2008 Saint Flour Lectures
Oracle Inequalities in Empirical Risk Minimization
and Sparse Recovery Problems

Vladimir Koltchinskii
School of Mathematics
Georgia Institute of Technology
Atlanta GA 30332-0160 USA
vlad@math.gatech.edu

September 15, 2009

Contents

Abstract

A number of problems in nonparametric statistics and learning theory can be formulated as penalized empirical risk minimization over large function classes with penalties depending on the complexity of the functions (decision rules) involved in the problem. The goal of mathematical analysis of such procedures is to prove "oracle inequalities" describing optimality properties of penalized empirical risk minimization with properly designed penalties as well as its adaptivity to unknown complexity of the problem. This requires a careful study of local properties of empirical, Rademacher and other stochastic processes indexed by function classes using the methods of high dimensional probability and asymptotic geometric analysis. Recently, this approach has proved to be especially useful in understanding of problems of recovery of a target function that has a sparse representation in a given large dictionary based on noisy measurements of this function at random locations.

Preface

The purpose of these lecture notes is to provide an introduction to the general theory of empirical risk minimization with an emphasis on excess risk bounds and oracle inequalities in penalized problems. In the recent years, there have been new developments in this area motivated by the study of new classes of methods in Machine Learning such as large margin classification methods (boosting, kernel machines). The main probabilistic tools involved in the analysis of these problems are concentration and deviation inequalities by Talagrand along with other methods of empirical processes theory (symmetrization inequalities, contraction inequality for Rademacher sums, entropy and generic chaining bounds). Sparse recovery based on ℓ_1-type penalization is another active area of research where the main problems can be stated in the framework of penalized empirical risk minimization and concentration inequalities and empirical processes tools proved to be very useful.

My interest in empirical processes started in the late 70s and early 80s. It was largely influenced by the work of Vapnik and Chervonenkis on Glivenko-Cantelli problem and on empirical risk minimization in pattern recognition, and, especially, by the results of Dudley on uniform central limit theorems. Talagrand's concentration inequality proved in the 90s was a major result with deep consequences in the theory of empirical processes and related areas of statistics, and it inspired many new approaches in analysis of empirical risk minimization problems.

Over the last years, the work of many people have had a profound impact on my own research and on my view of the subject of these notes. I was lucky to work together with several of them and to have numerous conversations and email exchanges with many others. I am especially thankful to Peter Bartlett, Lucien Birgé, Gilles Blanchard, Stephane Boucheron, Olivier Bousquet, Richard Dudley, Sara van de Geer, Evarist Giné, Gabor Lugosi, Pascal Massart, David Mason, Shahar Mendelson, Dmitry Panchenko, Alexandre Tsybakov, Aad van der Vaart, Jon Wellner and Joel Zinn.

I am thankful to the School of Mathematics, Georgia Institute of Technology and to the Department of Mathematics and Statistics, University of New Mexico where most of my work for the past several years have taken place.

The research described in these notes has been supported in part by NSF grants MSPA-MPS-0624841, DMS-0304861 and CCF-0808863.

I was working on these notes while visiting the Isaac Newton Institute for Mathematical Sciences in Cambridge in 2008. I am thankful to the Institute for its hospitality.

1 Introduction

1.1 Abstract Empirical Risk Minimization

Let $X, X_1, \dots, X_n, \dots$ be i.i.d. random variables defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ and taking values in a measurable space $(S, \mathcal{A})$ with common distribution P. Let P_n denote the empirical measure based on the sample $(X_1, \dots, X_n)$ of the first n observations:

$$P_n := n^{-1} \sum_{j=1}^{n} \delta_{X_j},$$

where $\delta_x,\ x \in S$ is the Diracs's measure. Let $\mathcal{F}$ be a class of measurable functions $f : S \mapsto \mathbb{R}$. In what follows, the values of a function $f \in \mathcal{F}$ will be interpreted as a "loss" associated with a certain "action" and the expectation of $f(X)$

$$\mathbb{E}f(X) = \int_S f dP = Pf$$

will be viewed as the risk of a certain "decision rule". We will be interested in the problem of risk minimization

$$Pf \longrightarrow \min,\ f \in \mathcal{F} \tag{1.1}$$

in the cases when the distribution P is unknown and has to be estimated based on the data $(X_1, \dots, X_n)$. Since the empirical measure P_n is a natural estimator of P, the true risk can be estimated by the corresponding empirical risk

$$n^{-1} \sum_{j=1}^{n} f(X_j) = \int_S f dP_n = P_n f$$

and the risk minimization problem has to be replaced by the empirical risk minimization:

$$P_n f \longrightarrow \min,\ f \in \mathcal{F}. \tag{1.2}$$

Many important methods of statistical estimation such as maximum likelihood and more general M-estimation are versions of empirical risk minimization. The general theory of empirical risk minimization has started with seminal paper of Vapnik and Chervonenkis [94] (see Vapnik [93] for more references) although some important ideas go back to much earlier work on asymptotic theory of M-estimation. Vapnik and Chervonenkis were motivated by applications of empirical risk minimization in pattern recognition and learning theory that required the development of the theory in a much more general framework than what was common in statistical literature. Their key idea was to relate the quality of the solution of empirical risk minimization problem to the accuracy of approximation of the true distribution P by the empirical distribution P_n uniformly over function classes representing losses of decision rules. Because of this, they have studied general Glivenko-Cantelli problems about convergence of $\|P_n - P\|_{\mathcal{F}}$ to 0, where

$$\|Y\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |Y(f)|$$

for $Y : \mathcal{F} \mapsto \mathbb{R}$. Vapnik and Chervonenkis introduced a number of important characteristics of complexity of function classes, such as VC-dimensions and random entropies, that control the accuracy of empirical approximation. These results along with the development of classical limit theorems in Banach spaces in the 60s and 70s led to the general theory of empirical processes that started with the pathbreaking paper by Dudley [41] on central limit theorems for empirical measures (see Dudley [42], Pollard [81], van der Vaart and Wellner [95]).

In the 90s, Talagrand studied isoperimetric inequalities for product measures and, in particular, he proved a striking uniform version of Bernstein inequality describing concentration of $\|P_n - P\|_{\mathcal{F}}$ around its expectation (see Talagrand [86, 87]). This was a real breakthrough in the theory of empirical processes and empirical risk minimization. At about the same time a concept of *oracle inequalities* has been developed in nonparametric statistics (see, e.g., Johnstone [52]). In modern statistics, it is common to deal with a multitude of possible models that describe the same data (for instance, a family of models for unknown regression functions of varying complexity). An oracle inequality is a bound on the risk of a statistical estimator that shows that the performance of the estimator is almost (often, up to numerical constants) as good as it would be if the statistician had an access to an oracle that knows what the best model for the target function is. It happened that concentration inequalities provide rather natural probabilistic tools needed to develop oracle inequalities in a number of statistical problems. In particular, Birgé and Massart [15], Barron, Birgé and Massart [5], and, more recently, Massart [73, 74] suggested a general approach to model selection in a variety of statistical problems such as density estimation, regression and classification that is based on penalized empirical risk minimization. They used Talagrand's concentration and deviation inequalities in a systematic way to establish a number of oracle inequalities showing some form of optimality of penalized empirical risk minimization as a model selection tool.

In the recent years, new important classes of algorithms in machine learning have been introduced that are based on empirical risk minimization. In particular, large margin classification algorithms, such as boosting and support vector machines (SVM), can be viewed as empirical risk minimization over infinite dimensional functional spaces with special convex loss functions. In an attempt to understand the nature of these classification methods and to explain their superb generalization performance, there has been another round of work on the abstract theory of empirical risk minimization. One of the main ideas was to use sup-norms or localized sup-norms of the Rademacher processes indexed by function classes to develop a general approach to measuring the complexities of these classes (see Koltchinskii [58], Bartlett, Boucheron and Lugosi [8], Koltchinskii and Panchenko [60], Bousquet, Koltchinskii and Panchenko [23], Bartlett, Bousquet and Mendelson [7], Lugosi and Wegkamp [70], Bartlett and Mendelson [9]). This resulted in rather flexible definitions of distribution dependent and data dependent complexities in an abstract framework as well as more specialized complexities reflecting relevant parameters of specific learning machines. Moreover, such complexities have been used as natural penalties in model selection methods. This approach provided a general explanation of fast convergence rates in classification and other learning problems, the phenomenon discovered and studied by several authors, in particular, by Mammen and Tsybakov [72] and in an influential paper by Tsybakov [91].

1.2 Excess Risk: Distribution Dependent Bounds

Definition 1.1 *Let*

$$\mathcal{E}(f) := \mathcal{E}_P(f) := \mathcal{E}_P(\mathcal{F}; f) := Pf - \inf_{g\in\mathcal{F}} Pg.$$

This quantity will be called the excess risk of $f \in \mathcal{F}$.

Let

$$\hat{f} = \hat{f}_n \in \mathrm{Argmin}_{f\in\mathcal{F}} P_n f$$

be a solution of the empirical risk minimization problem (1.2). The function $\hat{f}_n$ is used as an approximation of the solution of the true risk minimization problem (1.1) and its excess risk $\mathcal{E}_P(\hat{f}_n)$ is a natural measure of accuracy of this approximation.

It is of interest to find tight upper bounds on the excess risk of $\hat{f}_n$ that hold with a high probability. Such bounds usually depend on certain ”geometric” properties of the function class $\mathcal{F}$ and on various measures of its ”complexity” that determine the accuracy of approximation of the true risk Pf by the empirical risk $P_n f$ in a neighborhood of a proper size of the minimal set of the true risk.

In fact, it is rather easy to describe a general approach to derivation of such bounds in an abstract framework of empirical risk minimization discussed in these notes. This approach does give a correct answer in many specific examples. To be precise, define the δ-*minimal set* of the risk as

$$\mathcal{F}(\delta) := \mathcal{F}_P(\delta) := \{f : \mathcal{E}_P(f) \le \delta\}.$$

Suppose, for simplicity, that the infimum of the risk Pf is attained at $\bar{f} \in \mathcal{F}$ (the argument can be easily modified if the infimum is not attained in the class). Denote $\hat{\delta} := \mathcal{E}_P(\hat{f})$. Then $\hat{f}, \bar{f} \in \mathcal{F}(\hat{\delta})$ and $P_n\hat{f} \le P_n\bar{f}$. Therefore,

$$\hat{\delta} = \mathcal{E}_P(\hat{f}) = P(\hat{f} - \bar{f}) \le P_n(\hat{f} - \bar{f}) + (P - P_n)(\hat{f} - \bar{f}),$$

which implies

$$\hat{\delta} \le \sup_{f,g \in \mathcal{F}(\hat{\delta})} |(P_n - P)(f - g)|.$$

Imagine there exists a nonrandom upper bound

$$U_n(\delta) \ge \sup_{f,g \in \mathcal{F}(\delta)} |(P_n - P)(f - g)| \tag{1.3}$$

that holds uniformly in δ with a high probability. Then, with the same probability, the excess risk $\mathcal{E}_P(\hat{f})$ will be bounded by the largest solution of the inequality $\delta \le U_n(\delta)$. There are many different ways to construct upper bounds on the sup-norms of empirical processes. A very general approach is based on Talagrand's concentration inequalities. Assume for simplicity that functions in the class $\mathcal{F}$ take their values in the interval $[0, 1]$. Based on the $L_2(P)$-diameter $D_P(\mathcal{F}; \delta)$ of the δ-minimal set $\mathcal{F}(\delta)$ and the function

$$\phi_n(\mathcal{F}; \delta) := \mathbb{E} \sup_{f,g \in \mathcal{F}(\delta)} |(P_n - P)(f - g)|,$$

define

$$\bar{U}_n(\delta; t) := K\biggl(\phi_n(\mathcal{F}; \delta) + D(\mathcal{F}; \delta)\sqrt{\frac{t}{n}} + \frac{t}{n}\biggr).$$

Talagrand's concentration inequality then implies that with some numerical constant $K > 0$, for all $t > 0$,

$$\mathbb{P}\Bigl\{\sup_{f,g \in \mathcal{F}(\delta)} |(P_n - P)(f - g)| \ge \bar{U}_n(\delta; t)\Bigr\} \le e^{-t}.$$

This observation provides an easy way to construct a function $U_n(\delta)$ such that (1.3) holds with a high probability uniformly in δ (first, by defining such a function at a discrete set of the values of δ and then extending it to all the values by monotonicity). By solving the inequality $\delta \leq U_n(\delta)$, one can construct a bound $\bar{\delta}_n(\mathcal{F})$ such that the probability $\mathbb{P}\{\mathcal{E}_P(\hat{f}_n) \geq \bar{\delta}_n(\mathcal{F})\}$ is small. Thus, constructing an upper bound on the excess risk essentially reduces to solving a fixed point equation of the type $\delta = U_n(\delta)$. Such a fixed point method has been studied, for instance, in Massart [73], Koltchinskii and Panchenko [60], Bartlett, Bousquet and Mendelson [7], Koltchinskii [59] (and in several other papers of these authors).

In the case of P-Donsker classes $\mathcal{F}$,

$$\phi_n(\mathcal{F};\delta) \leq \mathbb{E}\|P_n - P\|_{\mathcal{F}} = O(n^{-1/2}),$$

which implies that

$$\bar{\delta}_n(\mathcal{F}) = O(n^{-1/2}).$$

Moreover, if the diameter $D(\mathcal{F};\delta)$ of the δ-minimal set tends to 0 as $\delta \to 0$ (which is typically the case if the risk minimization problem (1.1) has a unique solution), then, by asymptotic equicontinuity, we have

$$\lim_{\delta\to 0}\limsup_{n\to\infty} n^{1/2}\phi_n(\mathcal{F};\delta) = 0,$$

which allows to conclude that

$$\bar{\delta}_n(\mathcal{F}) = o(n^{-1/2}).$$

It happens that the bound $\bar{\delta}_n(\mathcal{F})$ is of asymptotically correct order as $n \to \infty$ in many specific examples of risk minimization problem in statistics and learning theory.

The bounds of this type are *distribution dependent* (i.e., they depend on the unknown distribution P).

1.3 Rademacher Processes and Data Dependent Bounds on Excess Risk

The next challenge is to construct *data dependent* upper confidence bounds on the excess risk $\mathcal{E}_P(\hat{f})$ of empirical risk minimizers that depend only on the sample $(X_1, \ldots, X_n)$, but do not depend explicitly on the unknown distribution P. Such bounds can be used in model selection procedures. Their construction usually requires the development of certain statistical estimates of the quantities involved in the definition of the distribution dependent bound $\bar{\delta}_n(\mathcal{F})$ based on the sample $(X_1, \ldots, X_n)$. Namely, we have to estimate the expectation of the local sup-norm of the empirical process $\phi_n(\mathcal{F};\delta)$ and the diameter of the δ-minimal set.

A natural way to estimate the empirical process is to replace it by the Rademacher process

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

where $\{\varepsilon_j\}$ are i.i.d. Rademacher random variables (i.e., they are symmetric Bernoulli random variables taking values $+1$ and -1 with probability $1/2$ each) that are also independent of the data $(X_1, \dots, X_n)$. The process $R_n(f),\ f \in \mathcal{F}$ depends only on the data (and on the independent sequence of Rademacher random variables that can be simulated). For each $f \in \mathcal{F}$, $R_n(f)$ is essentially the "correlation coefficient" between the values of the function f at data points and independent Rademacher noise. The fact that the sup-norm $\|R_n\|_{\mathcal{F}}$ of the Rademacher process is "large" means that there exists a function $f \in \mathcal{F}$ that fits the Rademacher noise very well. This usually means that the class of functions is too complex for the purposes of statistical estimation and performing empirical risk minimization over such a class is likely to lead to overfitting. Thus, the size of sup-norms or local sup-norms of the Rademacher process provides natural data dependent measures of complexity of function classes used in statistical estimation. Symmetrization inequalities well known in the theory of empirical processes show that the expected sup-norms of Rademacher processes are within a constant from the corresponding sup-norms of the empirical process. Moreover, using concentration inequalities, one can directly relate the sup-norms of these two processes.

The δ-minimal sets (the level sets) of the true risk involved in the construction of bounds $\bar{\delta}_n(\mathcal{F})$ can be estimated by the level sets of the empirical risk. This is based on *ratio type inequalities* for the excess risk, i.e., on bounding the following probabilities

$$\mathbb{P}\biggl\{\sup_{f\in\mathcal{F},\mathcal{E}_P(f)\geq\delta}\biggl|\frac{\mathcal{E}_{P_n}(f)}{\mathcal{E}_P(f)} - 1\biggr| \geq \varepsilon\biggr\}.$$

This problem is closely related to the study of ratio type empirical processes (see Giné, Koltchinskii and Wellner [49], Giné and Koltchinskii [50] and references therein). Finally, the $L_2(P)$-diameter of the δ-minimal sets of P can be estimated by the $L_2(P_n)$-diameter of the δ-minimal sets of P_n. Thus, we can estimate all the distribution dependent parameters involved in the construction of $\bar{\delta}_n(\mathcal{F})$ by their empirical versions and, as a result, construct data-dependent upper bounds on the excess risk $\mathcal{E}_P(\hat{f})$ that hold with a guaranteed high probability. The proofs of these facts heavily rely on Talagrand's concentration inequalities for empirical processes.

1.4 Penalized Empirical Risk Minimization and Oracle Inequalities

The data-dependent bounds on the excess risk can be used in general model selection techniques in abstract empirical risk minimization problems. In such problems, there is a need to deal with minimizing the risk over a very large class of functions $\mathcal{F}$, and there is a specified family (”a sieve”) of subclasses $\{\mathcal{F}_\alpha, \alpha \in A\}$ of varying complexity that are used to approximate functions from $\mathcal{F}$. Often, classes $\mathcal{F}_\alpha$ correspond to different statistical models. Instead of one empirical risk minimization problem (1.2) one has to deal now with a family of problems

$$P_n f \longrightarrow \min,\ f \in \mathcal{F}_\alpha,\ \alpha \in A, \tag{1.4}$$

that has a set of solutions $\{\hat{f}_{n,\alpha} : \alpha \in A\}$. In many cases, there is a natural way to measure the quality of the solution of each of the problems (1.4). For instance, it can be based on distribution dependent upper bounds $\bar{\delta}_n(\alpha) = \bar{\delta}_n(\mathcal{F}_\alpha)$ on the excess risk $\mathcal{E}_P(\mathcal{F}_\alpha; \hat{f}_{n,\alpha})$ discussed above. The goal of model selection is to provide a data driven (adaptive) choice $\hat{\alpha} = \hat{\alpha}(X_1, \ldots, X_n)$ of model index α such that the empirical risk minimization over the class $\mathcal{F}_{\hat{\alpha}}$ results in an estimator $\hat{f} = \hat{f}_{n,\hat{\alpha}}$ with the nearly ”optimal” excess risk $\mathcal{E}_P(\mathcal{F}; \hat{f})$. One of the most important approaches to model selection is based on penalized empirical risk minimization, i.e. on solving the following problem

$$\hat{\alpha} := \mathrm{argmin}_{\alpha \in A}\Biggl[\min_{f \in \mathcal{F}_\alpha} P_n f + \hat{\pi}_n(\alpha)\Biggr], \tag{1.5}$$

where $\hat{\pi}_n(\alpha), \alpha \in A$ are properly chosen complexity penalties. Often, $\hat{\pi}_n(\alpha)$ is designed as a data dependent upper bound on $\bar{\delta}_n(\alpha)$, the ”desired accuracy” of empirical risk minimization for the class $\mathcal{F}_\alpha$. This approach has been developed under several different names (Vapnik-Chervonenkis structural risk minimization, method of sieves, etc.). Sometimes, it is convenient to write penalized empirical risk minimization problem in the following form

$$\hat{f} := \mathrm{argmin}_{f \in \mathcal{F}}\Biggl[P_n f + \mathrm{pen}(n; f)\Biggr],$$

where $\mathrm{pen}(n; \cdot)$ is a real valued complexity penalty defined on $\mathcal{F}$. Denoting, for each $\alpha \in \mathbb{R}$,

$$\mathcal{F}_\alpha := \{f \in \mathcal{F} : \mathrm{pen}(n; f) = \alpha\}$$

and defining $\hat{\pi}_n(\alpha) = \alpha$, the problem can be again rewritten as (1.5).

The bounds on the excess risk of $\hat{f} = \hat{f}_{n,\hat{\alpha}}$ of the following type (with some constant C)

$$\mathcal{E}_P(\mathcal{F};\hat{f}) \leq C \inf_{\alpha\in A}\left[\inf_{f\in\mathcal{F}_\alpha}\mathcal{E}_P(f) + \bar{\delta}_n(\alpha)\right] \tag{1.6}$$

that hold with a high probability are often used to express the optimality of model selection. The meaning of these inequalities can be explained as follows. Imagine that the minimum of the true risk in the class $\mathcal{F}$ is attained in a subclass $\mathcal{F}_\alpha$ for some $\alpha = \alpha(P)$. If there were an oracle that knew the model index $\alpha(P)$, then with the help of the oracle one could achieve the excess risk at least as small as $\bar{\delta}_n(\alpha(P))$. The model selection method for which the inequality (1.6) holds is not using the help of the oracle. However, it follows from (1.6) that the excess risk of the resulting estimator is upper bounded by $C\bar{\delta}_n(\alpha(P))$ (which is within a constant of the performance of the oracle).

1.5 Concrete Empirical Risk Minimization Problems

Density estimation. The most popular method of statistical estimation, the maximum likelihood method, can be viewed as a special case of empirical risk minimization. Let μ be a σ-finite measure on $(S,\mathcal{A})$ and let $\mathcal{P}$ be a statistical model, i.e., $\mathcal{P}$ is a family of probability densities with respect to μ. In particular, $\mathcal{P}$ can be a parametric model with a parameter set Θ, $\mathcal{P} = \{p(\theta,\cdot) : \theta\in\Theta\}$. A maximum likelihood estimator of unknown density $p_*\in\mathcal{P}$ based on i.i.d. observations $X_1,\dots,X_n$ sampled from p_* is a solution of the following empirical risk minimization problem

$$n^{-1}\sum_{j=1}^n\Bigl(-\log p(X_j)\Bigr) \longrightarrow \min,\ p\in\mathcal{P}. \tag{1.7}$$

Another popular approach to density estimation is based on the following penalized empirical risk minimization problem

$$-\frac{2}{n}\sum_{j=1}^n p(X_j) + \|p\|^2_{L_2(\mu)} \longrightarrow \min,\ p\in\mathcal{P}. \tag{1.8}$$

This approach can be explained as follows. The best $L_2(\mu)$- approximation of the density p_* is obtained by solving

$$\|p-p_*\|^2_{L_2(\mu)} = -2\int_S pp_*d\mu + \|p\|^2_{L_2(\mu)} + \|p_*\|^2_{L_2(\mu)} \longrightarrow \min,\ p\in\mathcal{P}.$$

The integral $\int_S pp_*d\mu = \mathbb{E}p(X)$ can be estimated by $n^{-1}\sum_{j=1}^n p(X_j)$, leading to problem (1.8). Of course, in the case of complex enough models $\mathcal{P}$, there might be a need in complexity penalization in (1.7) and (1.8).

Prediction problems. Empirical risk minimization is especially useful in a variety of prediction problems. In these problems, the data consists of i.i.d. couples $(X_1, Y_1), \dots (X_n, Y_n)$ in $S \times T$ with common distribution P. Assume that $T \subset \mathbb{R}$. Given another couple (X, Y) sampled from P, the goal is to predict Y based on an observation of X. To formalize this problem, introduce a loss function $\ell : T \times \mathbb{R} \mapsto \mathbb{R}_+$. Given $g : S \mapsto \mathbb{R}$, denote $(\ell \bullet g)(x, y) := \ell(y, g(x))$, which will be interpreted as a loss suffered as a result of using $g(x)$ to predict y. Then the risk associated with "action" g is defined as

$$P(\ell \bullet g) = \mathbb{E}\ell(Y, g(X)).$$

Given a set $\mathcal{G}$ of possible actions g, we want to minimize the risk:

$$P(\ell \bullet g) \longrightarrow \min, \ g \in \mathcal{G}.$$

The risk can be estimated based on the data $(X_1, Y_1), \dots, (X_n, Y_n)$, which leads to the following empirical risk minimization problem

$$P_n(\ell \bullet g) = n^{-1} \sum_{j=1}^{n} \ell(Y_j, g(X_j)) \longrightarrow \min, \ g \in \mathcal{G}.$$

Introducing the notation $f := \ell \bullet g$ and setting $\mathcal{F} := \{\ell \bullet g : g \in \mathcal{G}\}$, one can rewrite the problems in the form (1.1), (1.2).

Regression and classification are two most common examples of prediction problems. In regression problems, the loss function is usually defined as $\ell(y; u) = \phi(y - u)$, where ϕ is, most often, nonnegative, even and convex function with $\phi(0) = 0$. The empirical risk minimization becomes

$$n^{-1} \sum_{j=1}^{n} \phi(Y_j - g(X_j)) \longrightarrow \min, \ g \in \mathcal{G}.$$

The choice $\phi(u) = u^2$ is by far the most popular and it means fitting the regression model using the least square method.

