L\|_{L_2(\Pi_n)}}\right). \tag{4.10}$$

Complexity assumption (4.1), together with the law of large numbers, gives the bound for covering numbers with respect to $L_2(\Pi)$(see the proof of Theorem 3.4 in [GK06]):

$$\log N\left(\mathcal{H};L_2(\Pi),u\right)\le\Omega(u). \tag{4.11}$$

Since $P_{L^\perp}h = h - P_L h$, (4.10) implies

$$N(P_{L^\perp}(\mathcal{H}); L_2(\Pi_n); u) \le N\left(\mathcal{H}; L_2(\Pi_n), u/2\right) N\left(\mathcal{H}; L_2(\Pi); \frac{u}{2\|U_L\|_{L_2(\Pi_n)}}\right).$$

Recalling the complexity conditions (4.1) and (4.11), we easily get

$$\log N(P_{L^\perp}(\mathcal{H}); L_2(\Pi_n); u) \le \Omega(u) + \Omega\left(\frac{u}{2\|U_L\|_{L_2(\Pi_n)}}\right).$$

It remains to use Theorem 3.1 from [GK06] to complete the proof. □

Lemma 4 *Under the assumptions of Theorem 2, there exists a constant $C > 0$ depending only on the loss such that with probability at least $1 - e^{-t}$ for all $\frac{1}{\sqrt{n}} \le \delta \le 1$, $\frac{1}{\sqrt{n}} \le \Delta \le 1$*

$$\begin{aligned}\alpha_n(\delta,\Delta) \le & C\left[\delta\sqrt{\frac{d+t_n}{n}} \bigvee \rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \right.\\ & \left.\Delta\sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d})}{n} \bigvee \frac{t_n}{n}\right] =: \hat{\beta}_n(\delta,\Delta).\end{aligned}$$

where $t_n := t + 4\log\log_2 n + 2\log 2$.

Proof: Recall that $\alpha_n(\delta,\Delta) := \sup\left\{|(P-P_n)(\ell'\bullet f_\lambda)(f_\lambda - f_{\lambda_\varepsilon})|,\ \lambda\in\Lambda(\delta,\Delta)\right\}$. The function $u \mapsto \ell'(y, f_{\lambda_\varepsilon}+u)u,\ |u|\le 2$ is Lipschitz with Lipschitz constant depending only on ℓ. Note that $\ell'(y, f_\lambda(\cdot))(f_\lambda(\cdot) - f_{\lambda_\varepsilon}(\cdot)) = \ell'(y, f_{\lambda_\varepsilon}+u)u|_{u=f_\lambda(\cdot)-f_{\lambda_\varepsilon}(\cdot)}$ This allows us to apply the symmetrization and contraction inequalities (see [vdVW96], Lemma 2.3.6 and Proposition A.3.2) which results in the following bound:

$$\mathbb{E}\alpha_n(\delta,\Delta) \le C\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}|R_n(f_\lambda - f_{\lambda_\varepsilon})|,$$

where $C > 0$ is a constant depending only on ℓ. Let P_L denote the orthogonal projection on a d-dimensional subspace L. The following representation is straightforward:

$$f_\lambda - f_{\lambda_\varepsilon} = P_L\left(f_\lambda - f_{\lambda_\varepsilon}\right) + \int\limits_{\mathcal{H}'} P_{L^\perp}(h)\left(\lambda(h) - \lambda_\varepsilon(h)\right)\mu(dh) + \int\limits_{\mathcal{H}\setminus\mathcal{H}'} P_{L^\perp}(h)\left(\lambda(h)-\lambda_\varepsilon(h)\right)\mu(dh). \tag{4.12}$$

Hence, it is enough to bound separately the expected supremum of the Rademacher process R_n for each term in the sum. For the first term, the standard bound on Rademacher processes indexed by a finite dimensional subspace (see, e.g., [Kol08], proposition 3.2) yields

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}\left|R_n\left(P_L(f_\lambda - f_{\lambda_\varepsilon})\right)\right| \le \delta\sqrt{\frac{d}{n}}. \tag{4.13}$$

To bound the remaining terms, we will use Lemma 3. First, due to linearity of the Rademacher process,

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}\left|R_n\left(\int\limits_{\mathcal{H}\setminus\mathcal{H}'}(\lambda-\lambda_\varepsilon)(h)P_{L^\perp}h\,\mu(dh)\right)\right| \le \Big(\Delta + \Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\Big)\mathbb{E}\sup_{h\in\mathcal{H}\setminus\mathcal{H}'}|R_n(P_{L^\perp}h)|. \tag{4.14}$$

We now use the bound of Lemma 3 with $\mathcal{H}\setminus\mathcal{H}'$ instead of $\mathcal{H}$ and with $\rho = 1$ to get

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}\left|R_n\left(\int\limits_{\mathcal{H}\setminus\mathcal{H}'}(\lambda-\lambda_\varepsilon)(h)P_{L^\perp}h\,\mu(dh)\right)\right| \le \tag{4.15}$$

$$C\Big(\Delta + \Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\Big)\left[\sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(1/\sqrt{d})}{n}\right].$$

Similarly,

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}\left|R_n\left(\int\limits_{\mathcal{H}'}(\lambda-\lambda_\varepsilon)P_{L^\perp}hd\mu(h)\right)\right|\le 2\mathbb{E}\sup_{h\in\mathcal{H}'}|R_n(P_{L^\perp}h)|$$

and using the bound of Lemma 3 with $\mathcal{H}'$ instead of $\mathcal{H}$ and with $\rho:=\rho(\mathcal{H}',L)$, we get