In the case of *binary classification* problems, $T := \{-1, 1\}$ and it is natural to consider a class $\mathcal{G}$ of binary functions (classifiers) $g : S \mapsto \{-1, 1\}$ and to use the binary loss $\ell(y; u) = I(y \neq u)$. The risk of a classifier g with respect to the binary loss

$$P(\ell \bullet g) = \mathbb{P}\{Y \neq g(X)\}$$

is just the probability to misclassify and, in learning theory, it is known as *the generalization error.* A binary classifier that minimizes the generalization error over all measurable binary functions is called *the Bayes classifier* and its generalization error is called *the Bayes risk.* The corresponding empirical risk

$$P_n(\ell \bullet g) = n^{-1} \sum_{j=1}^{n} I(Y_j \neq g(X_j))$$

is known as *the training error.* Minimizing the training error over $\mathcal{G}$

$$n^{-1} \sum_{j=1}^{n} I(Y_j \neq g(X_j)) \longrightarrow \min,\ g \in \mathcal{G}$$

is, usually, a computationally intractable problem (with an exception of very simple families of classifiers $\mathcal{G}$) since the functional to be minimized lacks convexity, smoothness or any other form of regularity.

Large margin classification. Large margin classification methods are based on the idea of considering real valued classifiers $g : S \mapsto \mathbb{R}$ instead of binary classifiers and replacing the binary loss by a convex "surrogate loss". A real valued classifier g can be easily transformed into binary: $g \mapsto \mathrm{sign}(g)$. Define $\ell(y, u) := \phi(yu)$, where $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ is a convex nonincreasing function such that $\phi(u) \geq I_{(-\infty,0]}(u), u \in \mathbb{R}$. The product $Yg(X)$ is called *the margin* of classifier g on the training example (X, Y). If $Yg(X) \geq 0$, (X, Y) is correctly classified by g, otherwise the example is misclassified. Given a convex set $\mathcal{G}$ of classifiers $g : S \mapsto \mathbb{R}$ the risk minimization problem becomes

$$P(\ell \bullet g) = \mathbb{E}\phi(Yg(X)) \longrightarrow \min,\ g \in \mathcal{G}$$

and its empirical version is

$$P_n(\ell \bullet g) = n^{-1} \sum_{j=1}^{n} \phi(Y_j g(X_j)) \longrightarrow \min,\ g \in \mathcal{G}, \tag{1.9}$$

which are convex optimization problems.

It is well known that, under very mild conditions on the "surrogate loss" ϕ (so called *classification calibration,*see, e.g., [10]) the solution g_* of the problem

$$P(\ell \bullet g) = \mathbb{E}\phi(Yg(X)) \longrightarrow \min,\ g : S \mapsto \mathbb{R}$$

possesses the property that $\mathrm{sign}(g_*)$ is the Bayes classifier. Thus, it becomes plausible that the empirical risk minimization problem (1.9) with a large enough and properly chosen convex function class $\mathcal{G}$ would have a solution $\hat{g}$ such that the generalization error of the binary classifier $\mathrm{sign}(\hat{g})$ is close enough to the Bayes risk. Because of the nature of the loss function (heavy penalization for negative and even small positive margins), the solution $\hat{g}$ tends to be a classifier with most of the margins on the training data positive and large, which explains the name "large margin classifiers".

Among common choices of the surrogate loss function are $\phi(u) = e^{-u}$ (the exponential loss), $\phi(u) = \log_2(1 + e^{-u})$ (the logit loss) and $\phi(u) = (1 - u) \vee 0$ (the hinge loss).

A possible choice of class $\mathcal{G}$ is

$$\mathcal{G} := \mathrm{conv}(\mathcal{H}) := \left\{ \sum_{j=1}^{N} \lambda_j h_j, N \geq 1, \lambda_j \geq 0, \sum_{j=1}^{N} \lambda_j h_j, h_j \in \mathcal{H} \right\},$$

where $\mathcal{H}$ is a given *base class* of classifiers. Usually, $\mathcal{H}$ consists of binary classifiers and it is a rather simple class class such that the direct minimization of the training error over $\mathcal{H}$ is computationally tractable. The problem (1.9) is then solved by a version of gradient descent algorithm in functional space. This leads to a family of classification methods called *boosting* (also, voting methods, ensemble methods, etc). Classifiers output by boosting are convex combinations of base classifiers and the whole method is often interpreted in machine learning literature as a way to combine simple base classifiers into more complex and powerful classifiers with a much better generalization performance.

Another popular approach is based on penalized empirical risk minimization in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ generated by a symmetric nonnegatively definite kernel $K : S \times S \mapsto \mathbb{R}$. For instance, using the square of the norm as a penalty results in the following problem:

$$n^{-1} \sum_{j=1}^{n} \phi(Y_j g(X_j)) + \varepsilon \|g\|_{\mathcal{H}_K}^2 \longrightarrow \min, \ g \in \mathcal{H}_K, \tag{1.10}$$

where $\varepsilon > 0$ is a regularization parameter. In the case of hinge loss $\phi(u) = (1 - u) \vee 0$ the method is called *support vector machine* (SVM). By the basic properties of RKHS, a function $g \in \mathcal{H}_K$ can be represented as $g(x) = \langle g, K(x, \cdot) \rangle_{\mathcal{H}_K}$. Because of this, it is very easy to conclude that the solution $\hat{g}$ of (1.10) must be in the linear span of functions $K(X_1, \cdot), \ldots, K(X_n, \cdot)$. Thus, the problem (1.10) is essentially a finite dimensional convex problem (in the case of hinge loss, it becomes a quadratic programming problem).

1.6 Sparse Recovery Problems

Let $\mathcal{H} = \{h_1, \dots, h_N\}$ be a given set of functions from S into $\mathbb{R}$ called *a dictionary.* Given $\lambda \in \mathbb{R}^N$, denote $f_\lambda = \sum_{j=1}^N \lambda_j h_j$. Suppose that a function $f_* \in \text{l.s.}(\mathcal{H})$ is observed at random points $X_1, \dots, X_n$ with common distribution Π,

$$Y_j = f_*(X_j),\ j = 1, \dots, n$$

being the observations. The goal is to find a representation of f_* in the dictionary, i.e., to find $\lambda \in \mathbb{R}^N$ such that

$$f_\lambda(X_j) = Y_j,\ j = 1, \dots, n. \tag{1.11}$$

In the case when the functions in the dictionary are not linearly independent, such a representation does not have to be unique. Moreover, if $N > n$, the system of linear equations (1.11) is underdetermined and the set

$$L := \Big\{\lambda \in \mathbb{R}^N : f_\lambda(X_j) = Y_j, j = 1, \dots, n\Big\}$$

is a nontrivial affine subspace of $\mathbb{R}^N$. However, even in this case, the following problem still makes sense:

$$\|\lambda\|_{\ell_0} = \sum_{j=1}^N I(\lambda_j \neq 0) \longrightarrow \min, \lambda \in L. \tag{1.12}$$

In other words, the goal is to find *the sparsest solution* of the linear system (1.11). In general, the sparse recovery problem (1.12) is not computationally tractable since solving such a nonconvex optimization problem essentially requires searching through all 2^N coordinate subspaces of $\mathbb{R}^N$ and then solving the corresponding linear systems. However, the following problem

$$\|\lambda\|_{\ell_1} = \sum_{j=1}^N |\lambda_j| \longrightarrow \min, \lambda \in L. \tag{1.13}$$

is convex, and, moreover, it is a linear programming problem. It happens that for some dictionaries $\mathcal{H}$ and distributions Π of design variables the solution of problem (1.13) is unique and coincides with the sparsest solution λ_* of the problem (1.12) (provided that $\|\lambda_*\|_{\ell_0}$ is sufficiently small). This fact is closely related to some problems in convex geometry concerning the neighborliness of convex polytopes.

More generally, one can study sparse recovery problems in the case when f_* does not necessarily belong to the linear span of the dictionary $\mathcal{H}$ and it is measured at random locations X_j with some errors. Given i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ and a loss function ℓ, this naturally leads to the study of the following penalized empirical risk minimization problem

$$\hat{\lambda}^{\varepsilon} := \text{argmin}_{\lambda \in \mathbb{R}^N} \left[P_n(\ell \bullet f_{\lambda}) + \varepsilon \|\lambda\|_{\ell_1} \right] \tag{1.14}$$

which is an empirical version of the problem

$$\lambda^{\varepsilon} := \text{argmin}_{\lambda \in \mathbb{R}^N} \left[P(\ell \bullet f_{\lambda}) + \varepsilon \|\lambda\|_{\ell_1} \right], \tag{1.15}$$

where $\varepsilon > 0$ is a regularization parameter. It is assumed that the loss function $\ell(y; u)$ is convex with respect to u which makes the optimization problems (1.14) and (1.15) convex. This framework includes sparse recovery in both regression and large margin classification contexts. In the case of regression with quadratic loss $\ell(y, u) = (y - u)^2$, this penalization method has been called LASSO in statistical literature. The sparse recovery algorithm (1.13) can be viewed as a version of (1.14) with quadratic loss and with $\varepsilon = 0$.

Another popular method of sparse recovery introduced recently by Candes and Tao [29] and called *the Dantzig selector* is based on solving the following linear programming problem

$$\hat{\lambda}^{\varepsilon} \in \text{Argmin}_{\lambda \in \hat{\Lambda}^{\varepsilon}} \|\lambda\|_{\ell_1},$$

where

$$\hat{\Lambda}^{\varepsilon} := \left\{ \lambda \in \mathbb{R}^N : \max_{1 \le k \le N} \left| n^{-1} \sum_{j=1}^{n} (f_{\lambda}(X_j) - Y_j) h_k(X_j) \right| \le \varepsilon/2 \right\}.$$

Note that the conditions defining the set $\hat{\Lambda}^{\varepsilon}$ are just necessary conditions of extremum in the LASSO-optimization problem

$$n^{-1} \sum_{j=1}^{n} (Y_j - f_{\lambda}(X_j))^2 + \varepsilon \|\lambda\|_{\ell_1} \longrightarrow \min, \ \lambda \in \mathbb{R}^N,$$

so, the Dantzig selector is closely related to LASSO.

We will also study some other types of penalties that can be used in sparse recovery problems such as $\|\lambda\|_{\ell_p}^p$ with a suitable value of $p > 1$ and entropy penalty $\sum_{j=1}^N \lambda_j \log \lambda_j$ that can be used for sparse recovery in the convex hull of the dictionary $\mathcal{H}$.

Our goal will be to establish oracle inequalities showing that the methods of this type allow one to find a sparse approximation of the target function (when it exists).

2 Empirical and Rademacher Processes

The empirical process is defined as

$$Z_n := n^{1/2}(P_n - P)$$

and it can be viewed as a random measure. However, more often, it has been viewed as a stochastic process indexed by a function class $\mathcal{F}$:

$$Z_n(f) = n^{1/2}(P_n - P)(f), f \in \mathcal{F}$$

(see Dudley [42] or van der Vaart and Wellner [95]).

The Rademacher process indexed by a class $\mathcal{F}$ was defined in Section 1.3 as

$$R_n(f) := n^{-1} \sum_{i=1}^{n} \varepsilon_i f(X_i),\ f \in \mathcal{F},$$

$\{\varepsilon_i\}$ being i.i.d. Rademacher random variables (i.e., ε_i takes the values $+1$ and -1 with probability 1/2 each) independent of $\{X_i\}$.

It should be mentioned that certain measurability assumptions are required in the study of empirical and Rademacher processes. In particular, under these assumptions, such quantities as $\|P_n - P\|_{\mathcal{F}}$ are properly measurable random variables. We refer to the books of Dudley [42], Chapter 5 and van der Vaart and Wellner [95], Section 1.7 for precise formulations of these measurability assumptions. Some of the bounds derived and used below hold even without the assumptions of this nature, if the expectation is replaced by outer expectation, as it is often done, for instance, in [95]. Another option is to "define"

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} := \sup\biggl\{\mathbb{E}\|P_n - P\|_{\mathcal{G}} : \mathcal{G} \subset \mathcal{F}, \mathcal{G} \text{ is finite}\biggr\},$$

which provides a simple way to get around the measurability difficulties. Such an approach has been frequently used by Talagrand (see, e.g., [88]). In what follows, it will be assumed that measurability problems have been resolved in one of these ways.

2.1 Symmetrization Inequalities

The following important inequality reveals close relationships between empirical and Rademacher processes.

Theorem 2.1 *For any class $\mathcal{F}$ of P-integrable functions and for any convex function $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$*

$$\mathbb{E}\Phi\Bigl(\frac{1}{2}\|R_n\|_{\mathcal{F}_c}\Bigr) \leq \mathbb{E}\Phi\Bigl(\|P_n - P\|_{\mathcal{F}}\Bigr) \leq \mathbb{E}\Phi\Bigl(2\|R_n\|_{\mathcal{F}}\Bigr),$$

where $\mathcal{F}_c := \{f - Pf : f \in \mathcal{F}\}$. In particular,

$$\frac{1}{2}\mathbb{E}\|R_n\|_{\mathcal{F}_c} \leq \mathbb{E}\|P_n - P\|_{\mathcal{F}} \leq 2\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

Proof. Assume that the random variables $X_1, \dots X_n$ are defined on a probability space $(\bar{\Omega}, \bar{\Sigma}, \bar{\mathbb{P}})$. We will also need two other probability spaces: $(\tilde{\Omega}, \tilde{\Sigma}, \tilde{\mathbb{P}})$ and $(\Omega_\varepsilon, \Sigma_\varepsilon, \mathbb{P}_\varepsilon)$. The main probability space on which all the random variables are defined will be denoted $(\Omega, \Sigma, \mathbb{P})$ and it will be the product space

$$(\Omega, \Sigma, \mathbb{P}) = (\bar{\Omega}, \bar{\Sigma}, \bar{\mathbb{P}}) \times (\tilde{\Omega}, \tilde{\Sigma}, \tilde{\mathbb{P}}) \times (\Omega_\varepsilon, \Sigma_\varepsilon, \mathbb{P}_\varepsilon).$$

The corresponding expectations will be denoted by $\bar{\mathbb{E}}, \tilde{\mathbb{E}}, \mathbb{E}_\varepsilon$ and $\mathbb{E}$. Let $(\tilde{X}_1, \dots, \tilde{X}_n)$ be an independent copy of $(X_1, \dots, X_n)$. Think of random variables $\tilde{X}_1, \dots, \tilde{X}_n$ as being defined on $(\tilde{\Omega}, \tilde{\Sigma}, \tilde{\mathbb{P}})$. Denote $\tilde{P}_n$ the empirical measure based on $(\tilde{X}_1, \dots, \tilde{X}_n)$ (it is an independent copy of P_n). Then $\tilde{\mathbb{E}}\tilde{P}_n f = Pf$ and, using Jensen's inequality,

$$\mathbb{E}\Phi\Bigl(\|P_n - P\|_{\mathcal{F}}\Bigr) = \bar{\mathbb{E}}\Phi\Bigl(\|P_n - \tilde{\mathbb{E}}\tilde{P}_n\|_{\mathcal{F}}\Bigr) = \bar{\mathbb{E}}\Phi\Bigl(\|\tilde{\mathbb{E}}(P_n - \tilde{P}_n)\|_{\mathcal{F}}\Bigr) \leq$$

$$\bar{\mathbb{E}}\tilde{\mathbb{E}}\Phi\Bigl(\|P_n - \tilde{P}_n\|_{\mathcal{F}}\Bigr) = \bar{\mathbb{E}}\tilde{\mathbb{E}}\Phi\Biggl(\Bigl\|n^{-1}\sum_{j=1}^n (\delta_{X_j} - \delta_{\tilde{X}_j})\Bigr\|_{\mathcal{F}}\Biggr).$$

Since $X_1, \dots, X_n, \tilde{X}_1, \dots, \tilde{X}_n$ are i.i.d., the distribution of $(X_1, \dots, X_n, \tilde{X}_1, \dots, \tilde{X}_n)$ is invariant with respect to all permutations of the components. In particular, one can switch any couple $X_j, \tilde{X}_j$. Because of this,

$$\bar{\mathbb{E}}\tilde{\mathbb{E}}\Phi\Biggl(\Bigl\|n^{-1}\sum_{j=1}^n (\delta_{X_j} - \delta_{\tilde{X}_j})\Bigr\|_{\mathcal{F}}\Biggr) = \bar{\mathbb{E}}\tilde{\mathbb{E}}\Phi\Biggl(\Bigl\|n^{-1}\sum_{j=1}^n \sigma_j(\delta_{X_j} - \delta_{\tilde{X}_j})\Bigr\|_{\mathcal{F}}\Biggr),$$

for arbitrary choice of $\sigma_j = +1$ or $\sigma_j = -1$. Define now i.i.d. Rademacher random variables on $(\Omega_\varepsilon, \Sigma_\varepsilon, \mathbb{P}_\varepsilon)$ (thus, independent of $(X_1, \dots, X_n, \tilde{X}_1, \dots, \tilde{X}_n)$). Then, we have

$$\bar{\mathbb{E}}\tilde{\mathbb{E}}\Phi\Biggl(\Bigl\|n^{-1}\sum_{j=1}^n (\delta_{X_j} - \delta_{\tilde{X}_j})\Bigr\|_{\mathcal{F}}\Biggr) = \mathbb{E}_\varepsilon\bar{\mathbb{E}}\tilde{\mathbb{E}}\Phi\Biggl(\Bigl\|n^{-1}\sum_{j=1}^n \varepsilon_j(\delta_{X_j} - \delta_{\tilde{X}_j})\Bigr\|_{\mathcal{F}}\Biggr)$$

and the proof can be completed as follows:

$$\mathbb{E}\Phi\Bigl(\|P_n - P\|_{\mathcal{F}}\Bigr) \le \mathbb{E}_\varepsilon \bar{\mathbb{E}}\tilde{\mathbb{E}}\Phi\biggl(\biggl\|n^{-1}\sum_{j=1}^n \varepsilon_j(\delta_{X_j} - \delta_{\tilde{X}_j})\biggr\|_{\mathcal{F}}\biggr) \le$$

$$\frac{1}{2}\mathbb{E}_\varepsilon \bar{\mathbb{E}}\Phi\biggl(2\biggl\|n^{-1}\sum_{j=1}^n \varepsilon_j\delta_{X_j}\biggr\|_{\mathcal{F}}\biggr) + \frac{1}{2}\mathbb{E}_\varepsilon \tilde{\mathbb{E}}\Phi\biggl(2\biggl\|n^{-1}\sum_{j=1}^n \varepsilon_j\delta_{\tilde{X}_j}\biggr\|_{\mathcal{F}}\biggr) = \mathbb{E}\Phi\Bigl(2\|R_n\|_{\mathcal{F}}\Bigr).$$

The proof of the lower bound is similar.

□

The upper bound is called the *symmetrization inequality* and the lower bound is often called the *desymmetrization inequality.* These inequalities were introduced to the theory of empirical processes by Giné and Zinn [47] (an earlier form of Rademacher symmetrization was used by Koltchinskii [57]) and Pollard [80]). The desymmetrization inequality is often used together with the following elementary lower bound (in the case of $\Phi(u) = u$)

$$\mathbb{E}\|R_n\|_{\mathcal{F}_c} \ge \mathbb{E}\|R_n\|_{\mathcal{F}} - \sup_{f\in\mathcal{F}}|Pf|\ \mathbb{E}|R_n(1)| \ge$$

$$\ge \mathbb{E}\|R_n\|_{\mathcal{F}} - \sup_{f\in\mathcal{F}}|Pf|\ \mathbb{E}^{1/2}|n^{-1}\sum_{j=1}^n \varepsilon_j|^2 \ge \mathbb{E}\|R_n\|_{\mathcal{F}} - \frac{\sup_{f\in\mathcal{F}}|Pf|}{\sqrt{n}}.$$

2.2 Comparison Inequalities for Rademacher Sums

Given a set $T \subset \mathbb{R}^n$ and i.i.d. Rademacher variables $\varepsilon_i, i = 1, 2, \dots,$ it is of interest to know how the expected value of the sup-norm of Rademacher sums indexed by T

$$R_n(T) := \mathbb{E}\sup_{t\in T}\biggl|\sum_{i=1}^n t_i\varepsilon_i\biggr|$$

depends on the geometry of the set T.

The following beautiful *comparison inequality* for Rademacher sums is due to Talagrand (see Ledoux and Talagrand [68], Theorem 4.12).

Theorem 2.2 *Let $T \subset \mathbb{R}^n$ and let $\varphi_i : \mathbb{R} \mapsto \mathbb{R},\ i = 1, \dots, n$ be functions such that $\varphi_i(0) = 0$ and*

$$|\varphi_i(u) - \varphi_i(v)| \le |u - v|,\ u, v \in \mathbb{R}$$

(i.e., φ_i is a contraction). For all convex nondecreasing functions $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$,

$$\mathbb{E}\Phi\biggl(\frac{1}{2}\sup_{t\in T}\biggl|\sum_{i=1}^n \varphi_i(t_i)\varepsilon_i\biggr|\biggr) \le \mathbb{E}\Phi\biggl(\sup_{t\in T}\biggl|\sum_{i=1}^n t_i\varepsilon_i\biggr|\biggr).$$

Proof. First, we prove that for a nondecreasing convex function $\Phi : \mathbb{R} \mapsto \mathbb{R}_+$ and for an arbitrary $A : T \mapsto \mathbb{R}$

$$\mathbb{E}\Phi\Big(\sup_{t\in T}\Big[A(t) + \sum_{i=1}^{n} \varphi_i(t_i)\varepsilon_i\Big]\Big) \leq \mathbb{E}\Phi\Big(\sup_{t\in T}\Big[A(t) + \sum_{i=1}^{n} t_i\varepsilon_i\Big]\Big). \tag{2.1}$$

We start with the case $n = 1$. Then, the bound is equivalent to the following

$$\mathbb{E}\Phi\Big(\sup_{t\in T}[t_1 + \varepsilon\varphi(t_2)]\Big) \leq \mathbb{E}\Phi\Big(\sup_{t\in T}[t_1 + \varepsilon t_2]\Big)$$

for an arbitrary set $T \subset \mathbb{R}^2$ and an arbitrary contraction φ. One can rewrite it as

$$\frac{1}{2}\Big(\Phi\Big(\sup_{t\in T}[t_1 + \varphi(t_2)]\Big) + \Phi\Big(\sup_{t\in T}[t_1 - \varphi(t_2)]\Big)\Big) \leq \frac{1}{2}\Big(\Phi\Big(\sup_{t\in T}[t_1 + t_2]\Big) + \Phi\Big(\sup_{t\in T}[t_1 - t_2]\Big)\Big).$$

If now $(t_1, t_2) \in T$ denote a point where $\sup_{t\in T}[t_1 + \varphi(t_2)]$ is attained and $(s_1, s_2) \in T$ is a point where $\sup_{t\in T}[t_1 - \varphi(t_2)]$ is attained, then it is enough to show that

$$\Phi\Big(t_1 + \varphi(t_2)\Big) + \Phi\Big(s_1 - \varphi(s_2)\Big) \leq \Phi\Big(\sup_{t\in T}[t_1 + t_2]\Big) + \Phi\Big(\sup_{t\in T}[t_1 - t_2]\Big)$$

(if the suprema are not attained, one can easily modify the argument). Clearly, we have the following conditions:

$$t_1 + \varphi(t_2) \geq s_1 + \varphi(s_2) \text{ and } t_1 - \varphi(t_2) \leq s_1 - \varphi(s_2).$$

First consider the case when $t_2 \geq 0, s_2 \geq 0$ and $t_2 \geq s_2$. In this case, we will prove that

$$\Phi\Big(t_1 + \varphi(t_2)\Big) + \Phi\Big(s_1 - \varphi(s_2)\Big) \leq \Phi\Big(t_1 + t_2\Big) + \Phi\Big(s_1 - s_2\Big), \tag{2.2}$$

which would imply the bound. Indeed, for

$$a := t_1 + \varphi(t_2), b := t_1 + t_2, c := s_1 - s_2, d := s_1 - \varphi(s_2),$$

we have $a \leq b$ and $c \leq d$ since

$$\varphi(t_2) \leq t_2, \quad \varphi(s_2) \leq s_2$$

(by the assumption that φ is a contraction and $\varphi(0) = 0$). We also have that

$$b - a = t_2 - \varphi(t_2) \geq s_2 - \varphi(s_2) = d - c,$$

because again φ is a contraction and $t_2 \geq s_2$. Finally, we have

$$a = t_1 + \varphi(t_2) \geq s_1 + \varphi(s_2) \geq s_1 - s_2 = c.$$

Since the function Φ is nondecreasing and convex, its increment over the interval $[a, b]$ is larger than its increment over the interval $[c, d]$ ($[a, b]$ is longer than $[c, d]$ and $a \geq c$), which is equivalent to (2.2).

If $t_2 \geq 0, s_2 \geq 0$ and $s_2 \geq t_2$, it is enough to use the change of notations $(t, s) \mapsto (s, t)$ and to replace φ with $-\varphi$.

The case $t_2 \leq 0, s_2 \leq 0$ can be now handled by using the transformation $(t_1, t_2) \mapsto (t_1, -t_2)$ and changing the function φ accordingly.

We have to consider the case $t_2 \geq 0, s_2 \leq 0$ (the only remaining case $t_2 \leq 0, s_2 \geq 0$ would again follow by switching the names of t and s and replacing φ with $-\varphi$). In this case, we have

$$\varphi(t_2) \leq t_2, \quad -\varphi(s_2) \leq -s_2,$$

which, in view of monotonicity of Φ, immediately implies

$$\Phi\Big(t_1 + \varphi(t_2)\Big) + \Phi\Big(s_1 - \varphi(s_2)\Big) \leq \Phi\Big(t_1 + t_2\Big) + \Phi\Big(s_1 - s_2\Big).$$

This completes the proof of (2.1) in the case $n = 1$.

In the general case, we have

$$\mathbb{E}\Phi\biggl(\sup_{t\in T}\biggl[A(t) + \sum_{i=1}^{n} \varphi_i(t_i)\varepsilon_i\biggr]\biggr) = \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_{n-1}}\mathbb{E}_{\varepsilon_n}\Phi\biggl(\sup_{t\in T}\biggl[A(t) + \sum_{i=1}^{n-1} \varphi_i(t_i)\varepsilon_i + \varepsilon_n\varphi(t_n)\biggr]\biggr).$$

The expectation $\mathbb{E}_{\varepsilon_n}$ (conditional on $\varepsilon_1, \dots, \varepsilon_{n-1}$) can be bounded using the result in the case $n = 1$. This yields (after changing the order of integration)

$$\mathbb{E}\Phi\biggl(\sup_{t\in T}\biggl[A(t) + \sum_{i=1}^{n} \varphi_i(t_i)\varepsilon_i\biggr]\biggr) \leq \mathbb{E}_{\varepsilon_n}\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_{n-1}}\Phi\biggl(\sup_{t\in T}\biggl[A(t) + \varepsilon_n t_n + \sum_{i=1}^{n-1} \varphi_i(t_i)\varepsilon_i\biggr]\biggr).$$

The proof of (2.1) can now be completed by an induction argument.

Finally, to prove the inequality of the theorem, it is enough to write

$$\mathbb{E}\Phi\biggl(\frac{1}{2}\sup_{t\in T}\biggl|\sum_{i=1}^{n} \varphi_i(t_i)\varepsilon_i\biggr|\biggr) = \mathbb{E}\Phi\biggl(\frac{1}{2}\biggl[\biggl(\sup_{t\in T}\sum_{i=1}^{n} \varphi_i(t_i)\varepsilon_i\biggr)_+ + \biggl(\sup_{t\in T}\sum_{i=1}^{n} \varphi_i(t_i)(-\varepsilon_i)\biggr)_+\biggr]\biggr) \leq$$

$$\frac{1}{2}\biggl[\mathbb{E}\Phi\biggl(\biggl(\sup_{t\in T}\sum_{i=1}^{n} \varphi_i(t_i)\varepsilon_i\biggr)_+\biggr) + \mathbb{E}\Phi\biggl(\biggl(\sup_{t\in T}\sum_{i=1}^{n} \varphi_i(t_i)(-\varepsilon_i)\biggr)_+\biggr)\biggr],$$

where $a_+ := a \vee 0$. Applying the inequality (2.1) to the function $u \mapsto \Phi(u_+)$, which is convex and nondecreasing, completes the proof.

□

We will frequently use a corollary of the above comparison inequality that provides upper bounds on the moments of the sup-norm of Rademacher process R_n on the class

$$\varphi \circ \mathcal{F} := \{\varphi \circ f : f \in \mathcal{F}\}$$

in terms of the corresponding moments of the sup-norm of R_n on $\mathcal{F}$ and Lipschitz constant of function φ.

Theorem 2.3 *Let $\varphi : \mathbb{R} \mapsto \mathbb{R}$ be a contraction satisfying the condition $\varphi(0) = 0$. For all convex nondecreasing functions $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$,*

$$\mathbb{E}\Phi\biggl(\frac{1}{2}\|R_n\|_{\varphi\circ\mathcal{F}}\biggr) \leq \mathbb{E}\Phi\Bigl(\|R_n\|_{\mathcal{F}}\Bigr).$$

In particular,

$$\mathbb{E}\|R_n\|_{\varphi\circ\mathcal{F}} \leq 2\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

The inequality of Theorem 2.3 will be called *the contraction inequality* for Rademacher processes.

A simple rescaling of the class $\mathcal{F}$ allows one to use the contraction inequality in the case of an arbitrary function φ satisfying the Lipschitz condition

$$|\varphi(u) - \varphi(v)| \leq L|u - v|$$

on an arbitrary interval (a, b) that contains the ranges of all the functions in $\mathcal{F}$. In this case, the last bound of Theorem 2.3 takes the form

$$\mathbb{E}\|R_n\|_{\varphi\circ\mathcal{F}} \leq 2L\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

This implies, for instance, that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\biggl|n^{-1}\sum_{i=1}^{n}\varepsilon_i f^2(X_i)\biggr| \leq 4U\mathbb{E}\sup_{f\in\mathcal{F}}\biggl|n^{-1}\sum_{i=1}^{n}\varepsilon_i f(X_i)\biggr| \tag{2.3}$$

provided that the functions in the class $\mathcal{F}$ are uniformly bounded by a constant U.

2.3 Concentration Inequalities

A well known, simple and useful concentration inequality for functions

$$Z = g(X_1, \ldots, X_n)$$

of independent random variables with values in arbitrary spaces is valid under so called *bounded difference condition* on g : there exist constants $c_j, j = 1, \dots, n$ such that for all $j = 1, \dots, n$ and all $x_1, x_2, \dots, x_j, x'_j, \dots, x_n$

$$\Big| g(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_n) - g(x_1, \dots, x_{j-1}, x'_j, x_{j+1}, \dots, x_n) \Big| \le c_j. \tag{2.4}$$

Theorem 2.4 Bounded difference inequality. *Under the condition (2.4),*

$$\mathbb{P}\{Z - \mathbb{E}Z \ge t\} \le \exp\Big\{ -\frac{2t^2}{\sum_{j=1}^n c_j^2} \Big\}$$

and

$$\mathbb{P}\{Z - \mathbb{E}Z \le -t\} \le \exp\Big\{ -\frac{2t^2}{\sum_{j=1}^n c_j^2} \Big\}.$$

A standard proof of this inequality is based on bounding the exponential moment $\mathbb{E}e^{\lambda(Z - \mathbb{E}Z)}$, using the following martingale difference representation

$$Z - \mathbb{E}Z = \sum_{j=1}^n \Big[\mathbb{E}(Z|X_1, \dots, X_j) - \mathbb{E}(Z|X_1, \dots, X_{j-1}) \Big],$$

then using Markov inequality and optimizing the resulting bound with respect to $\lambda > 0$.

In the case when $Z = X_1 + \cdots + X_n$, the bounded difference inequality coincides with Hoeffding inequality for sums of bounded independent random variables.

For a class $\mathcal{F}$ of functions uniformly bounded by a constant U, the bounded difference inequality immediately implies the following bounds for $\|P_n - P\|_{\mathcal{F}}$, providing a uniform version of Hoeffding inequality.

Theorem 2.5 *For all $t > 0$,*

$$\mathbb{P}\Big\{ \|P_n - P\|_{\mathcal{F}} \ge \mathbb{E}\|P_n - P\|_{\mathcal{F}} + \frac{tU}{\sqrt{n}} \Big\} \le \exp\{-t^2/2\}$$

and

$$\mathbb{P}\Big\{ \|P_n - P\|_{\mathcal{F}} \le \mathbb{E}\|P_n - P\|_{\mathcal{F}} - \frac{tU}{\sqrt{n}} \Big\} \le \exp\{-t^2/2\}.$$

Developing uniform versions of Bernstein's inequality happened to be a much harder problem that was solved in famous papers by Talagrand [86, 87] on concentration inequalities for product measures and empirical processes.

Theorem 2.6 Talagrand's inequality. *Let $X_1, \dots, X_n$ be independent random variables in S. For any class of functions $\mathcal{F}$ on S that is uniformly bounded by a constant $U > 0$ and for all $t > 0$*

$$\mathbb{P}\left\{\left|\left\|\sum_{i=1}^{n} f(X_i)\right\|_{\mathcal{F}} - \mathbb{E}\left\|\sum_{i=1}^{n} f(X_i)\right\|_{\mathcal{F}}\right| \geq t\right\} \leq K \exp\left\{-\frac{1}{K}\frac{t}{U}\log\left(1+\frac{tU}{V}\right)\right\},$$

where K is a universal constant and V is any number satisfying

$$V \geq \mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f^2(X_i).$$

Using symmetrization inequality and contraction inequality for the square (2.3), it is easy to show that in the case of i.i.d. random variables $X_1, \dots, X_n$ with distribution P

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f^2(X_i) \leq n\sup_{f\in\mathcal{F}} Pf^2 + 8U\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}. \tag{2.5}$$

The right hand side of this bound is a common choice of the quantity V involved in Talagrand's inequality. Moreover, in the case when $\mathbb{E}f(X) = 0$, the desymmetrization inequality yields

$$\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}} \leq 2\mathbb{E}\left\|\sum_{i=1}^{n} f(X_i)\right\|_{\mathcal{F}}.$$

As a result, one can use Talagrand's inequality with

$$V = n\sup_{f\in\mathcal{F}} Pf^2 + 16U\mathbb{E}\left\|\sum_{i=1}^{n} f(X_i)\right\|$$

and the size of $\left\|\sum_{i=1}^{n} f(X_i)\right\|_{\mathcal{F}}$ is now controlled it terms of its expectation only.

This form of Talagrand's inequality is especially convenient and there have been considerable efforts to find explicit and sharp values of the constants in such inequalities. In particular, we will frequently use the bounds proved by Bousquet [22] and Klein [54] (in fact, Klein and Rio [55] provide an improved version of this inequality). Namely, for a class $\mathcal{F}$ of measurable functions from S into $[0,1]$ (by a simple rescaling $[0,1]$ can be replaced by any bounded interval) the following bounds hold for all $t > 0$:

Bousquet bound.

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}} \geq \mathbb{E}\|P_n - P\|_{\mathcal{F}} + \sqrt{2\frac{t}{n}\left(\sigma_P^2(\mathcal{F}) + 2\mathbb{E}\|P_n - P\|_{\mathcal{F}}\right)} + \frac{t}{3n}\right\} \leq e^{-t}.$$

Klein-Rio bound

$$\mathbb{P}\Big\{\|P_n - P\|_{\mathcal{F}} \le \mathbb{E}\|P_n - P\|_{\mathcal{F}} - \sqrt{2\frac{t}{n}\Big(\sigma_P^2(\mathcal{F}) + 2\mathbb{E}\|P_n - P\|_{\mathcal{F}}\Big)} - \frac{t}{n}\Big\} \le e^{-t}.$$

Here

$$\sigma_P^2(\mathcal{F}) := \sup_{f\in\mathcal{F}}\Big(Pf^2 - (Pf)^2\Big).$$

Concentration inequalities can be also applied to the Rademacher process which can be viewed as an empirical process based on the sample $(X_1, \varepsilon_1), \dots, (X_n, \varepsilon_n)$ in the space $S \times \{-1, 1\}$ and indexed by the class of functions $\tilde{\mathcal{F}} := \{\tilde{f} : f \in \mathcal{F}\}$, where $\tilde{f}(x, u) := f(x)u,\ (x, u) \in S \times \{-1, 1\}$.

3 Bounding Expected Sup-Norms of Empirical and Rademacher Processes

In what follows, we will use a number of bounds on expectation of suprema of empirical and Rademacher processes. Because of symmetrization inequalities, the problems of bounding expected suprema for these two stochastic processes are equivalent. The bounds are usually based on various complexity measures of function classes (such as linear dimension, VC-dimension, shattering numbers, uniform covering numbers, random covering numbers, bracketing numbers, etc). It would be of interest to develop the bounds with precise dependence on such geometric parameters as the $L_2(P)$-diameter of the class. Combining the bounds on expected suprema with Talagrand's concentration inequalities yields exponential inequalities for the tail probabilities of sup-norms.

3.1 Subgaussian Processes

Recall that a random variable Y is called *subgaussian* with parameter σ^2, or $Y \in SG(\sigma^2)$, iff for all $\lambda \in \mathbb{R}$

$$\mathbb{E}e^{\lambda Y} \le e^{\lambda^2\sigma^2/2}.$$

Normal random variable with mean 0 and variance σ^2 belongs to $SG(\sigma^2)$. If ε is Rademacher r.v., then $\varepsilon \in SG(1)$.

The next proposition gives two simple and important properties of subgaussian random variables (see, e.g., [95], Section 2.2.1 for the proof of property (ii)).

Proposition 3.1 *(i) If $Y_1, \dots, Y_n$ are independent random variables and $Y_j \in SG(\sigma_j^2)$, then*

$$Y_1 + \dots + Y_n \in SG(\sigma_1^2 + \dots + \sigma_n^2).$$

(ii) For arbitrary $Y_1, \dots, Y_N,\ N \geq 2$ such that $Y_j \in SG(\sigma_j^2), j = 1, \dots, N$,

$$\mathbb{E} \max_{1 \leq j \leq N} |Y_j| \leq C \max_{1 \leq j \leq N} \sigma_j \sqrt{\log N},$$

where C is a numerical constant.

Let (T, d) be a pseudo-metric space and $Y(t), t \in T$ be a stochastic process. It is called subgaussian with respect to d iff, for all $t, s \in T$, $Y(t) - Y(s) \in SG(d^2(t, s))$.

Denote $D(T) = D(T, d)$ the diameter of the space T. Let $N(T, d, \varepsilon)$ be the ε-covering number of (T, d), i.e., the minimal number of balls of radius ε needed to cover T. Let $M(T, d, \varepsilon)$ be the ε-packing number of (T, d), i.e., the largest number of points in T separated from each other by at least a distance of ε. Obviously,

$$N(T, d, \varepsilon) \leq M(T, d, \varepsilon) \leq N(T, d, \varepsilon/2),\ \ \varepsilon \geq 0.$$

As always,

$$H(T, d, \varepsilon) = \log N(T, d, \varepsilon)$$

is called the ε-entropy of (T, d).

Theorem 3.1 (Dudley's entropy bounds). *If $Y(t), t \in T$ is a subgaussian process with respect to d, then the following bounds hold with some numerical constant $C > 0$:*

$$\mathbb{E} \sup_{t \in T} Y(t) \leq C \int_0^{D(T)} H^{1/2}(T, d, \varepsilon) d\varepsilon$$

and for all $t_0 \in T$

$$\mathbb{E} \sup_{t \in T} |Y(t) - Y(t_0)| \leq C \int_0^{D(T)} H^{1/2}(T, d, \varepsilon) d\varepsilon.$$

The proof is based on the well known *chaining method* (see, e.g., [68], Section 11.1) that also leads to more refined *generic chaining bounds* (see Talagrand [88]). For Gaussian processes, the following lower bound is also true (see [68], Section 3.3).

Theorem 3.2 (Sudakov's entropy bound). *If $Y(t), t \in T$ is a Gaussian process and*

$$d(t,s) := \mathbb{E}^{1/2}(X(t) - X(s))^2, \ t, s \in T,$$

then the following bound holds with some numerical constant $C > 0$:

$$\mathbb{E} \sup_{t \in T} Y(t) \geq C \sup_{\varepsilon > 0} \varepsilon H^{1/2}(T, d, \varepsilon).$$

In addition to Gaussian processes, Rademacher sums provide another important example of subgaussian processes.

Given $T \subset \mathbb{R}^n$, define

$$Y(t) := \sum_{i=1}^{n} \varepsilon_i t_i, \ t = (t_1, \ldots, t_n) \in T,$$

where $\{\varepsilon_i\}$ are i.i.d. Rademacher random variables. The stochastic process $Y(t), t \in T$ is called *the Rademacher sum* indexed by T. It is a subgaussian process with respect to the Euclidean distance in $\mathbb{R}^n$:

$$d(t,s) = \left(\sum_{i=1}^{n} (t_i - s_i)^2 \right)^{1/2}.$$

The following result by Talagrand is a version of Sudakov's type lower bound for Rademacher sums (see [68], Section 4.5).

Denote

$$R(T) := \mathbb{E}_\varepsilon \sup_{t \in T} \left| \sum_{i=1}^{n} \varepsilon_i t_i \right|.$$

Theorem 3.3 (Talagrand). *There exists a universal constant L such that*

$$R(T) \geq \frac{1}{L} \delta H^{1/2}(T, d, \delta) \tag{3.1}$$

whenever

$$R(T) \sup_{t \in T} \|t\|_{\ell_\infty} \leq \frac{\delta^2}{L}. \tag{3.2}$$

3.2 Finite Classes of Functions

Suppose $\mathcal{F}$ is a finite class of measurable functions uniformly bounded by a constant $U > 0$. Let $N := \mathrm{card}(\mathcal{F}) \geq 2$. Denote

$$\sigma^2 := \sup_{f \in \mathcal{F}} Pf^2.$$

Theorem 3.4 *There exist universal constants K_1, K_2 such that*

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq K_1 U \sqrt{\frac{\log N}{n}}.$$

and

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq K_2 \biggl[\sigma\sqrt{\frac{\log N}{n}} \bigvee U\frac{\log N}{n}.\biggr]$$

Proof. Conditionally on $X_1, \ldots, X_n$, the random variable

$$\sqrt{n} R_n(f) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j f(X_j),\ f \in \mathcal{F}$$

is subgaussian with parameter $\|f\|_{L_2(P_n)}$. Therefore, it follows from Proposition 3.1 that

$$\mathbb{E}_{\varepsilon}\|R_n\|_{\mathcal{F}} \leq K \sup_{f \in \mathcal{F}} \|f\|_{L_2(P_n)} \sqrt{\frac{\log N}{n}}.$$

The first bound now follows since

$$\sup_{f \in \mathcal{F}} \|f\|_{L_2(P_n)} \leq U.$$

To prove the second bound, denote

$$\mathcal{F}^2 := \{f^2 : f \in \mathcal{F}\}$$

and observe that

$$\sup_{f \in \mathcal{F}} \|f\|_{L_2(P_n)} \leq \sup_{f \in \mathcal{F}} \|f\|_{L_2(P)} + \sqrt{\|P_n - P\|_{\mathcal{F}^2}},$$

which implies

$$\mathbb{E} \sup_{f \in \mathcal{F}} \|f\|_{L_2(P_n)} \leq \sigma + \sqrt{\mathbb{E}\|P_n - P\|_{\mathcal{F}^2}}.$$

Using symmetrization and contraction inequalities, we get

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}^2} \leq 2\mathbb{E}\|R_n\|_{\mathcal{F}^2} \leq 8U\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

Hence,

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq K\mathbb{E}\sup_{f\in\mathcal{F}}\|f\|_{L_2(P_n)}\sqrt{\frac{\log N}{n}} \leq K\biggl(\sigma + \sqrt{8U\mathbb{E}\|R_n\|_{\mathcal{F}}}\biggr)\sqrt{\frac{\log N}{n}}.$$

The result now follows by bounding the solution with respect to $\mathbb{E}\|R_n\|_{\mathcal{F}}$ of the above inequality.

□

The result can be also deduced from the following theorem (it is enough to take $q = \log N$).

Theorem 3.5 *There exists a universal constants K such that for all $q \geq 2$*

$$\mathbb{E}^{1/q}\|R_n\|_{\mathcal{F}}^q \leq \mathbb{E}^{1/q}\|R_n\|_{\ell_q(\mathcal{F})}^q := \mathbb{E}^{1/q}\sum_{f\in\mathcal{F}}|R_n(f)|^q \leq$$

$$K\biggl[\sigma\frac{(q-1)^{1/2}N^{1/q}}{n^{1/2}}\bigvee U\frac{(q-1)N^{2/q}}{n}\biggr].$$

Proof. We will need the following simple property of Rademacher sums: for all $q \geq 2$,

$$\mathbb{E}^{1/q}\biggl|\sum_{i=1}^n \alpha_i\varepsilon_i\biggr|^q \leq (q-1)^{1/2}\biggl(\sum_{i=1}^n \alpha_i^2\biggr)^{1/2}$$

(see, e.g., de la Pena and Giné [32], p. 21). Using this inequality, we get

$$\mathbb{E}_\varepsilon\|R_n\|_{\mathcal{F}}^q \leq \sum_{f\in\mathcal{F}}\mathbb{E}_\varepsilon|R_n(f)|^q \leq (q-1)^{q/2}n^{-q/2}\sum_{f\in\mathcal{F}}\|f\|_{L_2(P_n)}^q \leq$$

$$(q-1)^{q/2}n^{-q/2}N\biggl(\sup_{f\in\mathcal{F}}P_nf^2\biggr)^{q/2} \leq (q-1)^{q/2}n^{-q/2}N\biggl(\sigma^2 + \|P_n - P\|_{\mathcal{F}^2}\biggr)^{q/2}.$$

This easily implies

$$\mathbb{E}^{1/q}\|R_n\|_{\mathcal{F}}^q \leq \mathbb{E}^{1/q}\sum_{f\in\mathcal{F}}|R_n(f)|^q \leq$$

$$(q-1)^{1/2}n^{-1/2}N^{1/q}2^{1/2-1/q}\biggl(\sigma + \mathbb{E}^{1/q}\|P_n - P\|_{\mathcal{F}^2}^{q/2}\biggr). \qquad (3.3)$$

It remains to use symmetrization and contraction inequalities to get

$$\mathbb{E}^{1/q}\|P_n - P\|_{\mathcal{F}^2}^{q/2} \leq 2U^{1/2}\mathbb{E}^{1/q}\|R_n\|_{\mathcal{F}}^{q/2} \leq 2U^{1/2}\sqrt{\mathbb{E}^{1/q}\|R_n\|_{\mathcal{F}}^q},$$

to substitute this bound into (3.3) and to solve the resulting inequality for $\mathbb{E}^{1/q}\|R_n\|_{\mathcal{F}}^q$ to complete the proof.

□

3.3 Shattering Numbers and VC-classes of sets

.

Let $\mathcal{C}$ be a class of subsets of S. Given a finite set $F \subset S$, denote

$$\Delta^{\mathcal{C}}(F) := \mathrm{card}\{\mathcal{C} \cap F\},$$

where

$$\mathcal{C} \cap F := \Big\{C \cap F : C \in \mathcal{C}\Big\}.$$

Clearly,

$$\Delta^{\mathcal{C}}(F) \leq 2^{\mathrm{card}(F)}.$$

If $\Delta^{\mathcal{C}}(F) = 2^{\mathrm{card}(F)}$, it is said that F is shattered by $\mathcal{C}$. The numbers $\Delta^{\mathcal{C}}(F)$ are called *the shattering numbers* of the class $\mathcal{C}$.

Define

$$m^{\mathcal{C}}(n) := \sup\Big\{\Delta^{\mathcal{C}}(F) : F \subset S, \mathrm{card}(F) \leq n\Big\}.$$

Clearly,

$$m^{\mathcal{C}}(n) \leq 2^n, \ n = 1, 2, \ldots$$

and if, for some n, $m^{\mathcal{C}}(n) < 2^n$, then $m^{\mathcal{C}}(k) < 2^k$ for all $k \geq n$.

Let

$$V(\mathcal{C}) := \min\{n \geq 1 : m^{\mathcal{C}}(n) < 2^n\}.$$

If $m^{\mathcal{C}}(n) = 2^n$ for all $n \geq 1$, set $V(\mathcal{C}) = \infty$. The number $V(\mathcal{C})$ is called the *Vapnik-Chervonenkis dimension (or the VC-dimension)* of class $\mathcal{C}$. If $V(\mathcal{C}) < +\infty$, then $\mathcal{C}$ is called the Vapnik-Chervonenkis class (or VC-class). It means that no set F of cardinality $n \geq V(\mathcal{C})$ is shattered by $\mathcal{C}$.

Denote

$$\binom{n}{\leq k} := \binom{n}{0} + \cdots + \binom{n}{k}.$$

The following lemma (proved independently in somewhat different forms by Sauer, Shelah, and also by Vapnik and Chervonenkis) is one of the main combinatorial facts related to VC-classes.

Theorem 3.6 (Sauer's Lemma). *Let $F \subset S$, $\mathrm{card}(F) = n$. If*

$$\Delta^{\mathcal{C}}(F) > \binom{n}{\leq k-1},$$

then there exists a subset $F' \subset F$, $\mathrm{card}(F') = k$ such that F' is shattered by $\mathcal{C}$.

The Sauer's Lemma immediately implies that, for a VC-class $\mathcal{C}$,

$$m^{\mathcal{C}}(n) \le \binom{n}{\le V(\mathcal{C})-1},$$

which can be further bounded by $\left(\frac{ne}{V(\mathcal{C})-1}\right)^{V(\mathcal{C})-1}$.

We will view P and P_n as functions defined on a class $\mathcal{C}$ of measurable sets $C \mapsto P(C), C \mapsto P_n(C)$ and the Rademacher process will be also indexed by sets:

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

For $Y : \mathcal{C} \mapsto \mathbb{R}$, we still write $\|Y\|_{\mathcal{C}} := \sup_{C\in\mathcal{C}}|Y(C)|$.

Denote $\mathcal{F} := \{I_C : C \in \mathcal{C}\}$.

Theorem 3.7 *There exists a numerical constant $K > 0$ such that*

$$\mathbb{E}\|P_n - P\|_{\mathcal{C}} \le K\mathbb{E}\sqrt{\frac{\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n}} \le K\sqrt{\frac{\mathbb{E}\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n}}.$$

The drawback of this result is that it does not take into account the "size" of the sets in class $\mathcal{C}$. A better bound is possible in the case when, for all $C \in \mathcal{C}$, $P(C)$ is small. We will derive such an inequality in which the size of $\mathbb{E}\|P_n - P\|_{\mathcal{C}}$ is controlled in terms of random shattering numbers $\Delta^{\mathcal{C}}(X_1,\dots,X_n)$ and of

$$\|P\|_{\mathcal{C}} = \sup_{C\in\mathcal{C}} P(C)$$

(and which implies the inequality of Theorem 3.7).