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}\left|R_n\left(\int\limits_{\mathcal{H}'}(\lambda-\lambda_\varepsilon)(h)P_{L^\perp}hd\mu(h)\right)\right|\le C\left[\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}}\bigvee\frac{U(L)\Omega(\rho/\sqrt{d})}{n}\right]. \tag{4.16}$$

Combining (4.13)–(4.16) results in the following bound:

$$\mathbb{E}\alpha_n(\delta,\Delta)\le C\left[\delta\sqrt{\frac{d}{n}}\bigvee\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}}\bigvee\right. \tag{4.17}$$

$$\left.\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\sqrt{\frac{\Omega(1/\sqrt{d})}{n}}\bigvee\Delta\sqrt{\frac{\Omega(1/\sqrt{d})}{n}}\bigvee\frac{U(L)\Omega(\rho/\sqrt{d})}{n}\right].$$

Talagrand's concentration inequality (see, e.g., [Bou02]) implies that with probability at least $1-e^{-s}$ and with a proper choice of constant $C>0$

$$\alpha_n(\delta,\Delta)\le\beta_n(\delta,\Delta,s):=2\left(\mathbb{E}\alpha_n(\delta,\Delta)+C\delta\sqrt{\frac{s}{n}}+C\frac{s}{n}\right). \tag{4.18}$$

We have to make the bound uniform with respect to $\frac{1}{\sqrt{n}}\le\delta\le 1,\ \frac{1}{\sqrt{n}}\le\Delta\le 1$. To this end, let

$$\delta_j=\Delta_j=\frac{1}{2^j},\quad t_{i,j}=t+2\log(i+1)+2\log(j+1)+2\log 2,i,j\ge 0. \tag{4.19}$$

Then, with probability at least

$$1-\sum_{i,j:\delta_i,\Delta_j\ge n^{-1/2}}\exp\{-t_{i,j}\}\ge 1-e^{-t-\log 4}(\sum_{j\ge 0}(j+1)^{-2})^2\ge 1-e^{-t},$$

for all i,j such that $\delta_i,\Delta_j\ge n^{-1/2}$ and all δ,Δ such that $\delta\in(\delta_{i+1},\delta_i],\ \Delta\in(\Delta_{j+1},\Delta_j]$, the following bounds hold: $\alpha(\delta,\Delta)\le\beta(\delta_i,\Delta_j,t_{i,j})$. Note that

$$t_{i,j}\le t+2\log 2+2\log\log_2\left(\frac{1}{\delta}\right)+2\log\log_2\left(\frac{1}{\Delta}\right),$$

$$\frac{2\log\log_2\left(\frac{1}{\Delta}\right)}{n}\le 2\frac{\log\log_2(n)}{n},\quad\frac{2\log\log_2\left(\frac{1}{\delta}\right)}{n}\le 2\frac{\log\log_2(n)}{n},$$

implying that $t_{i,j}\le t_n$. Thus, with probability at least $1-e^{-t}$, for all $\delta,\Delta\in[n^{-1/2},1]$

$$\alpha_n(\delta,\Delta)\le\hat\beta_n(\delta,\Delta):=C\left[\delta\sqrt{\frac{d+t_n}{n}}\bigvee\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}}\bigvee\right.$$

$$\left.\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\sqrt{\frac{\Omega(1/\sqrt{d})}{n}}\bigvee\Delta\sqrt{\frac{\Omega(1/\sqrt{d})}{n}}\bigvee\frac{U(L)\Omega(\rho/\sqrt{d})}{n}\bigvee\frac{t_n}{n}\right].$$

□

To complete the proof of the theorem, denote $\hat\delta:=\|f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon}\|_{L_2(\Pi)}$, and $\hat\Delta:=\hat\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')$. By Lemma 4, (4.9) and (2.4) of Proposition 3, the following inequalities hold with probability at least $1-e^{-t}$

$$c\hat\delta^2\le\hat\beta_n(\hat\delta,\hat\Delta), \tag{4.20}$$

$$\varepsilon\hat\Delta\le 2\varepsilon\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')+\hat\beta_n(\hat\delta,\hat\Delta), \tag{4.21}$$

provided that $\hat\delta\ge n^{-1/2},\hat\Delta\ge n^{-1/2}$. It remains to solve (4.20), (4.21) for $\hat\delta,\hat\Delta$ (using the assumption on ε) to get the desired bounds (in the cases when $\hat\delta<n^{-1/2}$ and/or $\hat\Delta<n^{-1/2}$ the derivation becomes even simpler).

□

References

[Bou02] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[GK06] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216, 2006.

[Kol08] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. *Lecture Notes for Ecole d'Eté de Probabilités de Saint-Flour*, 2008.

[Kol09a] V. Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359, 2009.

[Kol09b] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales Inst. H. Poincare, Probabilites et Statistique*, 45(1):7–57, 2009.

[KY08] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of 19th Annual Conference on Learning Theory(COLT 2008)*, pages 229–238, 2008.

[MPTJ07] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[MvdGB09] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 2009.

[RSSZ07] S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu. ℓ_1 regularization in infinite dimensional feature spaces. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 544–558. Springer, Berlin, 2007.

[vdVW96] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[Zha01] T. Zhang. Regularized winnow methods. In *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, pages 703–709. MIT Press, 2001.