Theorem 3.8 *There exists a numerical constant $K > 0$ such that*

$$\mathbb{E}\|P_n - P\|_{\mathcal{C}} \le K\|P\|_{\mathcal{C}}^{1/2}\mathbb{E}\sqrt{\frac{\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n}} \bigvee K\frac{\mathbb{E}\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n} \le$$

$$K\|P\|_{\mathcal{C}}^{1/2}\sqrt{\frac{\mathbb{E}\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n}} \bigvee K\frac{\mathbb{E}\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n}.$$

Proof. Let

$$T := \Big\{(I_C(X_1),\dots,I_C(X_n)) : C \in \mathcal{C}\Big\}.$$

Clearly,

$$\mathrm{card}(T) = \Delta^{\mathcal{C}}(X_1,\dots,X_n)$$

and

$$\mathbb{E}_{\varepsilon}\|R_n\|_{\mathcal{C}} = \mathbb{E}_{\varepsilon}\sup_{t\in T}\left|n^{-1}\sum_{i=1}^{n}\varepsilon_i t_i\right|.$$

For all $t \in T$, $n^{-1}\sum_{i=1}^{n}\varepsilon_i t_i$ is a subgaussian random variable with parameter $n^{-1}\|t\|_{\ell_2}$. Therefore, by Proposition 3.1,

$$\mathbb{E}_{\varepsilon}\sup_{t\in T}\left|n^{-1}\sum_{i=1}^{n}\varepsilon_i t_i\right| \le Kn^{-1}\sup_{t\in T}\|t\|_{\ell_2}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)}.$$

Note that

$$n^{-1}\sup_{t\in T}\|t\|_{\ell_2} = n^{-1/2}(\sup_{C\in\mathcal{C}} P_n(C))^{1/2}.$$

Hence,

$$\mathbb{E}_{\varepsilon}\|R_n\|_{\mathcal{C}} \le Kn^{-1/2}\mathbb{E}\|P_n\|_{\mathcal{C}}^{1/2}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)} \le$$

$$Kn^{-1/2}\mathbb{E}\sqrt{\|P_n-P\|_{\mathcal{C}}+\|P\|_{\mathcal{C}}}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)} \le$$

$$Kn^{-1/2}\mathbb{E}\sqrt{\|P_n-P\|_{\mathcal{C}}}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)} + Kn^{-1/2}\sqrt{\|P\|_{\mathcal{C}}}\mathbb{E}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)}.$$

By symmetrization inequality,

$$\mathbb{E}\|P_n-P\|_{\mathcal{C}} \le 2K\sqrt{2}n^{-1/2}\mathbb{E}\sqrt{\|P_n-P\|_{\mathcal{C}}}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)}+$$

$$2K\sqrt{2}n^{-1/2}\sqrt{\|P\|_{\mathcal{C}}}\mathbb{E}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)} \le$$

$$2Kn^{-1/2}\sqrt{\mathbb{E}\|P_n-P\|_{\mathcal{C}}}\sqrt{\mathbb{E}\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)}+$$

$$2Kn^{-1/2}\sqrt{\|P\|_{\mathcal{C}}}\mathbb{E}\sqrt{\log \Delta^{\mathcal{C}}(X_1,\dots,X_n)},$$

where we also used Cauchy-Schwarz inequality. It remains to solve the resulting inequality with respect to $\mathbb{E}\|P_n-P\|_{\mathcal{C}}$ (or just to upper bound its solution) to get the result.

□

In the case of VC-classes,

$$\log \Delta^{\mathcal{C}}(X_1,\dots,X_n) \le \log m^{\mathcal{C}}(n) \le KV(\mathcal{C})\log n$$

with some numerical constant $K > 0$. Thus, Theorem 3.8 yields the bound

$$\mathbb{E}\|P_n-P\|_{\mathcal{C}} \le K\left(\|P\|_{\mathcal{C}}^{1/2}\sqrt{\frac{V(\mathcal{C})\log n}{n}}\bigvee\frac{V(\mathcal{C})\log n}{n}\right).$$

However, this bound is not sharp: the logarithmic factor involved in it can be eliminated. To this end, the following bound on the covering numbers of a VC-class $\mathcal{C}$ is needed. For an arbitrary probability measure Q on $(S, \mathcal{A})$, define the distance

$$d_Q(C_1, C_2) = Q(C_1 \triangle C_2),\ C_1, C_2 \in \mathcal{C}.$$

Theorem 3.9 *There exists a universal constant $K > 0$ such that for any VC-class $\mathcal{C} \subset \mathcal{A}$ and for all probability measures Q on $(S, \mathcal{A})$*

$$N(\mathcal{C}; d_Q; \varepsilon) \le KV(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\frac{1}{\varepsilon}\right)^{V(\mathcal{C})-1},\quad \varepsilon \in (0,1).$$

This result is due to Haussler and it is an improvement of an earlier bound by Dudley (the proof and precise references can be found, e.g., in van der Vaart and Wellner [95]).

By Theorem 3.9, we get

$$N(\mathcal{C}; d_{P_n}; \varepsilon) \le KV(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\frac{1}{\varepsilon}\right)^{V(\mathcal{C})-1},\quad \varepsilon \in (0,1).$$

Using this fact one can prove the following inequality:

$$\mathbb{E}\|P_n - P\|_{\mathcal{C}} \le K\left(\|P\|_{\mathcal{C}}^{1/2}\sqrt{\log\frac{K}{\|P\|_{\mathcal{C}}}}\sqrt{\frac{V(\mathcal{C})}{n}} \bigvee \frac{V(\mathcal{C})\log\frac{K}{\|P\|_{\mathcal{C}}}}{n}\right).$$

We are not giving its proof here. However, in the next section, we establish more general results for VC-type classes of functions (see (3.13)) that do imply the above bound.

3.4 Upper Entropy Bounds

Let $N(\mathcal{F}; L_2(P_n); \varepsilon)$ denote the minimal number of $L_2(P_n)$-balls of radius ε covering $\mathcal{F}$. Denote

$$\sigma_n^2 := \sup_{f\in\mathcal{F}} P_n f^2.$$

Theorem 3.10 *The following bound holds with a numerical constant $C > 0$:*

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \le \frac{C}{\sqrt{n}}\mathbb{E}\int_0^{2^{1/2}\sigma_n}\sqrt{\log N(\mathcal{F}; L_2(P_n); \varepsilon)}d\varepsilon.$$

Proof. Conditionally on $X_1, \dots, X_n$, the process

$$\sqrt{n}R_n(f) = \frac{1}{\sqrt{n}}\sum_{j=1}^n \varepsilon_j f(X_j),\ f \in \mathcal{F}$$

is subgaussian with respect to the distance of the space $L_2(P_n)$. Hence, it follows from Theorem 3.1 that

$$\mathbb{E}_\varepsilon \|R_n\|_{\mathcal{F}} \leq C n^{-1/2} \int_0^{2^{1/2}\sigma_n} \sqrt{\log N(\mathcal{F}; L_2(P_n); \varepsilon)} d\varepsilon. \tag{3.4}$$

Taking expectation of both sides, yields the result.

□

Following Giné and Koltchinskii [50], we will derive from Theorem 3.10 several bounds under more special conditions on the random entropy. Assume that the functions in $\mathcal{F}$ are uniformly bounded by a constant $U > 0$ and let $F \leq U$ denote a measurable envelope of $\mathcal{F}$, i.e.

$$|f(x)| \leq F(x), x \in S, f \in \mathcal{F}.$$

We will assume that σ^2 is a number such that

$$\sup_{f \in \mathcal{F}} P f^2 \leq \sigma^2 \leq \|F\|^2_{L_2(P)}$$

Most often, we will use

$$\sigma^2 = \sup_{f \in \mathcal{F}} P f^2.$$

Let $H : [0, \infty) \mapsto [0, \infty)$ be a regularly varying function of exponent $0 \leq \alpha < 2$, strictly increasing for $u \geq 1/2$ and such that $H(u) = 0$ for $0 \leq u < 1/2$.

Theorem 3.11 *If, for all $\varepsilon > 0, n \geq 1$ and $\omega \in \Omega$,*

$$\log N(\mathcal{F}, L_2(P_n), \varepsilon) \leq H\left(\frac{\|F\|_{L_2(P_n)}}{\varepsilon}\right), \tag{3.5}$$

then there exists a constant $C > 0$, that depends only on H, such that

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq C\left[\frac{\sigma}{\sqrt{n}}\sqrt{H\Big(\frac{2\|F\|_{L_2(P)}}{\sigma}\Big)} \bigvee \frac{U}{n} H\Big(\frac{2\|F\|_{L_2(P)}}{\sigma}\Big)\right]. \tag{3.6}$$

In particular, if, for some $C_1 > 0$,

$$n\sigma^2 \geq C_1 U^2 H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right),$$

then

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq \frac{C\sigma}{\sqrt{n}}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right)} \tag{3.7}$$

with a constant $C > 0$ that depends only on H and C_1.

Proof. Without loss of generality, assume that $U = 1$ (otherwise the result follows by a simple rescaling of the class $\mathcal{F}$). Given function H, we will use constants $C_H > 0$, $D_H > 0$, $A_H > 0$ for which

$$\sup_{v\geq 1} \frac{\int_v^\infty u^{-2}\sqrt{H(u)}du}{v^{-1}\sqrt{H(v)}} \bigvee 1 \leq C_H, \quad \int_1^\infty u^{-2}\sqrt{H(u)}\ du \leq D_H$$

$$\sup_{v\geq 2} \frac{\log D_H v/(4C_H\sqrt{H(v)})}{v^2} \bigvee 1 \leq A_H.$$

The bound of Theorem 3.10 implies that with some numerical constant $C > 0$ (the value of C might change from place to place)

$$\begin{aligned}
&\mathbb{E}\|R_n\|_{\mathcal{F}} \leq Cn^{-1/2}\mathbb{E}\int_0^{2^{1/2}\sigma_n} \sqrt{\log N(\mathcal{F}, L_2(P_n), \varepsilon)}d\varepsilon \\
&\leq 2^{1/2}Cn^{-1/2}\mathbb{E}\int_0^{\sigma_n} \sqrt{H\left(\frac{\|F\|_{L_2(P_n)}}{\varepsilon}\right)}d\varepsilon \\
&\leq 2^{1/2}Cn^{-1/2}\mathbb{E}\int_0^{\sigma_n} \sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\varepsilon}\right)}d\varepsilon\ I\Big(\|F\|_{L_2(P_n)} \leq 2\|F\|_{L_2(P)}\Big) + \\
&2^{1/2}Cn^{-1/2}\mathbb{E}\int_0^{\sigma_n} \sqrt{H\left(\frac{\|F\|_{L_2(P_n)}}{\varepsilon}\right)}d\varepsilon\ I\Big(\|F\|_{L_2(P_n)} > 2\|F\|_{L_2(P)}\Big). \qquad (3.8)
\end{aligned}$$

It is very easy to bound the second term in the sum. First note that

$$\int_0^{\sigma_n} \sqrt{H\Big(\frac{\|F\|_{L_2(P_n)}}{\varepsilon}\Big)}d\varepsilon \leq \|F\|_{L_2(P_n)} \int_0^1 \sqrt{H(1/u)}du \leq D_H\|F\|_{L_2(P_n)}.$$

Then use Hölder's inequality and Bernstein's inequality to get

$$\begin{aligned}
&n^{-1/2}\mathbb{E}\left[\int_0^{\sigma_n} \sqrt{H\left(\frac{\|F\|_{L_2(P_n)}}{\varepsilon}\right)}d\varepsilon I\left(\|F\|_{L_2(P_n)} > 2\|F\|_{L_2(P)}\right)\right] \leq \\
&D_H n^{-1/2}\|F\|_{L_2(P)} \exp\left\{-\frac{9}{8}n\|F\|_{L_2(P)}^2\right\} \leq \frac{D_H}{2n}. \qquad (3.9)
\end{aligned}$$

Bounding the first term is slightly more complicated. Recall the notation

$$\mathcal{F}^2 := \{f^2 : f \in \mathcal{F}\}.$$

Using symmetrization and contraction inequalities, we get

$$\mathbb{E}\sigma_n^2 \leq \sigma^2 + \mathbb{E}\|P_n - P\|_{\mathcal{F}^2} \leq \sigma^2 + 2\mathbb{E}\|R_n\|_{\mathcal{F}^2} \leq \sigma^2 + 8\mathbb{E}\|R_n\|_{\mathcal{F}} =: B^2. \qquad (3.10)$$

Since, for nonincreasing h, the function

$$u \mapsto \int_0^u h(t)dt$$

is concave, we have, by the properties of H, that

$$\begin{aligned}
&n^{-1/2}\mathbb{E}\int_0^{\sigma_n}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\varepsilon}\right)}d\varepsilon\ I(\|F\|_{L_2(P_n)} \le 2\|F\|_{L_2(P)}) \le \\
&n^{-1/2}\mathbb{E}\int_0^{\sigma_n\wedge 2\|F\|_{L_2(P)}}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\varepsilon}\right)}d\varepsilon \\
&\le n^{-1/2}\int_0^{(\mathbb{E}\sigma_n^2)^{1/2}\wedge 2\|F\|_{L_2(P)}}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\varepsilon}\right)}d\varepsilon \\
&\le n^{-1/2}\int_0^{B\wedge 2\|F\|_{L_2(P)}}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\varepsilon}\right)}d\varepsilon \\
&\le C_H n^{-1/2}B\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{B\wedge 2\|F\|_{L_2(P)}}\right)}. \qquad (3.11)
\end{aligned}$$

Taking into account that

$$\sup_{f\in\mathcal{F}} Pf^2 \le \sigma^2 \le \|F\|_{L_2(P)}^2,$$

we deduce from inequality (3.11)

$$n^{-1/2}\mathbb{E}\left[\int_0^{\sigma_n}\sqrt{H\left(\frac{\|F\|_{L_2(P_n)}}{\varepsilon}\right)}d\varepsilon I\left(\|F\|_{L_2(P_n)} \le 2\|F\|_{L_2(P)}\right)\right]$$

$$\le C_H n^{-1/2}\sigma\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right)}$$

$$+\sqrt{8}C_H n^{-1/2}\sqrt{\mathbb{E}\|R_n\|_{\mathcal{F}}}\left(\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right)}\bigwedge\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\sqrt{8\mathbb{E}\|R_n\|_{\mathcal{F}}}\wedge 2\|F\|_{L_2(P)}}\right)}\right).$$

We will use the last bound together with inequalities (3.8) and (3.9). Denote

$$E := \mathbb{E}\|R_n\|_{\mathcal{F}}.$$

Then, we have either

$$E \le CD_H n^{-1},$$

or

$$E \leq CC_H \frac{\sigma}{\sqrt{n}} \sqrt{\frac{H(2\|F\|_{L_2(P)})}{\sigma}}$$

or

$$E \leq CC_H^2 n^{-1} \left[H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right) \bigwedge \left(H\left(\frac{\|F\|_{L_2(P)}}{\sqrt{2E}}\right) \bigvee H(1) \right) \right].$$

To complete the proof, it is enough to solve the resulting inequalities for E, using the following simple fact: if

$$\Psi(v) := v/H(1/\sqrt{v}),\ 0 < v \leq 1,$$

then

$$\Psi^{-1}(u) \leq u(H(1/\sqrt{u}) \vee 1),\ 0 < u \leq 1/H(1).$$

□

The next bounds follow from Theorem 3.11 with $\sigma^2 := \sup_{f\in\mathcal{F}} Pf^2$. If for some $A > 0, V > 0$ and for all $\varepsilon > 0$,

$$N(\mathcal{F}; L_2(P_n); \varepsilon) \leq \left(\frac{A\|F\|_{L_2(P_n)}}{\varepsilon}\right)^V, \tag{3.12}$$

then with some universal constant $C > 0$ (for $\sigma^2 \geq \text{const}\ n^{-1}$)

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq C\left[\sqrt{\frac{V}{n}}\sigma\sqrt{\log\frac{A\|F\|_{L_2(P)}}{\sigma}} \bigvee \frac{VU}{n}\log\frac{A\|F\|_{L_2(P)}}{\sigma}\right]. \tag{3.13}$$

If for some $A > 0, \rho \in (0,1)$ and for all $\varepsilon > 0$,

$$\log N(\mathcal{F}; L_2(P_n); \varepsilon) \leq \left(\frac{A\|F\|_{L_2(P_n)}}{\varepsilon}\right)^{2\rho}, \tag{3.14}$$

then

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq C\left[\frac{A^\rho\|F\|^\rho_{L_2(P)}}{\sqrt{n}}\sigma^{1-\rho} \bigvee \frac{A^{2\rho/(\rho+1)}\|F\|^{2\rho/(\rho+1)}_{L_2(P)} U^{(1-\rho)/(1+\rho)}}{n^{1/(1+\rho)}}\right]. \tag{3.15}$$

The inequalities of this type can be found in Talagrand [85], Einmahl and Mason [43], Giné and Guillou [48], Mendelson [76], Giné, Koltchinskii and Wellner [49]. Theorem 3.11 is given in Giné and Koltchinskii [50] (in a slightly more precise form).

A function class $\mathcal{F}$ is called VC-subgraph iff

$$\Big\{\{(x,t) : 0 \leq f(x) \leq t\} \cup \{(x,t) : 0 \geq f(x) \geq t\} : f \in \mathcal{F}\Big\}$$

is a VC-class. For a VC-subgraph class $\mathcal{F}$ the following bound holds with some constants $A, V > 0$ and for all probability measures Q on $(S, \mathcal{A})$:

$$N(\mathcal{F}; L_2(Q); \varepsilon) \leq \left(\frac{A\|F\|_{L_2(Q)}}{\varepsilon}\right)^V, \varepsilon > 0 \tag{3.16}$$

(see, e.g., van der Vaart and Wellner [95], Theorem 2.6.7). Of course, this *uniform covering numbers* condition does imply (3.12) and, as a consequence, (3.13).

We will call the function classes satisfying (3.12) *VC-type classes.*

If $\mathcal{H}$ is VC-type, then its convex hull $\mathrm{conv}(\mathcal{H})$ satisfies (3.14) with $\rho := \frac{V}{V+2}$ (see van der Vaart and Wellner [95], Theorem 2.6.9). More precisely, the following result holds.

Theorem 3.12 *Let $\mathcal{H}$ be a class of measurable functions on $(S, \mathcal{A})$ with a measurable envelope F and let Q be a probability measure on $(S, \mathcal{A})$. Suppose that $F \in L_2(Q)$ and*

$$N(\mathcal{H}; L_2(Q); \varepsilon) \leq \left(\frac{A\|F\|_{L_2(Q)}}{\varepsilon}\right)^V, \quad \varepsilon \leq \|F\|_{L_2(Q)}.$$

Then

$$\log N(\mathrm{conv}(\mathcal{H}); L_2(Q); \varepsilon) \leq \left(\frac{B\|F\|_{L_2(Q)}}{\varepsilon}\right)^{2V/(V+2)}, \quad \varepsilon \leq \|F\|_{L_2(Q)}$$

for some constant B that depends on A and V.

So, one can use the bound (3.15) for $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$. Note that in this bound the envelope F of the class $\mathcal{H}$ itself should be used rather than an envelope of a subset $\mathcal{F}$ of its convex hull (which might be smaller than F).

A number of other bounds on expected suprema of empirical and Rademacher processes (in particular, in terms of so called bracketing numbers) can be found in van der Vaart and Wellner [95], Dudley [42].

3.5 Lower Entropy Bounds

In this section, lower bounds on $\mathbb{E}\|R_n\|_{\mathcal{F}}$ expressed in terms of entropy of the class $\mathcal{F}$ will be proved. Again, we follow the paper by Giné and Koltchinskii [50]. Assume, for simplicity, that the functions in $\mathcal{F}$ are uniformly bounded by 1. In what follows, the function H satisfies the conditions of Theorem 3.11. Denote $\sigma^2 = \sup_{f \in \mathcal{F}} Pf^2$.

Under the notations of Section 3.4, we introduce the following condition: with some constant $c > 0$

$$\log N(\mathcal{F}, L_2(P), \sigma/2) \geq cH\left(\frac{\|F\|_{L_2(P)}}{\sigma}\right). \tag{3.17}$$

Theorem 3.13 *Let $\mathcal{F}$ satisfy condition (3.5). There exist a universal constant $B > 0$ and a constant C_1 that depends only on H such that*

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \geq B\frac{\sigma}{\sqrt{n}}\sqrt{\log N(\mathcal{F}, L_2(P), \sigma/2)} \tag{3.18}$$

provided that

$$n\sigma^2 \geq C_1 U^2 H\left(\frac{6\|F\|_{L_2(P)}}{\sigma}\right). \tag{3.19}$$

Moreover, if in addition (3.17) holds, then, for some constants C_2 depending only on c and C_3 depending only on H, and for all n for which condition (3.19) holds,

$$C_2\frac{\sigma}{\sqrt{n}}\sqrt{H\left(\frac{\|F\|_{L_2(P)}}{\sigma}\right)} \leq \mathbb{E}\|R_n\|_{\mathcal{F}} \leq C_3\frac{\sigma}{\sqrt{n}}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right)}. \tag{3.20}$$

Proof. Without loss of generality, we can assume that $U = 1$. The general case would follow by a simple rescaling. First note that, under the assumptions of the theorem, inequality (3.7) holds, so, we have with some constant C depending only on H

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq C\frac{\sigma}{\sqrt{n}}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right)}.$$

This already proves the right hand side of inequality (3.20).

It follows from Theorem 3.3 that

$$\mathbb{E}_\varepsilon\|R_n\|_{\mathcal{F}} \geq \frac{1}{8L}\frac{\sigma}{\sqrt{n}}\sqrt{\log N(\mathcal{F}, L_2(P_n), \sigma/8)}, \tag{3.21}$$

as soon as

$$\mathbb{E}_\varepsilon\|R_n\|_{\mathcal{F}} \leq \frac{\sigma^2}{64L}. \tag{3.22}$$

To use this result, we will derive a lower bound on the right hand side of (3.21) and an upper bound on the left hand side of (3.22) that hold with a high probability. Let us bound first the right hand side of (3.21).

Let

$$M := M(\mathcal{F}, L_2(P), \sigma/2)$$

(recall that $M(\mathcal{F}, L_2(P), \sigma/2)$ denotes the $\sigma/2$-packing number of the class $\mathcal{F} \subset L_2(P)$). We apply the law of large numbers to M functions in a maximal $\sigma/2$-separated subset of $\mathcal{F}$ and also to the envelope F. It implies that, for all $\varepsilon > 0$, there exists n and ω such that

$$M(\mathcal{F}, L_2(P), \sigma/2) \leq M(\mathcal{F}, L_2(P_n(\omega)), (1-\varepsilon)\sigma/2) \leq N(\mathcal{F}, L_2(P_n(\omega)), (1-\varepsilon)\sigma/4)$$

and

$$\|F\|_{L_2(P_n(\omega))} \leq (1+\varepsilon)\|F\|_{L_2(P)}.$$

Take $\varepsilon = 1/5$. Then, by (3.5),

$$M(\mathcal{F}, L_2(P), \sigma/2) \leq \exp\left\{H\left(\frac{6\|F\|_{L_2(P)}}{\sigma}\right)\right\}. \tag{3.23}$$

Let $f_1, \ldots, f_M$ be a maximal subset of $\mathcal{F}$ such that

$$P(f_i - f_j)^2 \geq \sigma^2/4 \text{ for all } 1 \leq i \neq j \leq M.$$

In addition, we have

$$P(f_i - f_j)^4 \leq 4P(f_i - f_j)^2 \leq 16\sigma^2.$$

Bernstein's inequality implies that

$$\mathbb{P}\left\{\max_{1\leq i\neq j\leq M}\left(nP(f_i-f_j)^2 - \sum_{k=1}^{n}(f_i-f_j)^2(X_k)\right) > \frac{8}{3}t + \sqrt{32tn\sigma^2}\right\} \leq M^2 e^{-t}.$$

Let $t = \delta n\sigma^2$. Since $P(f_i - f_j)^2 \geq \sigma^2/4$ and (3.23) holds, we get

$$\mathbb{P}\left\{\min_{1\leq i\neq j\leq M}\frac{1}{n}\sum_{k=1}^{n}(f_i-f_j)^2(X_k) \leq \sigma^2\left(1/4 - 8\delta/3 - \sqrt{32\delta}\right)\right\}$$

$$\leq \exp\left\{2H\left(\frac{3\|F\|_{L_2(P)}}{\sigma}\right) - \delta n\sigma^2\right\}.$$

For $\delta = 1/(32\cdot 8^3)$, this yields

$$\mathbb{P}\left\{\min_{1\leq i\neq j\leq M} P_n(f_i-f_j)^2 \leq \frac{\sigma^2}{16}\right\} \leq \exp\left\{H\left(\frac{6\|F\|_{L_2(P)}}{\sigma}\right) - \frac{n\sigma^2}{32\cdot 8^3}\right\}. \tag{3.24}$$

Denote

$$E_1 := \left\{M(\mathcal{F}, L_2(P_n), \sigma/4) \geq M\right\}.$$

On this event,

$$N(\mathcal{F}, L_2(P_n), \sigma/8) \geq M(\mathcal{F}, L_2(P_n), \sigma/4) \geq M = M(\mathcal{F}, L_2(P), \sigma/2) \geq N(\mathcal{F}, L_2(P), \sigma/2)$$

and

$$\mathbb{P}(E_1) \geq 1 - \exp\left\{H\left(\frac{6\|F\|_{L_2(P)}}{\sigma}\right) - \frac{n\sigma^2}{32\cdot 8^3}\right\}. \tag{3.25}$$

Using symmetrization and contraction inequalities and conditions (3.19), we have

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}^2} \leq 2\mathbb{E}\|R_n\|_{\mathcal{F}^2} \leq 8\mathbb{E}\|R_n\|_{\mathcal{F}} \leq C\frac{\sigma}{\sqrt{n}}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\sigma}\right)} \leq 6\sigma^2 \qquad (3.26)$$

(with a proper choice of constant C_1 in (3.19)). Next, Bousquet's version of Talagrand's inequality yields the bound

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}^2} \geq 6\sigma^2 + \sigma\sqrt{\frac{26t}{n}} + \frac{t}{3n}\right\} \leq e^{-t}.$$

We take $t = 26n\sigma^2$. Then

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}^2} \geq 41\sigma^2\right\} \leq \exp\{-26n\sigma^2\}.$$

Denote

$$E_2 := \left\{\sigma_n^2 = \sup_{f\in\mathcal{F}} P_n f^2 < 42\sigma^2\right\}. \qquad (3.27)$$

Then

$$\mathbb{P}(E_2) > 1 - \exp\{-26n\sigma^2\}. \qquad (3.28)$$

Also, by Bernstein's inequality, the event

$$E_3 = \{\|F\|_{L_2(P_n)} \leq 2\|F\|_{L_2(P)}\} \qquad (3.29)$$

has probability

$$\mathbb{P}(E_3) \geq 1 - \exp\left\{-\frac{9}{4}n\|F\|_{L_2(P)}^2\right\}. \qquad (3.30)$$

On the event $E_2 \cap E_3$, (3.4) and (3.19) yields that with some constant C depending only on H (C might change its value from place to place):

$$\begin{aligned}
\mathbb{E}_\varepsilon\|R_n\|_{\mathcal{F}} &\leq \frac{C}{\sqrt{n}}\int_0^{\sqrt{2}\sigma_n}\sqrt{H\left(\frac{\|F\|_{L_2(P_n)}}{\varepsilon}\right)}d\varepsilon \\
&\leq \frac{C}{\sqrt{n}}\int_0^{\sqrt{84}\sigma}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\varepsilon}\right)}d\varepsilon \leq \frac{C}{\sqrt{n}}\int_0^{2\sigma}\sqrt{H\left(\frac{2\|F\|_{L_2(P)}}{\varepsilon}\right)}d\varepsilon \\
&\leq C\frac{\sigma}{\sqrt{n}}\sqrt{H\left(\frac{\|F\|_{L_2(P)}}{\sigma}\right)} < \frac{\sigma^2}{64L}
\end{aligned} \qquad (3.31)$$

(again, with a proper choice of constant C_1 in (3.19)). It follows from (3.21)-(3.31) that

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \geq \frac{1}{8L}\frac{\sigma}{\sqrt{n}}\sqrt{\log N(\mathcal{F}, L_2(P), \sigma/2)}\mathbb{P}(E_1 \cap E_2 \cap E_3) \qquad (3.32)$$

and that

$$\mathbb{P}(E_1 \cap E_2 \cap E_3) \geq$$
$$1 - \exp\Big\{H\Big(\frac{6\|F\|_{L_2(P)}}{\sigma}\Big) - \frac{n\sigma^2}{32 \cdot 8^3}\Big\} - \exp\{-26n\sigma^2\} - \exp\{-9n\sigma^2/4\}.$$

This last probability is larger than 1/2 by condition (3.19) with a proper value of C_1. Thus, (3.32) implies inequality (3.18). The left hand side of inequality (3.20) now follows from (3.18) and (3.17), completing the proof.

□

3.6 Function Classes in Hilbert Spaces

Suppose that L is a finite dimensional subspace of $L_2(P)$ with $\dim(L) = d$. Denote

$$\psi_L(x) := \frac{1}{\sqrt{d}} \sup_{f \in L, \|f\|_{L_2(P)} \leq 1} |f(x)|.$$

We will use the following L_p-version of Hoffmann-Jørgensen inequality: for all independent mean zero random variables $Y_j,\ j = 1, \dots, n$ with values in a Banach space B and with $\mathbb{E}\|Y_j\|^p < +\infty$ for some $p \geq 1$,

$$\mathbb{E}^{1/p}\Big\|\sum_{j=1}^n Y_j\Big\|^p \leq K_p\Big(\mathbb{E}\Big\|\sum_{j=1}^n Y_j\Big\| + \mathbb{E}^{1/p}\Big(\max_{1\leq i\leq n} \|Y_i\|\Big)^p\Big), \tag{3.33}$$

where K_p is a constant depending only on p (see Ledoux and Talagrand [68], Theorem 6.20).

Proposition 3.2 *Let*

$$\mathcal{F} := \{f \in L : \|f\|_{L_2(P)} \leq R\}.$$

Then

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq \mathbb{E}^{1/2}\|R_n\|_{\mathcal{F}}^2 = R\sqrt{\frac{d}{n}}.$$

Moreover, there exists a universal constant K such that whenever

$$\mathbb{E} \max_{1\leq i\leq n} \psi_L^2(X_i) \leq \frac{n}{K^2},$$

we have

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \geq \frac{1}{K} R\sqrt{\frac{d}{n}}.$$

Proof. Let $\phi_1, \dots, \phi_d$ be an orthonormal basis of L. Then

$$\|R_n\|_{\mathcal{F}} := \sup_{f \in L, \|f\|_{L_2(P)} \le R} |R_n(f)| = \sup\biggl\{\biggl|R_n\biggl(\sum_{j=1}^d \alpha_j \phi_j\biggr)\biggr| : \sum_{j=1}^d \alpha_j^2 \le R^2\biggr\} =$$

$$\sup\biggl\{\biggl|\sum_{j=1}^d \alpha_j R_n(\phi_j)\biggr| : \sum_{j=1}^d \alpha_j^2 \le R^2\biggr\} = R\biggl(\sum_{j=1}^d R_n^2(\phi_j)\biggr)^{1/2}.$$

Therefore,

$$\mathbb{E}\|R_n\|_{\mathcal{F}}^2 = R^2 \sum_{j=1}^d \mathbb{E} R_n^2(\phi_j),$$

and the first statement follows since

$$\mathbb{E} R_n^2(\phi_j) = \frac{P\phi_j^2}{n}\frac{1}{n},\ j = 1, \dots, n.$$

The proof of the second statement follows from the first statement and inequality (3.33), which immediately yields

$$R\sqrt{\frac{d}{n}} = \mathbb{E}^{1/2}\|R_n\|_{\mathcal{F}}^2 \le K_2\biggl(\mathbb{E}\|R_n\|_{\mathcal{F}} + R\sqrt{\frac{d}{n}}\frac{1}{\sqrt{n}}\mathbb{E}^{1/2} \max_{1 \le i \le n} \psi_L^2(X_i)\biggr),$$

and the result follows with $K = 2K_2$.

□

Let K be a symmetric nonnegatively definite square integrable kernel on $S \times S$ and let $\mathcal{H}_K$ be the corresponding *reproducing kernel Hilbert space (RKHS),* i.e., $\mathcal{H}_K$ is the completion of the linear span of functions $\{K(x, \cdot) : x \in S\}$ with respect to the following inner product:

$$\biggl\langle \sum_i \alpha_i K(x_i, \cdot), \sum_j \beta_j K(y_i, \cdot) \biggr\rangle_K = \sum_{i,j} \alpha_i \beta_j K(x_i, y_j).$$

Let

$$\mathcal{F} := \{f \in \mathcal{H}_K : \|f\|_K \le 1 \text{ and } \|f\|_{L_2(P)} \le r\}$$

Let A_K denote the linear integral operator from $L_2(P)$ into $L_2(P)$ with kernel K,

$$A_K f(x) = \int_S K(x, y) f(y) P(dy),$$

and let $\{\lambda_i\}$ denote its eigenvalues arranged in decreasing order and $\{\phi_i\}$ denote the corresponding $L_2(P)$-orthonormal eigenfunctions.

The following result is due to Mendelson [77].

Proposition 3.3 *There exist universal constants $C_1, C_2 > 0$ such that*

$$C_1\biggl(n^{-1}\sum_{j=1}^{\infty}(\lambda_j \wedge r^2)\biggr)^{1/2} \leq \mathbb{E}^{1/2}\|R_n\|_{\mathcal{F}}^2 \leq C_2\biggl(n^{-1}\sum_{j=1}^{\infty}(\lambda_j \wedge r^2)\biggr)^{1/2}.$$

In addition, there exists a universal constant C such that

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \geq \frac{1}{C}\biggl(n^{-1}\sum_{j=1}^{\infty}(\lambda_j \wedge r^2)\biggr)^{1/2} - \frac{\sqrt{\sup_{x\in S} K(x,x)}}{n}.$$

Proof. By the well known properties of RKHS,

$$\mathcal{F} = \biggl\{\sum_{k=1}^{\infty} c_k\phi_k : c = (c_1, c_2, \dots) \in \mathcal{E}_1 \cap \mathcal{E}_2\biggr\},$$

where

$$\mathcal{E}_1 := \biggl\{c : \sum_{k=1}^{\infty}\frac{c_k^2}{\lambda_k} \leq 1\biggr\} \text{ and } \mathcal{E}_2 := \biggl\{c : \sum_{k=1}^{\infty}\frac{c_k^2}{r^2} \leq 1\biggr\}.$$

In other words, the set $\mathcal{E}_1$ is the ellipsoid in ℓ_2 (with the center at the origin) with "half-axes" $\sqrt{\lambda_k}$ and $\mathcal{E}_2$ is the ellipsoid with "half-axes" r (a ball of radius r). Let

$$\mathcal{E} := \biggl\{c : \sum_{k=1}^{\infty}\frac{c_k^2}{\lambda_k \wedge r^2} \leq 1\biggr\}$$

denote the ellipsoid with "half-axes" $\sqrt{\lambda_k} \wedge r$. A straightforward argument shows that

$$\mathcal{E} \subset \mathcal{E}_1 \cap \mathcal{E}_2 \subset \sqrt{2}\mathcal{E}.$$

Hence,

$$\sup_{c\in\mathcal{E}}\biggl|R_n\biggl(\sum_{k=1}^{\infty} c_k\phi_k\biggr)\biggr| \leq \|R_n\|_{\mathcal{F}} \leq \sqrt{2}\sup_{c\in\mathcal{E}}\biggl|R_n\biggl(\sum_{k=1}^{\infty} c_k\phi_k\biggr)\biggr|.$$

Also, we have

$$\sup_{c\in\mathcal{E}}\biggl|R_n\biggl(\sum_{k=1}^{\infty} c_k\phi_k\biggr)\biggr|^2 = \sup_{c\in\mathcal{E}}\biggl|\sum_{k=1}^{\infty}\frac{c_k}{\sqrt{\lambda_k}\wedge r}\Bigl(\sqrt{\lambda_k}\wedge r\Bigr)R_n(\phi_k)\biggr|^2 = \sum_{k=1}^{\infty}\Bigl(\lambda_k \wedge r^2\Bigr)R_n^2(\phi_k).$$

Hence,

$$\mathbb{E}\sup_{c\in\mathcal{E}}\biggl|R_n\biggl(\sum_{k=1}^{\infty} c_k\phi_k\biggr)\biggr|^2 = \sum_{k=1}^{\infty}\Bigl(\lambda_k \wedge r^2\Bigr)\mathbb{E}R_n^2(\phi_k).$$

Since $P\phi_k^2 = 1$, $\mathbb{E}R_n^2(\phi_k) = \frac{1}{n}$, so, we get

$$\mathbb{E}\sup_{c\in\mathcal{E}}\left|R_n\left(\sum_{k=1}^{\infty} c_k\phi_k\right)\right|^2 = n^{-1}\sum_{k=1}^{\infty}(\lambda_k \wedge r^2),$$

and the first bound follows.

The proof of the second bound is based on the observation that

$$\sup_{f\in\mathcal{F}}|f(x)| \le \sqrt{\sup_{x\in S} K(x,x)}$$

and on the same application of Hoffmann-Jørgensen inequality as in the previous proposition.

□

A similar result with the identical proof holds for data-dependent Rademacher complexity $\mathbb{E}_\varepsilon\|R_n\|_{\mathcal{F}}$. In this case, let $\{\lambda_i^{(n)}\}$ be the eigenvalues (arranged in decreasing order) of the random matrix $\left(n^{-1}K(X_i, X_j)\right)_{i,j=1}^{n}$ (equivalently, of the integral operator from $L_2(P_n)$ into $L_2(P_n)$ with kernel K).

Proposition 3.4 *There exist universal constants $C_1, C_2 > 0$ such that*

$$C_1\left(n^{-1}\sum_{j=1}^{n}(\lambda_j^{(n)} \wedge r^2)\right)^{1/2} \le \mathbb{E}_\varepsilon^{1/2}\|R_n\|_{\mathcal{F}}^2 \le C_2\left(n^{-1}\sum_{j=1}^{n}(\lambda_j^{(n)} \wedge r^2)\right)^{1/2}.$$

In addition, there exists a universal constant C such that

$$\mathbb{E}_\varepsilon\|R_n\|_{\mathcal{F}} \ge \frac{1}{C}\left(n^{-1}\sum_{j=1}^{n}(\lambda_j^{(n)} \wedge r^2)\right)^{1/2} - \frac{\sqrt{\sup_{x\in S} K(x,x)}}{n}.$$

4 Excess Risk Bounds

In this section, we develop distribution dependent and data dependent upper bounds on the excess risk $\mathcal{E}_P(\hat{f}_n)$ of an empirical risk minimizer

$$\hat{f}_n := \mathrm{argmin}_{f\in\mathcal{F}} P_n f.$$

We will assume that such a minimizer exists (a simple modification of the results is possible if $\hat{f}_n$ is an approximate solution of (1.2)). Our approach to this problem has been already outlined in the Introduction and it is closely related to the recent work of Massart [73], Koltchinskii and Panchenko [60], Bartlett, Bousquet and Mendelson [7], Bousquet, Koltchinskii and Panchenko [23], Koltchinskii [59], Bartlett and Mendelson [9].

4.1 Distribution Dependent Bounds and Ratio Bounds for Excess Risk

To simplify the matter, assume that the functions in $\mathcal{F}$ take their values in $[0,1]$. Recall that the set

$$\mathcal{F}_P(\delta) := \Big\{ f \in \mathcal{F} : \mathcal{E}_P(f) \le \delta \Big\}$$

is called the δ-minimal set of the risk P. In particular, $\mathcal{F}_P(0)$ is its minimal set.

Define $\rho_P : L_2(P) \times L_2(P) \mapsto [0, +\infty)$ such that

$$\rho_P^2(f,g) \ge P(f-g)^2 - (P(f-g))^2, \ f, g \in L_2(P).$$

Usually, ρ_P is also a (pseudo)metric, such as

$$\rho_P^2(f,g) = P(f-g)^2 \text{ or } \rho_P^2(f,g) = P(f-g)^2 - (P(f-g))^2.$$

Under the notations of the Introduction,

$$D(\delta) := D_P(\mathcal{F};\delta) := \sup_{f,g \in \mathcal{F}(\delta)} \rho_P(f,g)$$

is the ρ_P-diameter of the δ-minimal set. Also, denote

$$\mathcal{F}'(\delta) := \Big\{ f - g : f, g \in \mathcal{F}(\delta) \Big\}$$

and

$$\phi_n(\delta) := \phi_n(\mathcal{F}; P; \delta) := \mathbb{E}\|P_n - P\|_{\mathcal{F}'(\delta)}.$$

Let $\{\delta_j\}_{j \ge 0}$ be a decreasing sequence of positive numbers with $\delta_0 = 1$ and let $\{t_j\}_{j\ge 0}$ be a sequence of positive numbers. For $\delta \in (\delta_{j+1}, \delta_j]$, define

$$U_n(\delta) := \phi_n(\delta_j) + \sqrt{2\frac{t_j}{n}(D^2(\delta_j) + 2\phi_n(\delta_j))} + \frac{t_j}{2n}. \tag{4.1}$$

Finally, denote

$$\delta_n(\mathcal{F}; P) := \sup\{\delta \in (0,1] : \delta \le U_n(\delta)\}.$$

It is easy to check that

$$\delta_n(\mathcal{F}, P) \le U_n(\delta_n(\mathcal{F}, P)).$$

Obviously, the definitions of U_n and $\delta_n(\mathcal{F}, P)$ depend on the choice of $\{\delta_j\}$ and $\{t_j\}$.

We start with the following simple inequality that provides a distribution dependent upper bound on the excess risk $\mathcal{E}_P(\hat{f}_n)$.

Theorem 4.1 *For all $\delta \geq \delta_n(\mathcal{F}; P)$,*

$$\mathbb{P}\{\mathcal{E}(\hat{f}_n) > \delta\} \leq \sum_{\delta_j \geq \delta} e^{-t_j}.$$

Proof. It is enough to assume that $\delta > \delta_n(\mathcal{F}; P)$ (otherwise, the result follows by continuity). Denote $\hat{\delta} := \mathcal{E}(\hat{f}_n)$. If $\hat{\delta} \geq \delta \geq \varepsilon > 0$ and $g \in \mathcal{F}(\varepsilon)$, we have

$$\hat{\delta} = P\hat{f}_n - \inf_{g \in \mathcal{F}} Pg \leq P(\hat{f}_n - g) + \varepsilon \leq P_n(\hat{f}_n - g) + (P - P_n)(f - g) + \varepsilon \leq \|P_n - P\|_{\mathcal{F}'(\hat{\delta})} + \varepsilon.$$

By letting $\varepsilon \to 0$, this gives $\hat{\delta} \leq \|P_n - P\|_{\mathcal{F}'(\hat{\delta})}$. Denote

$$E_{n,j} := \Big\{ \|P_n - P\|_{\mathcal{F}'(\delta_j)} \leq U_n(\delta_j) \Big\}.$$

It follows from Bousquet's version of Talagrand's inequality that $\mathbb{P}(E_{n,j}) \geq 1 - e^{-t_j}$. Let

$$E_n := \bigcap_{\delta_j \geq \delta} E_{n,j}.$$

Then

$$\mathbb{P}(E_n) \geq 1 - \sum_{\delta_j \geq \delta} e^{-t_j}.$$

On the event E_n, for all $\sigma \geq \delta$, $\|P_n - P\|_{\mathcal{F}'(\sigma)} \leq U_n(\sigma)$, which holds by the definition of $U_n(\delta)$ and monotonicity of the function $\delta \mapsto \|P_n - P\|_{\mathcal{F}'(\delta)}$. Thus, on the event $\{\hat{\delta} \geq \delta\} \bigcap E_n$, we have

$$\hat{\delta} \leq \|P_n - P\|_{\mathcal{F}'(\hat{\delta})} \leq U_n(\hat{\delta}),$$

which implies that $\delta \leq \hat{\delta} \leq \delta_n(\mathcal{F}; P)$, contradicting the assumption that $\delta > \delta_n(\mathcal{F}; P)$. Therefore, we must have $\{\hat{\delta} \geq \delta\} \subset E_n^c$, and the result follows.

□

We now turn to uniform bounds on the ratios of the excess empirical risk of a function $f \in \mathcal{F}$ to its true excess risk. The excess empirical risk is defined as

$$\hat{\mathcal{E}}_n(f) := \mathcal{E}_{P_n}(f).$$

Given $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$, denote

$$\psi^\flat(\delta) := \sup_{\sigma \geq \delta} \frac{\psi(\sigma)}{\sigma}$$

and

$$\psi^{\sharp}(\varepsilon) := \inf\biggl\{\delta > 0 : \psi^{\flat}(\delta) \le \varepsilon\biggr\}.$$

These transformations will be called the $\flat$-transform and the $\sharp$-transform of ψ, respectively.

It happens that, with a high probability, the quantity

$$\sup_{f\in\mathcal{F},\mathcal{E}(f)\ge\delta}\left|\frac{\hat{\mathcal{E}}_n(f)}{\mathcal{E}(f)} - 1\right|$$

can be bounded from above by the function $\delta \mapsto V_n(\delta) := U_n^{\flat}(\delta)$.

Theorem 4.2 *For all $\delta \ge \delta_n(\mathcal{F}; P)$,*

$$\mathbb{P}\biggl\{\sup_{f\in\mathcal{F},\mathcal{E}(f)\ge\delta}\left|\frac{\hat{\mathcal{E}}_n(f)}{\mathcal{E}(f)} - 1\right| > V_n(\delta)\biggr\} \le \sum_{\delta_j\ge\delta} e^{-t_j}.$$

Proof. Consider the event E_n defined in the proof of Theorem 4.1. For this event

$$\mathbb{P}(E_n) \ge 1 - \sum_{\delta_j\ge\delta} e^{-t_j},$$

so, it is enough to prove that the inequality

$$\sup_{f\in\mathcal{F},\mathcal{E}(f)\ge\delta}\left|\frac{\hat{\mathcal{E}}_n(f)}{\mathcal{E}(f)} - 1\right| \le V_n(\delta)$$

holds on the event E_n. To this end, note that on this event, by the proof of Theorem 4.1, $\hat{f}_n \in \mathcal{F}(\delta)$. For all $f \in \mathcal{F}$ such that $\sigma := \mathcal{E}(f) \ge \delta$, for arbitrary $\varepsilon \in (0, \delta)$ and $g \in \mathcal{F}(\varepsilon)$, the following bounds hold:

$$\sigma = \mathcal{E}(f) \le Pf - Pg + \varepsilon \le P_n f - P_n g + (P - P_n)(f - g) + \varepsilon \le$$

$$\hat{\mathcal{E}}_n(f) + \|P_n - P\|_{\mathcal{F}'(\sigma)} + \varepsilon \le \hat{\mathcal{E}}_n(f) + U_n(\sigma) + \varepsilon \le \hat{\mathcal{E}}_n(f) + V_n(\delta)\sigma + \varepsilon,$$

which means that on the event E_n the condition $\mathcal{E}(f) \ge \delta$ implies that

$$\hat{\mathcal{E}}_n(f) \ge \Bigl(1 - V_n(\delta)\Bigr)\mathcal{E}(f).$$

Similarly, on the E_n, the condition $\sigma := \mathcal{E}(f) \ge \delta$ implies that

$$\hat{\mathcal{E}}_n(f) = P_n f - P_n\hat{f}_n \le Pf - P\hat{f}_n + (P_n - P)(f - \hat{f}_n) \le$$

$$\leq \mathcal{E}(f) + U_n(\sigma) \leq \mathcal{E}(f) + V_n(\delta)\sigma = \Bigl(1 + V_n(\delta)\Bigr)\mathcal{E}(f),$$

and the result follows.

□

A convenient choice of sequence $\{\delta_j\}$ is $\delta_j := q^{-j},\ j \geq 0$ with some fixed $q > 1$. If $t_j = t > 0,\ j \geq 0$, the corresponding functions $U_n(\delta)$ and $V_n(\delta)$ will be denoted by $U_n(\delta; t)$ and $V_n(\delta; t)$, and $\delta_n(\mathcal{F}; P)$ will be denoted by $\delta_n(t)$.

The following corollary is obvious.

Corollary 4.1 *For all $t > 0$ and for all $\delta \geq \delta_n(t)$,*

$$\mathbb{P}\{\mathcal{E}(\hat{f}_n) \geq \delta\} \leq \biggl(\log_q \frac{q}{\delta}\biggr) e^{-t}$$

and

$$\mathbb{P}\biggl\{\sup_{f\in\mathcal{F},\mathcal{E}(f)\geq\delta}\biggl|\frac{\hat{\mathcal{E}}_n(f)}{\mathcal{E}(f)} - 1\biggr| > V_n(\delta; t)\biggr\} \leq \biggl(\log_q \frac{q}{\delta}\biggr) e^{-t}.$$

It follows from the definition of $\delta_n(t)$ that $\delta_n(t) \geq \frac{t}{n}$. Because of this, the probabilities in Corollary 4.1 can be bounded from above by $\log_q \frac{n}{t} \exp\{-t\}$ (which depends neither on the class $\mathcal{F}$, nor on P). Most often, the logarithmic factor in front of the exponent does not create a problem: in typical applications $\delta_n(t)$ is upper bounded by $\delta_n + \frac{t}{n}$, where δ_n is larger than $\frac{\log\log n}{n}$. Adding $\log\log n$ to t is enough to eliminate the impact of the logarithm. However, if $\delta_n = O(n^{-1})$, the presence of the logarithmic factor would result in a suboptimal bound. To tackle this difficulty, we will use a slightly different choice of $\{\delta_j\}$, $\{t_j\}$.

For $q > 1$ and $t > 0$, denote

$$V_n^t(\sigma) := 2q\biggl[\phi_n^\flat(\sigma) + \sqrt{(D^2)^\flat(\sigma)}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\biggr],\ \sigma > 0.$$

Let

$$\sigma_n^t := \sigma_n^t(\mathcal{F}; P) := \inf\{\sigma : V_n^t(\sigma) \leq 1\}.$$

Theorem 4.3 *For all $t > 0$*

$$\mathbb{P}\{\mathcal{E}(\hat{f}_n) > \sigma_n^t\} \leq C_q e^{-t}$$

and for all $\sigma \geq \sigma_n^t$

$$\mathbb{P}\biggl\{\sup_{f\in\mathcal{F},\mathcal{E}(f)\geq\sigma}\biggl|\frac{\hat{\mathcal{E}}_n(f)}{\mathcal{E}(f)} - 1\biggr| > V_n^t(\sigma)\biggr\} \leq C_q e^{-t},$$

where

$$C_q := \frac{q}{q-1} \vee e.$$

Proof. Let $\sigma > \sigma_n^t$. Take $\delta_j = q^{-j}, j \geq 0$ and $t_j := t\frac{\delta_j}{\sigma}$ for some $t > 0, \sigma > 0$. The function $U_n(\delta)$, the quantity $\delta_n(\mathcal{F}, P)$, etc, now correspond to this choice of the sequences $\{\delta_j\}$, $\{t_j\}$. Then, it is easy to verify that for all $\delta \geq \sigma$

$$\frac{U_n(\delta)}{\delta} \leq 2q\biggl[\sup_{\delta_j \geq \sigma} \frac{\phi_n(\delta_j)}{\delta_j} + \sup_{\delta_j \geq \sigma} \frac{D(\delta_j)}{\sqrt{\delta_j}}\sqrt{\frac{t\delta_j}{n\sigma\delta_j}} + \frac{t\delta_j}{n\sigma\delta_j}\biggr]$$
$$\leq 2q\biggl[\sup_{\delta \geq \sigma} \frac{\phi_n(\delta)}{\delta} + \sup_{\delta \geq \sigma} \frac{D(\delta)}{\sqrt{\delta}}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\biggr] =$$
$$2q\biggl[\phi_n^\flat(\sigma) + \sqrt{(D^2)^\flat(\sigma)}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\biggr] = V_n^t(\sigma). \tag{4.2}$$

Since $\sigma > \sigma_n^t$ and the function V_n^t is strictly decreasing, we have $V_n^t(\sigma) < 1$ and, for all $\delta > \sigma_n^t$,

$$U_n(\delta) \leq V_n^t(\sigma)\delta < \delta.$$

Therefore, $\sigma_n^t \geq \delta_n(\mathcal{F}; P)$. It follows from Theorem 4.1 that

$$\mathbb{P}\{\mathcal{E}(\hat{f}_n) \geq \sigma\} \leq \sum_{\delta_j \geq \sigma} e^{-t_j}.$$

The right hand side can be now bounded as follows:

$$\sum_{\delta_j \geq \sigma} e^{-t_j} = \sum_{\delta_j \geq \sigma} \exp\biggl\{-t\frac{\delta_j}{\sigma}\biggr\} \leq \sum_{j \geq 0} e^{-tq^j} =$$
$$e^{-t} + \frac{q}{q-1}\sum_{j=1}^{\infty} q^{-j}e^{-tq^j}(q^j - q^{j-1}) \leq e^{-t} + \frac{1}{q-1}\int_1^{\infty} e^{-tx}dx =$$
$$e^{-t} + \frac{1}{q-1}\frac{1}{t}e^{-t} \leq \frac{q}{q-1}e^{-t},\ t \geq 1. \tag{4.3}$$

This implies the first bound for $t \geq 1$ and it is trivial for $t \leq 1$ because of the definition of constant C_q.

To prove the second bound use Theorem 4.2 and note that, by (4.2), $V_n(\sigma) \leq V_n^t(\sigma)$. The result follows from Theorem 4.2 and (4.3).

□

The result of Lemma 4.1 below is due to Massart [73, 74] (we formulate it in a slightly different form). Suppose that $\mathcal{F}$ is a class of measurable functions from S into

$[0,1]$ and $f_* : S \mapsto [0,1]$ is a measurable function such that with some numerical constant $D > 0$

$$D(Pf - Pf_*) \geq \rho_P^2(f, f_*) \geq P(f - f_*)^2 - (P(f - f_*))^2, \tag{4.4}$$

where ρ_P is a (pseudo)metric. The assumptions of this type are frequently used in model selection problems (see Section 6.3). They describe the link between the excess risk (or the approximation error) $Pf - P_*$ and the variance of the "loss" $f - f_*$. This particular form of bound (4.4) is typical in regression problems with L_2-loss (see Section 5.1): the link function in this case is just the square. In some other problems, such as classification under "low noise" assumption other link functions are also used (see Section 5.3).

Assume, for simplicity, that the infimum of Pf over $\mathcal{F}$ is attained at a function $\bar{f} \in \mathcal{F}$ (the result can be easily modified if this is not the case). Let

$$\omega_n(\delta) := \omega_n(\mathcal{F}; \bar{f}; \delta) := \mathbb{E} \sup_{f \in \mathcal{F}, \rho_P^2(f, \bar{f}) \leq \delta} |(P_n - P)(f - \bar{f})|.$$

Lemma 4.1 *There exists a constant $K > 0$ such that for all $\varepsilon \in (0,1]$ and for all $t > 0$*

$$\sigma_n^t(\mathcal{F}; P) \leq \varepsilon(\inf_{\mathcal{F}} Pf - Pf_*) + \frac{1}{D}\omega_n^\sharp\left(\frac{\varepsilon}{KD}\right) + \frac{KD}{\varepsilon}\frac{t}{n}.$$

Proof. Note that

$$\phi_n(\delta) = \mathbb{E}\|P_n - P\|_{\mathcal{F}'(\delta)} \leq 2\mathbb{E} \sup_{f \in \mathcal{F}(\delta)} |(P_n - P)(f - \bar{f})|.$$

For $f \in \mathcal{F}(\delta)$,

$$\rho_P(f, \bar{f}) \leq \rho_P(f, f_*) + \rho_P(\bar{f}, f_*) \leq \sqrt{D(Pf - Pf_*)} + \sqrt{D(P\bar{f} - Pf_*)} \leq$$

$$\leq \sqrt{D(Pf - P\bar{f})} + 2\sqrt{D(P\bar{f} - Pf_*)} \leq \sqrt{D\delta} + 2\sqrt{D\Delta} \leq \sqrt{2D(\delta + 4\Delta)},$$

where

$$\Delta := P\bar{f} - Pf_* = \inf_{\mathcal{F}} Pf - Pf_*.$$

As a result, it follows that

$$D(\delta) \leq 2\sqrt{D}(\sqrt{\delta} + 2\sqrt{\Delta}) \leq \sqrt{8D(\delta + 4\Delta)}$$

and

$$\phi_n(\delta) \leq 2\omega_n\Big(2D(\delta + 4\Delta)\Big).$$

We will now bound the functions $\phi_n^\flat(\sigma)$ and $(D^2)^\flat(\sigma)$ involved in the definition of $V_n^t(\sigma)$ (see the proof of Theorem 4.3). Denote $\tau := \frac{\Delta}{\sigma}$. Then

$$\phi_n^\flat(\sigma) = \sup_{\delta \geq \sigma} \frac{\phi_n(\delta)}{\delta} \leq 2 \sup_{\delta \geq \sigma} \frac{\omega_n\Big(2D(1+4\tau)\delta\Big)}{\delta} = 4D(1+4\tau)\omega_n^\flat\Big(2D(1+4\tau)\sigma\Big)$$

and also

$$(D^2)^\flat(\sigma) = \sup_{\delta \geq \sigma} \frac{D^2(\delta)}{\delta} \leq \sup_{\delta \geq \sigma} \frac{8D(\delta+4\Delta)}{\delta} \leq 8D(1+4\tau).$$

Therefore,

$$V_n^t(\sigma) \leq 2q\Bigg[4D(1+4\tau)\omega_n^\flat\Big(2D(1+4\tau)\sigma\Big) + 2\sqrt{2D}\sqrt{1+4\tau}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\Bigg].$$

Suppose that, for some $\varepsilon \in (0,1]$, we have $\sigma \geq \varepsilon\Delta$ implying that $\tau \leq \frac{1}{\varepsilon}$. Then we can upper bound $V_n^t(\sigma)$ as follows:

$$V_n^t(\sigma) \leq 2q\Bigg[\frac{20D}{\varepsilon}\omega_n^\flat\Big(2D\sigma\Big) + 2\sqrt{10}\sqrt{\frac{tD}{n\varepsilon\sigma}} + \frac{t}{n\sigma}\Bigg].$$

As soon as

$$\sigma \geq \frac{1}{2D}\omega_n^\sharp\Big(\frac{\varepsilon}{KD}\Big) \bigvee \frac{KDt}{n\varepsilon}$$

with a sufficiently large K, the right hand side of the last bound can be made smaller than 1. Thus, σ_n^t is upper bounded either by $\varepsilon\Delta$, or by the expression

$$\frac{1}{2D}\omega_n^\sharp\Big(\frac{\varepsilon}{KD}\Big) \bigvee \frac{KDt}{n\varepsilon},$$

which implies the bound of the lemma.

□

Remark. By increasing the value of the constant K it is easy to upper bound the quantity $\sup\{\sigma : V_n^t(\sigma) \leq 1/2\}$ in exactly the same way.

The next statement follows immediately from Lemma 4.1 and Theorem 4.3.

Proposition 4.1 *There exists a large enough constant $K > 0$ such that for all $\varepsilon \in (0,1]$ and all $t > 0$*

$$\mathbb{P}\Big\{P\hat{f} - Pf_* \geq (1+\varepsilon)(\inf_{\mathcal{F}} Pf - Pf_*) + \frac{1}{D}\omega_n^\sharp\Big(\frac{\varepsilon}{KD}\Big) + \frac{KD}{\varepsilon}\frac{t}{n}\Big\} \leq C_q e^{-t}.$$

Let us call $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ a function of concave type if it is nondecreasing and $u \mapsto \frac{\psi(u)}{u}$ is decreasing. If, in addition, for some $\gamma \in (0,1)$, $u \mapsto \frac{\psi(u)}{u^\gamma}$ is decreasing, ψ will be called a function of strictly concave type (with exponent γ). In particular, if $\psi(u) := \varphi(u^\gamma)$, or $\psi(u) := \varphi^\gamma(u)$, where φ is a nondecreasing strictly concave function with $\varphi(0) = 0$, then ψ is of concave type for $\gamma = 1$ and of strictly concave type for $\gamma < 1$.

Proposition 4.2 *Let $\delta_j := q^{-j}, j \geq 0$ for some $q > 1$. If ψ is a function of strictly concave type with some exponent $\gamma \in (0,1)$, then*

$$\sum_{j:\delta_j \geq \delta} \frac{\psi(\delta_j)}{\delta_j} \leq c_{\gamma,q} \frac{\psi(\delta)}{\delta},$$

where $c_{\gamma,q}$ is a constant depending only on q, γ.

Proof. Note that

$$\sum_{j:\delta_j \geq \delta} \frac{\psi(\delta_j)}{\delta_j} = \sum_{j:\delta_j \geq \delta} \frac{\psi(\delta_j)}{\delta_j^\gamma \delta_j^{1-\gamma}} \leq \frac{\psi(\delta)}{\delta^\gamma} \sum_{j:\delta_j \geq \delta} \frac{1}{\delta_j^{1-\gamma}} =$$

$$= \frac{\psi(\delta)}{\delta} \sum_{j:\delta_j \geq \delta} \left(\frac{\delta}{\delta_j}\right)^{1-\gamma} \leq \frac{\psi(\delta)}{\delta} \sum_{j \geq 0} q^{-j(1-\gamma)} = c_{\gamma,q} \frac{\psi(\delta)}{\delta}.$$

□

Assume that $\phi_n(\delta) \leq \check{\phi}_n(\delta)$ and $D(\delta) \leq \check{D}(\delta),\ \delta > 0$, where $\check{\phi}_n$ is a function of strictly concave type with some exponent $\gamma \in (0,1)$ and $\check{D}$ is a concave type function. Define

$$\check{U}_n(\delta; t) := \check{U}_{n,t}(\delta) := \check{K}\left(\check{\phi}_n(\delta) + \check{D}(\delta)\sqrt{\frac{t}{n}} + \frac{t}{n}\right)$$

with some numerical constant $\check{K}$. Then $\check{U}_n(\cdot; t)$ is also a function of strictly concave type.. In this case, it is natural to define

$$\check{V}_n(\delta; t) := \check{U}^\flat_{n,t}(\delta) = \frac{\check{U}_n(\delta; t)}{\delta} \quad \text{and} \quad \check{\delta}_n(t) := \check{U}^\sharp_{n,t}(1).$$

Theorem 4.4 *There exists a constant $\check{K}$ in the definition of the function $\check{U}_n(\delta; t)$ such that for all $t > 0$*

$$\mathbb{P}\{\mathcal{E}(\hat{f}_n) \geq \check{\delta}_n(t)\} \leq e^{-t}$$

and for all $\delta \geq \check{\delta}_n(t)$,

$$\mathbb{P}\left\{\sup_{f \in \mathcal{F}, \mathcal{E}(f) \geq \delta} \left|\frac{\hat{\mathcal{E}}_n(f)}{\mathcal{E}(f)} - 1\right| \geq \check{V}_n(\delta; t)\right\} \leq e^{-t}.$$

Proof. It is similar to the proof of Theorem 4.2, but now our goal is to avoid using the concentration inequality repeatedly for each value of δ_j since this leads to a logarithmic factor. The trick was previously used in Massart [73] and in Ph.D. dissertation of Bousquet (see also Bartlett, Bousquet and Mendelson [7]). Define

$$\mathcal{G}_\delta := \bigcup_{\sigma\geq\delta} \frac{\delta}{\sigma}\Big\{f-g : f,g\in\mathcal{F}(\sigma)\Big\}.$$

Then the functions in $\mathcal{G}_\delta$ are bounded by 1 and

$$\sigma_P(\mathcal{G}_\delta) \leq \sup_{\sigma\geq\delta}\frac{\delta}{\sigma}\sup_{f,g\in\mathcal{F}(\sigma)}\sigma_P(f-g) \leq \delta\sup_{\sigma\geq\delta}\frac{\check{D}(\sigma)}{\sigma} \leq \check{D}(\delta),$$

since $\check{D}$ is of concave type. Also, since $\check{\phi}_n$ is of strictly concave type, Proposition 4.2 gives

$$\mathbb{E}\|P_n-P\|_{\mathcal{G}_\delta} = \mathbb{E}\sup_{j:\delta_j\geq\delta}\sup_{\sigma\in(\delta_{j+1},\delta_j]}\frac{\delta}{\sigma}\|P_n-P\|_{\mathcal{F}'(\sigma)} \leq$$

$$\leq q\sum_{j:\delta_j\geq\delta}\frac{\delta}{\delta_j}\mathbb{E}\|P_n-P\|_{\mathcal{F}'(\delta_j)} \leq q\delta\sum_{j:\delta_j\geq\delta}\frac{\check{\phi}_n(\delta_j)}{\delta_j} \leq qc_{\gamma,q}\check{\phi}_n(\delta).$$

Now Talagrand's concentration inequality implies that there exists an event E of probability $\mathbb{P}(E)\geq 1-e^{-t}$ such that on this event $\|P_n-P\|_{\mathcal{G}_\delta}\leq\check{U}_n(\delta;t)$ (the constant $\check{K}$ in the definition of $\check{U}_n(\delta;t)$ should be chosen properly). Then, on the event E, for all $\sigma\geq\delta$,

$$\|P_n-P\|_{\mathcal{F}'(\sigma)} \leq \frac{\sigma}{\delta}\check{U}_n(\delta;t) \leq \check{V}_n(\delta;t)\sigma.$$

The rest repeats the proof of theorems 4.2 and 4.1. □

In the next theorem, we consider empirical risk minimization problems over Donsker classes of functions under the assumption that, roughly speaking, the true risk has unique minimum and, as a consequence, the δ-minimal sets $\mathcal{F}(\delta)$ shrink to a set consisting of a single function as $\delta\to 0$. Essentially, it will be shown that in such cases the excess risk is of the order $o_{\mathbb{P}}(n^{-1/2})$.

Theorem 4.5 *If $\mathcal{F}$ is a P-Donsker class and*

$$D_P(\mathcal{F};\delta)\to 0 \text{ as } n\to\infty,$$

then

$$\mathcal{E}_P(\hat{f}_n) = o_{\mathbb{P}}(n^{-1/2}) \text{ as } n\to\infty.$$

Proof. If $\mathcal{F}$ is a P-Donsker class, then the sequence of empirical processes

$$Z_n(f) := n^{1/2}(P_n f - Pf), f \in \mathcal{F}$$

is asymptotically equicontinuous, i.e., for all $\varepsilon > 0$

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \mathbb{P}\Bigl\{\sup_{\rho_P(f,g)\le\delta, f,g\in\mathcal{F}} \Bigl|Z_n(f) - Z_n(g)\Bigr| \ge \varepsilon\Bigr\} = 0.$$

(see, e.g., van der Vaart and Wellner [95], Section 2.1.2). This also implies (in the case of uniformly bounded classes, by an application of Talagrand's concentration inequality) that

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \mathbb{E} \sup_{\rho_P(f,g)\le\delta, f,g\in\mathcal{F}} \Bigl|Z_n(f) - Z_n(g)\Bigr| = 0.$$

Since $D_P(\mathcal{F};\delta) \to 0$ as $\delta \to 0$, it follows that

$$\lim_{\delta\to 0}\limsup_{n\to\infty} n^{1/2}\phi_n(\mathcal{F};P;\delta) = \lim_{\delta\to 0}\limsup_{n\to\infty} n^{1/2}\mathbb{E}\|P_n - P\|_{\mathcal{F}'(\delta)} \le$$
$$\lim_{\delta\to 0}\limsup_{n\to\infty} \mathbb{E} \sup_{\rho_P(f,g)\le D(\mathcal{F};\delta), f,g\in\mathcal{F}} \Bigl|Z_n(f) - Z_n(g)\Bigr| = 0. \tag{4.5}$$

Let now $\{\delta_j\}$ be a decreasing sequence such that $D(\delta_j) \le e^{-(j+1)}$ and let

$$t_j := t + 2\log\log\frac{1}{D(\delta_j)} \le t + 2\log(j+1).$$

Let δ_n^t denote the corresponding quantity $\delta_n(\mathcal{F};P)$ and U_n^t the corresponding function U_n (as they were defined before Theorem 4.1). Then it follows from Theorem 4.1 that

$$\mathbb{P}\{\mathcal{E}_P(\hat{f}_n) > \delta_n^t\} \le \sum_{\delta_j\ge\delta_n^t} e^{-t_j} \le \sum_{j\ge 0} e^{-t_j} \le \sum_{j\ge 0} e^{-t-2\log(j+1)} = \sum_{j\ge 1} j^{-2}e^{-t} \le 2e^{-t}.$$

The definition of U_n^t and the relationship (4.5) imply that, for all $t > 0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty} n^{1/2}U_n^t(\delta) = 0.$$

We also have

$$U_n^t(1) = O(n^{-1/2})$$

since

$$\phi_n(1) \le 2\mathbb{E}\|P_n - P\|_{\mathcal{F}} = O(n^{-1/2})$$

for a Donsker class $\mathcal{F}$ and $D(1) < +\infty$. Therefore, the definition of δ_n^t implies that

$$\delta_n^t \le U_n^t(\delta_n^t) \le U_n^t(1) \to 0 \text{ as } n \to \infty$$

and

$$\limsup_{n\to\infty} n^{1/2}\delta_n^t \leq \limsup_{n\to\infty} n^{1/2}U_n^t(\delta_n^t) = \lim_{\delta\to 0}\limsup_{n\to\infty} n^{1/2}U_n^t(\delta) = 0.$$

As a result, for all $t > 0$,

$$\delta_n^t = o(n^{-1/2}).$$

Then, it is easy to show that there is a choice of $t = \tau_n \to \infty$ (slowly enough) such that

$$\delta_n^{\tau_n} = o(n^{-1/2}).$$

The claim of the theorem now follows from the bound

$$\mathbb{P}\{\mathcal{E}_P(\hat{f}_n) > \delta_n^{\tau_n}\} \leq 2e^{-\tau_n} \to 0 \text{ as } n \to \infty.$$

□

There is another version of the proof that is based on Theorem 4.3.

The condition $D(\mathcal{F};\delta) \to 0$ as $\delta \to 0$ is quite natural when the true risk minimization problem (1.1) has unique solution. In this case, such quantities as $\delta_n(\mathcal{F}; P)$ often give correct (in a minimax sense) convergence rate for the excess risk in risk minimization problems. However, if the minimum in (1.1) is not unique, the diameter $D(\delta)$ of the δ-minimal set is bounded away from 0. In such cases, $\delta_n(\mathcal{F}; P)$ is bounded from below by $c\sqrt{\frac{1}{n}}$. At the same time, the optimal convergence rate of the excess risk to 0 is often better than this (in fact, it can be close to n^{-1}, e.g., in classification problems).

4.2 Rademacher Complexities and Data Dependent Bounds on Excess Risk

In a variety of statistical problems, it is crucial to have data dependent upper and lower confidence bounds on the sup-norm of the empirical process $\|P_n - P\|_{\mathcal{F}}$ for a given function class $\mathcal{F}$. This random variable is a natural measure of the accuracy of approximation of unknown distribution P by its empirical distribution P_n. However, $\|P_n - P\|_{\mathcal{F}}$ depends on the unknown distribution P and, hence, it can not be used directly. It happens that it is easy to construct rather simple upper and lower bounds on $\|P_n - P\|_{\mathcal{F}}$ in terms of the sup-norm of Rademacher process $\|R_n\|_{\mathcal{F}}$. The last random variable depends only on the data $X_1, \dots, X_n$ and on random signs $\varepsilon_1, \dots, \varepsilon_n$ that are independent of $X_1, \dots, X_n$ and are easy to simulate. Thus, $\|R_n\|_{\mathcal{F}}$ can be used as a data dependent complexity measure of the class $\mathcal{F}$ that allows one to estimate the accuracy of approximation of P by P_n based on the data. This bootstrap type approach was introduced independently in Koltchinskii [58] and Bartlett, Boucheron and Lugosi [8] and it was used to develop a general method of model selection and complexity regularization in learning theory. It is based on the following simple bounds. Their proof is very elementary and relies only on the symmetrization and bounded difference inequalities.

Assume that the functions in the class $\mathcal{F}$ are uniformly bounded by a constant $U > 0$.

Theorem 4.6 *For all* $t > 0$,

$$\mathbb{P}\biggl\{\|P_n - P\|_{\mathcal{F}} \geq 2\|R_n\|_{\mathcal{F}} + \frac{3tU}{\sqrt{n}}\biggr\} \leq \exp\biggl\{-\frac{t^2}{2}\biggr\}$$

and

$$\mathbb{P}\biggl\{\|P_n - P\|_{\mathcal{F}} \leq \frac{1}{2}\|R_n\|_{\mathcal{F}} - \frac{2tU}{\sqrt{n}} - \frac{U}{2\sqrt{n}}\biggr\} \leq \exp\biggl\{-\frac{t^2}{2}\biggr\}.$$

Proof. Denote

$$Z_n := \|P_n - P\|_{\mathcal{F}} - 2\|R_n\|_{\mathcal{F}}.$$

Then, by symmetrization inequality, $\mathbb{E}Z_n \leq 0$ and applying bounded difference inequality to random variable Z_n easily yields

$$Z_n \geq \mathbb{E}Z_n + \frac{3tU}{\sqrt{n}} \leq \exp\biggl\{-\frac{t^2}{2}\biggr\},$$

which implies the first bound.

The second bound is proved similarly by considering the random variable

$$Z_n := \|P_n - P\|_{\mathcal{F}} - \frac{1}{2}\|R_n\|_{\mathcal{F}} - \frac{U}{2\sqrt{n}}$$

and using symmetrization and bounded difference inequalities.

□

Note that other versions of bootstrap, most notably, the classical Efron's bootstrap, can be also used in a similar way (see Fromont [44]).

The major drawback of Theorem 4.6 is that the error term does not take into account the size of the variance of functions in the class $\mathcal{F}$. In some sense, this is a data dependent version of uniform Hoeffding inequality and what is often needed is a data dependent version of uniform Bernstein type inequality. We provide such a result below. It can be viewed as a **statistical version** of Talagrand's concentration inequality. Recently, Giné and Nickl [51] used some inequalities of similar nature in adaptive density estimation.

Denote

$$\sigma_P^2(\mathcal{F}) := \sup_{f\in\mathcal{F}} Pf^2 \text{ and } \sigma_n^2(\mathcal{F}) := \sup_{f\in\mathcal{F}} P_n f^2.$$

Theorem 4.7 *There exists a numerical constant $K > 0$ such that for all $t \geq 1$ with probability at least $1 - e^{-t}$ the following bounds hold:*

$$\Bigl| \|R_n\|_{\mathcal{F}} - \mathbb{E}\|R_n\|_{\mathcal{F}} \Bigr| \leq K\biggl[\sqrt{\frac{t}{n}\Bigl(\sigma_n^2(\mathcal{F}) + U\|R_n\|_{\mathcal{F}}\Bigr)} + \frac{tU}{n}\biggr], \tag{4.6}$$

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq K\biggl[\|R_n\|_{\mathcal{F}} + \sigma_n(\mathcal{F})\sqrt{\frac{t}{n}} + \frac{tU}{n}\biggr], \tag{4.7}$$

$$\sigma_P^2(\mathcal{F}) \leq K\biggl(\sigma_n^2(\mathcal{F}) + U\|R_n\|_{\mathcal{F}} + \frac{tU}{n}\biggr) \tag{4.8}$$

and

$$\sigma_n^2(\mathcal{F}) \leq K\biggl(\sigma_P^2(\mathcal{F}) + U\mathbb{E}\|R_n\|_{\mathcal{F}} + \frac{tU}{n}\biggr). \tag{4.9}$$

Also, for all $t \geq 1$ with probability at least $1 - e^{-t}$

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} \leq K\biggl[\|R_n\|_{\mathcal{F}} + \sigma_n(\mathcal{F})\sqrt{\frac{t}{n}} + \frac{tU}{n}\biggr] \tag{4.10}$$

and

$$\Bigl| \|P_n - P\|_{\mathcal{F}} - \mathbb{E}\|P_n - P\|_{\mathcal{F}} \Bigr| \leq K\biggl[\sqrt{\frac{t}{n}\Bigl(\sigma_n^2(\mathcal{F}) + U\|R_n\|_{\mathcal{F}}\Bigr)} + \frac{tU}{n}\biggr]. \tag{4.11}$$

Proof. It is enough to consider the case when $U = 1/2$. The general case then follows by rescaling. Using Talagrand's concentration inequality (to be specific, Klein-Rio bound), we claim that on an event E of probability at least $1 - e^{-t}$

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq \|R_n\|_{\mathcal{F}} + \sqrt{\frac{2t}{n}\Bigl(\sigma_P^2(\mathcal{F}) + 2\mathbb{E}\|R_n\|_{\mathcal{F}}\Bigr)} + \frac{t}{n}, \tag{4.12}$$

which implies that

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq \|R_n\|_{\mathcal{F}} + \sigma_P(\mathcal{F})\sqrt{\frac{2t}{n}} + \frac{t}{n} + 2\sqrt{\frac{1}{2}\mathbb{E}\|R_n\|_{\mathcal{F}}\frac{2t}{n}} \leq$$

$$\leq \|R_n\|_{\mathcal{F}} + \sigma_P(\mathcal{F})\sqrt{\frac{2t}{n}} + \frac{t}{n} + \frac{1}{2}\mathbb{E}\|R_n\|_{\mathcal{F}} + \frac{2t}{n},$$

or

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq 2\|R_n\|_{\mathcal{F}} + 2\sqrt{2}\sigma_P(\mathcal{F})\sqrt{\frac{t}{n}} + \frac{6t}{n}. \tag{4.13}$$

We will now upper bound $\sigma_P^2(\mathcal{F})$ in terms of $\sigma_n^2(\mathcal{F})$. Denote $\mathcal{F}^2 := \{f^2 : f \in \mathcal{F}\}$. Again, we apply Talagrand's concentration inequality (namely, Bousquet's bound) and show that on an event F of probability at least $1 - e^{-t}$

$$\sigma_P^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} Pf^2 \le \sup_{f\in\mathcal{F}} P_n f^2 + \|P_n - P\|_{\mathcal{F}^2} \le$$

$$\le \sigma_n^2(\mathcal{F}) + \mathbb{E}\|P_n - P\|_{\mathcal{F}^2} + \sqrt{\frac{2t}{n}\Big(\sigma_P^2(\mathcal{F}) + 2\mathbb{E}\|P_n - P\|_{\mathcal{F}^2}\Big)} + \frac{t}{3n},$$

where we also used the fact that

$$\sup_{f\in\mathcal{F}^2} \mathrm{Var}_P(f^2) \le \sup_{f\in\mathcal{F}} Pf^2 = \sigma_P^4(\mathcal{F}) < \sup_{f\in\mathcal{F}} Pf^2 = \sigma_P^2(\mathcal{F})$$

since the functions from $\mathcal{F}$ are uniformly bounded by $U = 1/2$. Using symmetrization inequality and then contraction inequality for Rademacher processes, we get

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}^2} \le 2\mathbb{E}\|R_n\|_{\mathcal{F}^2} \le 8\mathbb{E}\|R_n\|_{\mathcal{F}^2}.$$

Hence,

$$\sigma_P^2(\mathcal{F}) \le \sigma_n^2(\mathcal{F}) + 8\mathbb{E}\|R_n\|_{\mathcal{F}} + \sigma_P(\mathcal{F})\sqrt{\frac{2t}{n}} + 2\sqrt{\frac{8t}{n}\mathbb{E}\|R_n\|_{\mathcal{F}}} + \frac{t}{3n} \le$$

$$\le \sigma_n^2(\mathcal{F}) + 9\mathbb{E}\|R_n\|_{\mathcal{F}} + \sigma_P(\mathcal{F})\sqrt{\frac{2t}{n}} + \frac{9t}{n},$$

where the inequality $2\sqrt{ab} \le a + b,\ a, b \ge 0$ was applied. Next we use bound (4.13) on $\mathbb{E}\|R_n\|_{\mathcal{F}}$ to get

$$\sigma_P^2(\mathcal{F}) \le \sigma_n^2(\mathcal{F}) + 18\|R_n\|_{\mathcal{F}} + 19\sigma_P(\mathcal{F})\sqrt{\frac{2t}{n}} + \frac{100t}{n}.$$

As before, we bound the term $19\sigma_P(\mathcal{F})\sqrt{\frac{2t}{n}} = 2 \times 19\frac{\sigma_P(\mathcal{F})}{\sqrt{2}}\sqrt{\frac{t}{n}}$ using the inequality $2ab \le a^2 + b^2$, which gives

$$\sigma_P^2(\mathcal{F}) \le \frac{1}{2}\sigma_P^2(\mathcal{F}) + \sigma_n^2(\mathcal{F}) + 18\|R_n\|_{\mathcal{F}} + \frac{500t}{n}.$$

As a result, the following bound holds on the event $E \cap F$:

$$\sigma_P^2(\mathcal{F}) \le 2\sigma_n^2(\mathcal{F}) + 36\|R_n\|_{\mathcal{F}} + \frac{1000t}{n}. \tag{4.14}$$

It also implies

$$\sigma_P(\mathcal{F}) \le \sqrt{2}\sigma_n(\mathcal{F}) + 6\sqrt{\|R_n\|_{\mathcal{F}}} + 32\sqrt{\frac{t}{n}}.$$

We use this bound on $\sigma_P(\mathcal{F})$ in terms of $\sigma_n(\mathcal{F})$ to derive from (4.13) that

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq 2\|R_n\|_{\mathcal{F}} + 4\sigma_n(\mathcal{F})\sqrt{\frac{t}{n}} +$$

$$+12\sqrt{2}\sqrt{\|R_n\|_{\mathcal{F}}}\sqrt{\frac{t}{n}} + \frac{100t}{n} \leq 3\|R_n\|_{\mathcal{F}} + 4\sigma_n(\mathcal{F})\sqrt{\frac{t}{n}} + \frac{172t}{n}.$$

The last bound holds on the same event $E \cap F$ of probability at least $1-2e^{-t}$. This implies inequalities (4.7) and (4.8) of the theorem. Inequality (4.6) follows from Talagrand's inequality, specifically, from combination of Klein-Rio inequality (4.12), the following application of Bousquet's inequality

$$\|R_n\|_{\mathcal{F}} \leq \mathbb{E}\|R_n\|_{\mathcal{F}} + \sqrt{\frac{2t}{n}\Big(\sigma_P^2(\mathcal{F}) + 2\mathbb{E}\|R_n\|_{\mathcal{F}}\Big)} + \frac{t}{3n} \tag{4.15}$$

and bounds (4.7), (4.8) that have been already proved. The proof of the next inequality (4.9) is another application of symmetrization, contraction and Talagrand's concentration and is similar to the proof of (4.8). The last two bounds follow from the inequalities for the Rademacher process and symmetrization inequality.

Under the assumption $t \geq 1$, the exponent in the expression for probability can be written as e^{-t} without a constant in front of it. The constant can be removed by increasing the value of K.

□

We will use the above tools to construct data dependent bounds on the excess risk. As in the previous section, we assume that the functions in the class $\mathcal{F}$ are uniformly bounded by 1. First we show that the δ-minimal sets of the risk can be estimated by the δ-minimal sets of the empirical risk provided that δ is not too small, which is a consequence of Theorem 4.2. Let

$$\hat{\mathcal{F}}_n(\delta) := \mathcal{F}_{P_n}(\delta)$$

be the δ-minimal set of P_n.

Lemma 4.2 *Let $\delta_n^{\diamond}$ be a number such that $\delta_n^{\diamond} \geq U_n^{\sharp}\left(\frac{1}{2}\right)$. There exists an event of probability at least $1 - \sum_{\delta_j \geq \delta_n^{\diamond}} e^{-t_j}$ such that on this event, for all $\delta \geq \delta_n^{\diamond}$,*

$$\mathcal{F}(\delta) \subset \hat{\mathcal{F}}_n(3\delta/2) \text{ and } \hat{\mathcal{F}}_n(\delta) \subset \mathcal{F}(2\delta).$$

Proof. It easily follows from the definitions that $\delta_n^\diamond \geq \delta_n(\mathcal{F}; P)$. Denote

$$E_n := \bigcap_{\delta_j \geq \delta_n^\diamond} E_{n,j},$$

where $E_{n,j}$ are the events defined in the proof of Theorem 4.1. Then

$$\mathbb{P}(E_n) \geq 1 - \sum_{\delta_j \geq \delta_n^\diamond} e^{-t_j}.$$

It follows from the proof of Theorem 4.2, that, on the event E_n, for all $f \in \mathcal{F}$ with $\mathcal{E}(f) \geq \delta_n^\diamond$,

$$\frac{1}{2} \leq \frac{\hat{\mathcal{E}}_n(f)}{\mathcal{E}(f)} \leq \frac{3}{2}.$$

By the proof of Theorem 4.2, on the same event

$$\|P_n - P\|_{\mathcal{F}'(\delta_n^\diamond)} \leq U_n(\delta_n^\diamond).$$

Therefore, on the event E_n,

$$\mathcal{E}(f) \leq 2\hat{\mathcal{E}}_n(f) \vee \delta_n^\diamond, \; f \in \mathcal{F}, \tag{4.16}$$

which implies that, for all $\delta \geq \delta_n^\diamond$, $\hat{\mathcal{F}}_n(\delta) \subset \mathcal{F}(2\delta)$. On the other hand, on the same event E_n, for all $f \in \mathcal{F}$, the assumption $\mathcal{E}(f) \geq \delta_n^\diamond$ implies that $\hat{\mathcal{E}}_n(f) \leq \frac{3}{2}\mathcal{E}(f)$ and the assumption $\mathcal{E}(f) \leq \delta_n^\diamond$ implies that

$$\hat{\mathcal{E}}_n(f) \leq \mathcal{E}(f) + \|P_n - P\|_{\mathcal{F}'(\delta_n^\diamond)} \leq \mathcal{E}(f) + U_n(\delta_n^\diamond) \leq \delta_n^\diamond + V_n(\delta_n^\diamond)\delta_n^\diamond \leq \frac{3}{2}\delta_n^\diamond.$$

Thus, for all $f \in \mathcal{F}$,

$$\hat{\mathcal{E}}_n(f) \leq \frac{3}{2}\Big(\mathcal{E}(f) \vee \delta_n^\diamond\Big), \tag{4.17}$$

which implies that on the event E_n, for all $\delta \geq \delta_n^\diamond$, $\mathcal{F}(\delta) \subset \hat{\mathcal{F}}_n(3\delta/2)$.

□

Now we are ready to define an empirical version of excess risk bounds. It will be convenient to use the following definition of ρ_P :

$$\rho_P^2(f, g) := P(f - g)^2.$$

Given a decreasing sequence $\{\delta_j\}$ of positive numbers with $\delta_0 = 1$ and a sequence $\{t_j\}$ of real numbers, $t_j \geq 1$, define

$$\bar{U}_n(\delta) := \bar{K}\left(\phi_n(\delta_j) + D(\delta_j)\sqrt{\frac{t_j}{n}} + \frac{t_j}{n}\right), \quad \delta \in (\delta_{j+1}, \delta_j], j \geq 0,$$

where $\bar{K} = 2$. Comparing this with the definition (4.1) of the function U_n, it is easy to check that $U_n(\delta) \le \bar{U}_n(\delta), \delta \in (0, 1]$. As a consequence, if we define $\bar{\delta}_n := \bar{U}_n^{\sharp}(1/2)$, then $\delta_n(\mathcal{F}; P) \le \bar{\delta}_n$.

Empirical versions of the functions D and ϕ_n are defined by the following relationships:

$$\hat{D}_n(\delta) := \sup_{f,g \in \hat{\mathcal{F}}_n(\delta)} \rho_{P_n}(f, g) \quad \text{and} \quad \hat{\phi}_n(\delta) := \|R_n\|_{\hat{\mathcal{F}}_n'(\delta)}.$$

Also, let

$$\hat{U}_n(\delta) := \hat{K}\biggl(\hat{\phi}_n(\hat{c}\delta_j) + \hat{D}_n(\hat{c}\delta_j)\sqrt{\frac{t_j}{n}} + \frac{t_j}{n}\biggr), \quad \delta \in (\delta_{j+1}, \delta_j], j \ge 0,$$

$$\tilde{U}_n(\delta) := \tilde{K}\biggl(\phi_n(\tilde{c}\delta_j) + D(\tilde{c}\delta_j)\sqrt{\frac{t_j}{n}} + \frac{t_j}{n}\biggr), \quad \delta \in (\delta_{j+1}, \delta_j], j \ge 0,$$

where $2 \le \hat{K} \le \tilde{K}, \ \hat{c}, \tilde{c} \ge 1$ are numerical constants. Define

$$\bar{V}_n(\delta) := \bar{U}_n^{\flat}(\delta), \ \hat{V}_n(\delta) := \hat{U}_n^{\flat}(\delta), \ \tilde{V}_n(\delta) := \bar{U}_n^{\flat}(\delta)$$

and

$$\hat{\delta}_n := \hat{U}_n^{\sharp}(1/2), \quad \tilde{\delta}_n := \tilde{U}_n^{\sharp}(1/2).$$

The constants in the definitions of the functions $\bar{U}_n$ and $\tilde{U}_n$ can be chosen in such a way that for all δ $U_n(\delta) \le \bar{U}_n(\delta) \le \tilde{U}_n(\delta)$, which yields the bound $\delta_n(\mathcal{F}; P) \le \bar{\delta}_n \le \tilde{\delta}_n$. Since the definitions of the functions $U_n, \bar{U}_n, \tilde{U}_n$ differ only in the constants, it is plausible that the quantities $\delta_n(\mathcal{F}; P), \bar{\delta}_n, \tilde{\delta}_n$ are of the same order (in fact, it can be checked in numerous examples).

We will prove that with a high probability, for all δ, $\bar{U}_n(\delta) \le \hat{U}_n(\delta) \le \tilde{U}_n(\delta)$, so, $\hat{U}_n$ provides a data-dependent upper bound on $\bar{U}_n$ and $\tilde{U}_n$ provides a distribution dependent upper bound on $\hat{U}_n$. This implies that, with a high probability, $\bar{\delta}_n \le \hat{\delta}_n \le \tilde{\delta}_n$, which provides a data dependent bound $\hat{\delta}_n$ on the excess risk $\mathcal{E}_P(\hat{f}_n)$ which is of correct size (up to a constant) in many cases.

Theorem 4.8 *With the above notations,*

$$\mathbb{P}\Bigl\{\bar{\delta}_n \le \hat{\delta}_n \le \tilde{\delta}_n\Bigr\} \ge 1 - 3\sum_{\delta_j \ge \bar{\delta}_n} \exp\{-t_j\}.$$

Proof. The proof follows from the inequalities of Theorem 4.7 and Lemma 4.2 in a rather straightforward way. Note that $\bar{\delta}_n \geq U_n^{\sharp}(1/2)$, so we can use it as $\delta_n^{\diamond}$ in Lemma 4.2. Denote H the event introduced in the proof of this lemma (it was called E_n in the proof). Then

$$\mathbb{P}(H) \geq 1 - \sum_{\delta_j \geq \bar{\delta}_n} e^{-t_j}$$

and, on the event H,

$$\mathcal{F}(\delta) \subset \hat{\mathcal{F}}_n(3\delta/2) \text{ and } \hat{\mathcal{F}}_n(\delta) \subset \mathcal{F}(2\delta)$$

for all $\delta \geq \bar{\delta}_n$.

First, the values of δ and t will be fixed. At the end, the resulting bounds will be used for $\delta = \delta_j$ and $t = t_j$. We will apply the inequalities of Theorem 4.7 to the function class $\mathcal{F}'(\delta)$. It easily follows from bound (4.10) that there exists an event $F = F(\delta)$ of probability at least $1 - e^{-t}$ such that, on the event $H \cap F$,

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}'(\delta)} \leq K\left[\|R_n\|_{\hat{\mathcal{F}}'_n(3/2\delta)} + \hat{D}_n\Big(\frac{3}{2}\delta\Big)\sqrt{\frac{t}{n}} + \frac{t}{n}\right]$$

with a properly chosen K. Recalling the definition of $\bar{U}_n$ and $\hat{U}_n$, the last bound immediately implies that with a straightforward choice of numerical constants $\hat{K}, \hat{c}$, the inequality $\bar{U}_n(\delta) \leq \hat{U}_n(\delta)$. holds on the event $H \cap F$.

Quite similarly, using the inequalities of Theorem 4.7 (in particular, using bound (4.9) to control the "empirical" diameter $\hat{D}(\delta)$ in terms of the "true" diameter $D(\delta)$) and also the desymmetrization inequality, it is easy to see that there exists an event $G = G(\delta)$ of probability at least $1 - e^{-t}$ such that the inequality $\hat{U}_n(\delta) \leq \tilde{U}_n(\delta)$ holds on $H \cap G$ with properly chosen numerical constants $\tilde{K}, \tilde{c}$ in the definition of $\tilde{U}_n$.

Using the resulting inequalities for $\delta = \delta_j \geq \bar{\delta}_n$ yields

$$\mathbb{P}(E) \geq 1 - 3\sum_{\delta_j \geq \bar{\delta}_n} \exp\{-t_j\},$$

where

$$E := \left\{\forall \delta_j \geq \bar{\delta}_n : \ \bar{U}_n(\delta_j) \leq \hat{U}_n(\delta_j) \leq \tilde{U}_n(\delta_j)\right\} \supset \bigcup_{j:\delta_j \geq \bar{\delta}_n} (H \cap F(\delta_j) \cap G(\delta_j)).$$

By the definitions of $\bar{U}_n, \hat{U}_n$ and $\tilde{U}_n$, this also implies that, on the event E,

$$\bar{U}_n(\delta) \leq \hat{U}_n(\delta) \leq \tilde{U}_n(\delta)$$

for all $\delta \geq \bar{\delta}_n$. By simple properties of $\sharp$-transform, we conclude that $\bar{\delta}_n \leq \hat{\bar{\delta}}_n \leq \tilde{\bar{\delta}}_n$ on the event E, which completes the proof.

□

It is easily seen from the proof of Theorem 4.8 and from the definitions and constructions of the events involved in this proof as well as in the proofs of Theorem 4.2 and Lemma 4.2 that on an event E of probability at least $1-p$, where $p = 3\sum_{\delta_j \geq \bar{\delta}_n} e^{-t_j}$, the following conditions hold:

(i) $\bar{\delta}_n \leq \hat{\bar{\delta}}_n \leq \tilde{\bar{\delta}}_n$;

(ii) $\mathcal{E}(\hat{f}) \leq \bar{\delta}_n$;

(iii) for all $f \in \mathcal{F}$,

$$\mathcal{E}(f) \leq 2\hat{\mathcal{E}}_n(f) \vee \bar{\delta}_n$$

and

$$\hat{\mathcal{E}}_n(f) \leq \frac{3}{2}\Big(\mathcal{E}(f) \vee \bar{\delta}_n\Big);$$

(iv) for all $\delta \geq \bar{\delta}_n$,

$$\|P_n - P\|_{\mathcal{F}'(\delta)} \leq U_n(\delta).$$

Sometimes it is convenient to deal with different triples $(\bar{\delta}_n, \hat{\bar{\delta}}_n, \tilde{\bar{\delta}}_n)$ (defined in terms of various complexity measures of the class $\mathcal{F}$) that still satisfy conditions (i)-(iv) with a high probability. In fact, to satisfy conditions (ii)-(iv) it is enough to choose $\bar{\delta}_n$ in such a way that

(v) $\bar{\delta}_n \geq U_n^{\sharp}(1/2)$.

This is reflected in the following definition.

Definition 4.1 *Suppose sequences $\{\delta_j\}$, $\{t_j\}$ and the corresponding function U_n are given. We will call $\bar{\delta}_n$ that depends on $\mathcal{F}$ and P an* **admissible distribution dependent bound** *on excess risk iff it satisfies condition (v), and, as a consequence, also conditions (ii)-(iv). If (ii)-(iv) hold on an event E such that $\mathbb{P}(E) \geq 1-p$, then $\bar{\delta}_n$ will be called an admissible bound of confidence level $1-p$. A triple $(\bar{\delta}_n, \hat{\bar{\delta}}_n, \tilde{\bar{\delta}}_n)$, such that $\bar{\delta}_n$ and $\tilde{\bar{\delta}}_n$ depend on $\mathcal{F}$ and P, $\hat{\bar{\delta}}_n$ depends on $\mathcal{F}$ and $X_1, \ldots, X_n$, and, for some $p \in (0,1)$, conditions (i)-(v) hold on an event E with $\mathbb{P}(E) \geq 1-p$, will be called a* **triple bound** *on the excess risk of confidence level $1-p$.*

Such triple bounds will be used later in model selection methods based on penalized empirical risk minimization.

We conclude this section with a simple example showing that *in the multiple minima case* the distribution dependent excess risk bounds developed in the previous section are not always sharp. Moreover, there is a difficulty in estimation of the level sets of the risk (the δ-minimal sets), which was of importance in constructing data dependent excess risk bounds. Some more subtle geometric characteristics of the class $\mathcal{F}$ that can be used in such cases to recover the correct convergence rates were suggested in Koltchinskii [59]. However, the development of the theory of excess risk bounds in the multiple minima case remains an open problem.

Recall the definition of $\delta_n(t)$ in Corollary 4.1.

Proposition 4.3 *Let $S := \{0,1\}^{N+1}$ and P be the uniform distribution on $\{0,1\}^{N+1}$. Let $\mathcal{F} := \{f_j : 1 \le j \le N+1\}$, where*

$$f_j(x) = x_j, \quad x = (x_1, \dots, x_{N+1}) \in \{0,1\}^{N+1}.$$

Then the following statements hold for an empirical risk minimizer $\hat{f}$:
(i) $\mathcal{E}_P(\hat{f}) = 0$;
(ii) with some $c > 0$,

$$\delta_n(t) \ge c\biggl(\sqrt{\frac{\log N}{n}} + \sqrt{\frac{t}{n}}\biggr);$$

(iii) for any $\varepsilon > 0$ there exists N_0 such that, for $N_0 \le N \le \sqrt{n}$ and for $\delta = 0.25\sqrt{\frac{\log N}{n}}$, the inclusion $\mathcal{F}(0) \subset \hat{\mathcal{F}}_n(\delta)$ **does not hold** *with probability at least $1 - \varepsilon$.*

Proof. For $k \ne j$, $P(f_k - f_j)^2 = 1/2$. Thus, $D_P(\mathcal{F}; \delta) = 1/2$. At the same time,

$$\phi_n(\delta) = \mathbb{E} \sup_{f,g \in \mathcal{F}} |(P_n - P)(f - g)| = \mathbb{E} \max_{1 \le k,j \le N} |(P_n - P)(f_k - f_j)|.$$

It is easy to check that the last expectation is of the order $c\sqrt{\frac{\log N}{n}}$. It implies that the value of $\delta_n(t)$ is of the order $c\biggl(\sqrt{\frac{\log N}{n}} + \sqrt{\frac{t}{n}}\biggr)$. The excess risk $\mathcal{E}(f)$ is equal to 0 for all $f \in \mathcal{F}$. In particular, $\mathcal{E}(\hat{f}_n) = 0$. Thus, the bound $\delta_n(t)$ is not sharp.

To show that (iii) holds, note that

$$\mathbb{P}\Bigl\{\mathcal{F}(0) \subset \hat{\mathcal{F}}_n(\delta)\Bigr\} = \mathbb{P}\Bigl\{\hat{\mathcal{F}}_n(\delta) = \mathcal{F}\Bigr\} =$$

$$\mathbb{P}\Bigl\{\forall j, 1 \le j \le N+1 : P_n f_j \le \min_{1 \le k \le N+1} P_n f_k + \delta\Bigr\} \le$$

$$\leq \mathbb{P}\Big\{\forall j, 1 \leq j \leq N : P_n f_j \leq P_n f_{N+1} + \delta\Big\} = \mathbb{P}\Big\{\forall j, 1 \leq j \leq N : \nu_{n,j} \leq \nu_n + \delta n\Big\},$$

where $\nu_n, \nu_{n,j}, 1 \leq j \leq N$ are i.i.d. binomial random variables with parameters n and $1/2$. Therefore,

$$\mathbb{P}\Big\{\mathcal{F}(0) \subset \hat{\mathcal{F}}_n(\delta)\Big\} \leq \sum_{k=0}^{n} \mathbb{P}\{\nu_n = k\}\mathbb{P}\Big\{\forall j, 1 \leq j \leq N : \nu_{n,j} \leq k + \delta n \Big| \nu_n = k\Big\} =$$

$$\sum_{k=0}^{n} \mathbb{P}\{\nu_n = k\} \prod_{j=1}^{N} \mathbb{P}\{\nu_{n,j} \leq k + \delta n\} = \sum_{k=0}^{n} \mathbb{P}\{\nu_n = k\}\mathbb{P}^N\{\nu_n \leq k + \delta n\} \leq$$

$$\mathbb{P}\{\nu_n > \bar{k}\} + \mathbb{P}^N\{\nu_n \leq \bar{k} + \delta n\},$$

where $0 \leq \bar{k} \leq n$. Let $\bar{k} = \frac{n}{2} + n\delta$. By Bernstein's inequality,

$$\mathbb{P}\{\nu_n > \bar{k}\} \leq \exp\Big\{-\frac{n\delta^2}{4}\Big\} = (\log N)^{-2^{-6}}.$$

On the other hand, by the normal approximation of binomial distribution (Φ being the standard normal distribution function)

$$\mathbb{P}\{\nu_n \leq \bar{k} + \delta n\} \leq \Phi(4\delta\sqrt{n}) + n^{-1/2} = \Phi(\sqrt{\log N}) + n^{-1/2}.$$

Under the condition $N_0 \leq N \leq \sqrt{n}$ this yields, for a large enough N_0,

$$\mathbb{P}\{\mathcal{F}(0) \subset \hat{\mathcal{F}}_n(\delta)\} \leq \varepsilon,$$

and the result follows.

□

5 Examples of Excess Risk Bounds in Prediction Problems

Let (X, Y) be a random couple in $S \times T$, $T \subset \mathbb{R}$ with distribution P. The distribution of X will be denoted by Π. Assume that the random variable X is "observable" and Y is to be predicted based on an observation of X. Let $\ell : T \times \mathbb{R} \mapsto \mathbb{R}$ be a loss function. Given a function $g : S \mapsto \mathbb{R}$, the quantity $(\ell \bullet g)(x, y) := \ell(y, g(x))$ is interpreted as a loss suffered when $g(x)$ is used to predict y. The problem of optimal prediction can be viewed as a risk minimization

$$\mathbb{E}\ell(Y, g(X)) = P(\ell \bullet g) \longrightarrow \min,\ g : S \mapsto \mathbb{R}.$$

Since the distribution P and the risk function $g \mapsto P(\ell \bullet g)$ are unknown, the risk minimization problem is usually replaced by the empirical risk minimization

$$P_n(\ell \bullet g) = n^{-1} \sum_{j=1}^{n} \ell(Y_j, g(X_j)) \longrightarrow \min, \ g \in \mathcal{G},$$

where $\mathcal{G}$ is a given class of functions $g : S \mapsto \mathbb{R}$ and $(X_1, Y_1), \dots, (X_n, Y_n)$ is a sample of i.i.d. copies of (X, Y) ("training data"). Obviously, this can be viewed as a special case of abstract empirical risk minimization problems discussed in the previous section. In this case, the class $\mathcal{F}$ is the "loss class" $\mathcal{F} := \{\ell \bullet g : g \in \mathcal{G}\}$ and the goal of this section is to derive excess risk bounds for concrete examples of loss functions and function classes frequently used in Statistics and Learning Theory.

Let μ_x denote a version of conditional distribution of Y given $X = x$. The following representation of the risk holds under very mild regularity assumptions:

$$P(\ell \bullet g) = \int_S \int_T \ell(y; g(x)) \mu_x(dy) \Pi(dx)$$

Given a probability measure μ on T, let

$$u_\mu \in \mathrm{Argmin}_{u \in \bar{\mathbb{R}}} \int_T \ell(y; u) \mu(dy).$$

Define

$$g_*(x) := u_{\mu_x} = \mathrm{argmin}_{u \in \bar{\mathbb{R}}} \int_T \ell(y; u) \mu_x(dy).$$

Assume that the function g_* is well defined and properly measurable. Then, for all g, $P(\ell \bullet g) \geq P(\ell \bullet g_*)$ so, g_* is a point of *global* minimum of $P(\ell \bullet g)$.

Let

$$\hat{g}_n := \mathrm{argmin}_{g \in \mathcal{G}} P_n(\ell \bullet g)$$

be a solution of the corresponding empirical risk minimization problem (for simplicity, assume its existence).

The following assumption on the loss function ℓ is often used in the analysis of the problem: there exists a function $D(u, \mu) \geq 0$ such that for all measures $\mu = \mu_x, \ x \in S$

$$\int_T (\ell(y, u) - \ell(y, u_\mu))^2 \mu(dy) \leq D(u, \mu) \int_T (\ell(y, u) - \ell(y, u_\mu)) \mu(dy). \tag{5.1}$$

In the case when the functions in the class $\mathcal{G}$ take their values in the interval $[-M/2, M/2]$ and

$$D(u, \mu_x), \ |u| \leq M/2, x \in S$$

is uniformly bounded by a constant $D > 0$, it immediately follows from (5.1) (just by plugging in $u = g(x)$, $\mu = \mu_x$ and integrating with respect to Π.) that, for all $g \in \mathcal{G}$,

$$P(\ell \bullet g - \ell \bullet g_*)^2 \leq DP(\ell \bullet g - \ell \bullet g_*). \tag{5.2}$$

As a consequence, if $g_* \in \mathcal{G}$, then the $L_2(P)$-diameter of the δ-minimal set of $\mathcal{F}$ is bounded as follows:

$$D(\mathcal{F}; \delta) \leq 2(D\delta)^{1/2}.$$

Moreover, even if $g_* \notin \mathcal{G}$, the condition (4.4) still holds for the loss class $\mathcal{F}$ with $f_* = \ell \bullet g_*$, providing a link between the excess risk (approximation error) and the variance of the "excess loss" and opening a way for Massart's type penalization methods (see sections 4.1, 6.3). The idea to control variance in terms of expectation has been extensively used in Massart [73] (and even in a much earlier work of Birgé and Massart) as well as in the learning theory literature (Mendelson [76], Bartlett, Jordan and McAuliffe [10], Blanchard, Lugosi and Vayatis [17], Bartlett, Bousquet and Mendelson [7]).

5.1 Regression with Quadratic Loss

We start with regression problems with bounded response and with quadratic loss. To be specific, assume that Y takes values in $T = [0, 1]$ and $\ell(y, u) := (y - u)^2,\ y \in T, u \in \mathbb{R}$. The minimum of the risk

$$P(\ell \bullet g) = \mathbb{E}(Y - g(X))^2$$

over the set of all measurable functions $g : S \mapsto \mathbb{R}$ is attained at the regression function

$$g_*(x) := \eta(x) := \mathbb{E}(Y|X = x).$$

If $\mathcal{G}$ is a class of measurable functions from S into $[0, 1]$ such that $g_* \in \mathcal{G}$, then it is easy to check that for all $g \in \mathcal{G}$

$$\mathcal{E}_P(\ell \bullet g) = \|g - g_*\|^2_{L_2(\Pi)}.$$

In general, the excess risk is given by

$$\mathcal{E}_P(\ell \bullet g) = \|g - g_*\|^2_{L_2(\Pi)} - \inf_{h \in \mathcal{G}} \|h - g_*\|^2_{L_2(\Pi)}.$$

The following lemma provides an easy way to bound the excess risk from below in the case of a *convex class* $\mathcal{G}$ and $\bar{g} := \mathrm{argmin}_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)}$.

Lemma 5.1 *If $\mathcal{G}$ is a convex class of functions, then*

$$2\mathcal{E}_P(\ell \bullet g) \geq \|g - \bar{g}\|^2_{L_2(\Pi)}.$$

Proof. Note that the identity

$$\frac{u^2 + v^2}{2} - \left(\frac{u+v}{2}\right)^2 = \frac{(u-v)^2}{4}$$

implies that

$$\frac{(g - g_*)^2 + (\bar{g} - g_*)^2}{2} = \left(\frac{g + \bar{g}}{2} - g_*\right)^2 + \frac{(g - \bar{g})^2}{4}.$$

Integrating with respect to Π yields

$$\frac{\|g - g_*\|^2_{L_2(\Pi)} + \|\bar{g} - g_*\|^2_{L_2(\Pi)}}{2} = \left\|\frac{g + \bar{g}}{2} - g_*\right\|^2_{L_2(\Pi)} + \frac{\|g - \bar{g}\|^2_{L_2(\Pi)}}{4}.$$

Since $\mathcal{G}$ is convex and $g, \bar{g} \in \mathcal{G}$, we have $\frac{g+\bar{g}}{2} \in \mathcal{G}$ and

$$\left\|\frac{g + \bar{g}}{2} - g_*\right\|^2_{L_2(\Pi)} \geq \|\bar{g} - g_*\|^2_{L_2(\Pi)}.$$

Therefore,

$$\frac{\|g - g_*\|^2_{L_2(\Pi)} + \|\bar{g} - g_*\|^2_{L_2(\Pi)}}{2} \geq \|\bar{g} - g_*\|^2_{L_2(\Pi)} + \frac{\|g - \bar{g}\|^2_{L_2(\Pi)}}{4},$$

implying the result.

□

As before, we denote $\mathcal{F} := \{\ell \bullet g : g \in \mathcal{G}\}$. It follows from Lemma 5.1 that

$$\mathcal{F}(\delta) \subset \{\ell \bullet g : \|g - \bar{g}\|^2_{L_2(\Pi)} \leq 2\delta\}.$$

Also, for all functions $g_1, g_2 \in \mathcal{G}$ and all $x \in S, y \in T$,

$$\Big|(\ell \bullet g_1)(x,y) - (\ell \bullet g_2)(x,y)\Big| = \Big|(y - g_1(x))^2 - (y - g_2(x))^2\Big|$$

$$= |g_1(x) - g_2(x)||2y - g_1(x) - g_2(x)| \leq 2|g_1(x) - g_2(x)|,$$

which implies

$$P\Big(\ell \bullet g_1 - \ell \bullet g_2\Big)^2 \leq 4\|g_1 - g_2\|^2_{L_2(\Pi)}.$$

Hence

$$D(\delta) \le 2\sup\Big\{\|g_1 - g_2\|_{L_2(\Pi)} : \|g_1 - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta, \|g_2 - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta\Big\} \le 4\sqrt{2}\sqrt{\delta}.$$

In addition, by symmetrization inequality,

$$\phi_n(\delta) = \mathbb{E}\|P_n - P\|_{\mathcal{F}'(\delta)} \le 2\mathbb{E}\|R_n\|_{\mathcal{F}'(\delta)} \le$$

$$2\mathbb{E}\sup\Big\{\Big|R_n(\ell \bullet g_1 - \ell \bullet g_2)\Big| : g_1, g_2 \in \mathcal{G}, \|g_1 - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta, \|g_2 - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta\Big\} \le$$

$$4\mathbb{E}\sup\Big\{\Big|R_n(\ell \bullet g - \ell \bullet \bar{g})\Big| : g \in \mathcal{G}, \|g - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta\Big\},$$

and since $\ell(y, \cdot)$ is Lipschitz with constant 2 on the interval $[0, 1]$ one can use the contraction inequality to get

$$\phi_n(\delta) \le 16\mathbb{E}\sup\{|R_n(g - \bar{g})| : g \in \mathcal{G}, \|g - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta\} =: \psi_n(\delta).$$

As a result, we get

$$\phi_n^\flat(\sigma) \le \psi_n^\flat(\sigma)$$

and

$$\sqrt{(D^2)^\flat(\sigma)} \le 4\sqrt{2}.$$

This yields an upper bound on the quantity σ_n^t involved in Theorem 4.3:

$$\sigma_n^t \le K\bigg(\psi_n^\sharp\Big(\frac{1}{2q}\Big) + \frac{t}{n}\bigg).$$

Thus, the following statement is a corollary of Theorem 4.3.

Theorem 5.1 *Let $\mathcal{G}$ be a convex class of functions from S into $[0, 1]$ and let $\hat{g}$ denotes the least square estimator of the regression function*

$$\hat{g} := \text{argmin}_{g \in \mathcal{G}} n^{-1} \sum_{j=1}^n (Y_j - g(X_j))^2.$$

Then, there exist constants $K > 0, C > 0$ such that for all $t > 0$,

$$\mathbb{P}\bigg\{\|\hat{g} - g_*\|^2_{L_2(\Pi)} \ge \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + K\bigg(\psi_n^\sharp\Big(\frac{1}{2q}\Big) + \frac{t}{n}\bigg)\bigg\} \le Ce^{-t}.$$

A slightly weaker result holds in the case when the class $\mathcal{G}$ is not necessarily convex. It follows from Lemma 4.1. Note that the condition

$$4(P(\ell \bullet g) - P(\ell \bullet g_*)) = 4\|g - g_*\|^2_{L_2(\Pi)} =: \rho^2_P(\ell \bullet g, \ell \bullet g_*) \geq P(\ell \bullet g - \ell \bullet g_*)^2$$

is satisfied for all functions $g : S \mapsto [0,1]$. Also,

$$\omega_n(\delta) = \mathbb{E} \sup_{4\|g-\bar{g}\|^2_{L_2(\Pi)} \leq \delta} \Big|(P_n - P)(\ell \bullet g - \ell \bullet \bar{g})\Big| \leq \frac{1}{2}\psi_n(\delta/8)$$

(by symmetrization and contraction inequalities, and by notation).

Therefore, the following result holds.

Theorem 5.2 *Let $\mathcal{G}$ be a class of functions from S into $[0,1]$ and let $\hat{g}$ denote the least square estimator of the regression function. Then, there exist constants $K > 0, C > 0$ such that for all $t > 0$,*

$$\mathbb{P}\bigg\{\|\hat{g} - g_*\|^2_{L_2(\Pi)} \geq (1+\varepsilon) \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + \frac{1}{4}\psi^\sharp_n\Big(\frac{\varepsilon}{K}\Big) + \frac{Kt}{n\varepsilon}\bigg\} \leq Ce^{-t}.$$

Clearly, these results hold (with different constants) if the functions in $\mathcal{G}$ take their values in an arbitrary bounded interval.

Example 1. Finite dimensional classes. Suppose that $L \subset L_2(\Pi)$ is a finite dimensional linear space with $\dim(L) = d < \infty$ and let $\mathcal{G} \subset L$ be a convex class of functions taking values in a bounded interval (for simplicity, $[0,1]$). It follows from Proposition 3.2 that

$$\psi_n(\delta) \leq C\sqrt{\frac{d\delta}{n}}$$

with some constant $C > 0$. Hence,

$$\psi^\sharp_n\Big(\frac{1}{2q}\Big) \leq K\frac{d}{n}$$

and Theorem 5.1 implies that

$$\mathbb{P}\bigg\{\|\hat{g} - g_*\|^2_{L_2(\Pi)} \geq \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + K\Big(\frac{d}{n} + \frac{t}{n}\Big)\bigg\} \leq Ce^{-t}$$

with some constant $K > 0$.

Example 2. Reproducing kernel Hilbert spaces (RKHS). Suppose $\mathcal{G}$ is the unit ball in RKHS $\mathcal{H}_K$:

$$\mathcal{G} := \{h : \|h\|_{\mathcal{H}_K} \leq 1\}.$$

Denote $\{\lambda_k\}$ the eigenvalues of the integral operator from $L_2(\Pi)$ into $L_2(\Pi)$ with kernel K. Then Proposition 3.3 implies that

$$\psi_n(\delta) \leq C\biggl(n^{-1}\sum_{j=1}^{\infty}(\lambda_j \wedge \delta)\biggr)^{1/2}.$$

The function

$$\delta \mapsto \biggl(n^{-1}\sum_{j=1}^{\infty}(\lambda_j \wedge \delta)\biggr)^{1/2} =: \gamma_n(\delta)$$

is strictly convex and, as a result,

$$\gamma_n^{\flat}(\delta) = \frac{\gamma_n(\delta)}{\delta}$$

is strictly decreasing. By a simple computation, Theorem 5.1 yields

$$\mathbb{P}\biggl\{\|\hat{g} - g_*\|^2_{L_2(\Pi)} \geq \inf_{g\in\mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + K\biggl(\gamma_n^{\sharp}(1) + \frac{t}{n}\biggr)\biggr\} \leq Ce^{-t}$$

with some constant $K > 0$.

Example 3. VC-subgraph classes. Suppose that $\mathcal{G}$ is a VC-subgraph class of functions $g : S \mapsto [0,1]$ of VC-dimension V. Then the function $\psi_n(\delta)$ can be upper bounded using (3.13):

$$\psi_n(\delta) \leq C\biggl[\sqrt{\frac{V\delta}{n}\log\frac{1}{\delta}} \bigvee \frac{V}{n}\log\frac{1}{\delta}\biggr].$$

Therefore

$$\psi_n^{\sharp}(\varepsilon) \leq \frac{CV}{n\varepsilon^2}\log\frac{n\varepsilon^2}{V}.$$

Theorem 5.2 implies

$$\mathbb{P}\biggl\{\|\hat{g} - g_*\|^2_{L_2(\Pi)} \geq (1+\varepsilon)\inf_{g\in\mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + K\biggl(\frac{V}{n\varepsilon^2}\log\frac{n\varepsilon^2}{V} + \frac{t}{n\varepsilon}\biggr)\biggr\} \leq Ce^{-t}.$$

Example 4. Entropy conditions. In the case when the entropy of the class $\mathcal{G}$ (random, uniform, bracketing, etc.) is bounded by $O(\varepsilon^{-2\rho})$ for some $\rho \in (0,1)$, we typically have

$$\psi_n^{\sharp}(\varepsilon) = O\Bigl(n^{-1/(1+\rho)}\Bigr).$$

For instance, if (3.14) holds, then it follows from (3.15) (with $F \equiv U = 1$ for simplicity) that

$$\psi_n(\delta) \leq K\biggl(\frac{A^{\rho}}{\sqrt{n}}\delta^{(1-\rho)/2} \bigvee \frac{A^{2\rho/(\rho+1)}}{n^{1/(1+\rho)}}\biggr).$$

Therefore,

$$\psi_n^\sharp(\varepsilon) \le \frac{CA^{2\rho/(1+\rho)}}{(n\varepsilon^2)^{1/(1+\rho)}}.$$

In this case Theorem 5.2 gives the bound

$$\mathbb{P}\biggl\{\|\hat g - g_*\|^2_{L_2(\Pi)} \ge (1+\varepsilon)\inf_{g\in\mathcal{G}}\|g-g_*\|^2_{L_2(\Pi)} + K\biggl(\frac{A^{2\rho/(1+\rho)}}{(n\varepsilon^2)^{1/(1+\rho)}} + \frac{t}{n\varepsilon}\biggr)\biggr\} \le Ce^{-t}.$$

Example 5. Convex hulls. If

$$\mathcal{G} := \mathrm{conv}(\mathcal{H}) := \biggl\{\sum_j \lambda_j h_j : \sum_j |\lambda_j| \le 1, h_j \in \mathcal{H}\biggr\}$$

is the symmetric convex hull of a given VC-type class $\mathcal{H}$ of measurable functions from S into $[0,1]$, then the condition of the previous example is satisfied with $\rho := \frac{V}{V+2}$. This yields

$$\psi_n^\sharp(\varepsilon) \le \biggl(\frac{K(V)}{n\varepsilon^2}\biggr)^{\frac{1}{2}\frac{2+V}{1+V}}$$

and Theorem 5.1 yields

$$\mathbb{P}\biggl\{\|\hat g - g_*\|^2_{L_2(\Pi)} \ge \inf_{g\in\mathcal{G}}\|g-g_*\|^2_{L_2(\Pi)} + K\biggl(\biggl(\frac{1}{n}\biggr)^{\frac{1}{2}\frac{2+V}{1+V}} + \frac{t}{n}\biggr)\biggr\} \le Ce^{-t}$$

with some constant $K>0$ depending on V.

5.2 Empirical Risk Minimization with Convex Loss

A standard assumption on the loss function ℓ that makes the empirical risk minimization problem computationally tractable is that $\ell(y,\cdot)$ is a convex function for all $y \in T$. Assuming, in addition, that $\mathcal{G}$ is a convex class of functions, the convexity of the loss implies that the empirical risk $\mathcal{G} \ni g \mapsto P_n(\ell \bullet g)$ is a convex functional and the empirical risk minimization is a convex minimization problem. We will call the problems of this type *convex risk minimization.* The least squares and the L_1-regression as well as some of the methods of large margin classification (such as boosting) are examples of convex risk minimization.

The convexity assumption also simplifies the analysis of empirical risk minimization problems. In particular, it makes easier proving the existence of the minimal point g_*, checking condition (5.1), etc.. In this section, we extend the results for L_2-regression to this more general framework.

Assume the functions in $\mathcal{G}$ take their values in $[-M/2, M/2]$. We will need the following assumptions on the loss function ℓ : ℓ satisfies the Lipschitz condition with some $L > 0$

$$\forall y \in T \ \forall u, v \in [-M/2, M/2] \ \ |\ell(y,u) - \ell(y,v)| \leq L|u - v| \tag{5.3}$$

and also the following assumption on convexity modulus of ℓ holds with some $\Lambda > 0$:

$$\forall y \in T \ \forall u, v \in [-M/2, M/2] \ \ \frac{\ell(y,u) + \ell(y,v)}{2} - \ell\biggl(y; \frac{u+v}{2}\biggr) \geq \Lambda|u - v|^2. \tag{5.4}$$

Note that, if g_* is bounded by $M/2$, conditions (5.3) and (5.4) imply (5.1) with $D(u,\mu) \leq \frac{L^2}{2\Lambda}$. To see this, it is enough to use (5.4) with $v = u_\mu$, $\mu = \mu_x$ and to integrate it with respect to μ to get, for the function $L(u) := \int_T \ell(y,u)\mu(dy)$, (note that the minimum of L is attained at u_μ)

$$\frac{L(u) - L(u_\mu)}{2} = \frac{L(u) + L(u_\mu)}{2} - L(u_\mu) \geq \frac{L(u) + L(u_\mu)}{2} - L\biggl(\frac{u + u_\mu}{2}\biggr) \geq \Lambda|u - u_\mu|^2$$

and then to use the Lipschitz condition to get

$$\int_T |\ell(y,u) - \ell(y,u_\mu)|^2 \mu(dy) \leq L^2|u - u_\mu|^2.$$

This nice and simple trick, based on strict convexity, has been used repeatedly in the theory (see, for instance, Bartlett, Jordan and McAuliffe [10]). We will use it in the proof of Theorem 5.3.

Theorem 5.3 *Suppose that $\mathcal{G}$ is a convex class of functions taking values in $[-M/2, M/2]$. Assume that the minimum of $P(\ell \bullet g)$ over $\mathcal{G}$ is attained at $\bar{g} \in \mathcal{G}$ and*

$$\omega_n(\delta) := \mathbb{E} \sup_{g \in \mathcal{G}, \|g - \bar{g}\|^2_{L_2(\Pi)} \leq \delta} |R_n(g - \bar{g})|.$$

Denote

$$\hat{g} := \mathrm{argmin}_{g \in \mathcal{G}} P_n(\ell \bullet g).$$

Then there exist constants $K > 0, C > 0, c > 0$ such that

$$\mathbb{P}\biggl\{P(\ell \bullet \hat{g}) \geq \inf_{g \in \mathcal{G}} P(\ell \bullet g) + K\biggl(\Lambda\omega_n^\sharp\biggl(\frac{c\Lambda}{L}\biggr) + \frac{L^2}{\Lambda}\frac{t}{n}\biggr)\biggr\} \leq Ce^{-t}, \ t > 0.$$

Proof. Note that by Lipschitz condition (5.3), for all $g_1, g_2 \in \mathcal{G}$,

$$P|\ell \bullet g_1 - \ell \bullet g_2|^2 \le L^2 \|g_1 - g_2\|^2_{L_2(\Pi)}.$$

On the other hand, by (5.4), for all $g \in \mathcal{G}, x \in S, y \in T$,

$$\frac{\ell(y, g(x)) + \ell(y, \bar{g}(x))}{2} \ge \ell\biggl(y; \frac{g(x) + \bar{g}(x)}{2}\biggr) + \Lambda |g(x) - \bar{g}(x)|^2.$$

Integrating this inequality and observing that $\frac{g+\bar{g}}{2} \in \mathcal{G}$ and hence

$$P\biggl(\ell \bullet \biggl(\frac{g+\bar{g}}{2}\biggr)\biggr) \ge P(\ell \bullet \bar{g}),$$

yields

$$\frac{P(\ell \bullet g) + P(\ell \bullet \bar{g})}{2} \ge P(\ell \bullet \bar{g}) + \Lambda \Pi |g - \bar{g}|^2,$$

or

$$P(\ell \bullet g) - P(\ell \bullet \bar{g}) \ge 2\Lambda \Pi |g - \bar{g}|^2.$$

For the loss class $\mathcal{F} = \{\ell \bullet g : g \in \mathcal{G}\}$, this gives the following upper bound on the $L_2(P)$-diameter of the δ-minimal set $\mathcal{F}(\delta) : D^2(\delta) \le \frac{2\delta}{\Lambda}$. By symmetrization and contraction inequalities, it is easy to bound

$$\phi_n(\delta) = \mathbb{E}\|P_n - P\|_{\mathcal{F}'(\delta)}$$

in terms of $\omega_n(\delta)$:

$$\phi_n(\delta) \le CL\omega_n\biggl(\frac{\delta}{2\Lambda}\biggr).$$

By a simple computation, the quantity σ_n^t used in Theorem 4.3 is bounded as follows:

$$\sigma_n^t \le K\biggl(\Lambda \omega_n^\sharp\biggl(\frac{c\Lambda}{L}\biggr) + \frac{L^2}{\Lambda}\frac{t}{n}\biggr).$$

Under the additional assumption that ℓ is uniformly bounded by 1 in $T \times [-M/2, M/2]$, Theorem 4.3 implies the result. To get rid of the extra assumption, suppose that ℓ is uniformly bounded by D on $T \times [-M/2, M/2]$. Then the result holds for the loss function ℓ/D. For this loss function, L and Λ are replaced by L/D and Λ/D, and the expression

$$\Lambda \omega_n^\sharp\biggl(\frac{c\Lambda}{L}\biggr) + \frac{L^2}{\Lambda}\frac{t}{n}$$

becomes

$$\Lambda/D\omega_n^\sharp\biggl(\frac{c\Lambda/D}{L/D}\biggr) + \frac{L^2/D^2}{\Lambda/D}\frac{t}{n} = \frac{1}{D}\biggl(\Lambda\omega_n^\sharp\biggl(\frac{c\Lambda}{L}\biggr) + \frac{L^2}{\Lambda}\frac{t}{n}\biggr),$$

so the result follows by rescaling.

□

As an example, consider the case when $\mathcal{G} := M\mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a base class of functions from S into $[-1/2, 1/2]$. There are many powerful functional gradient descent type algorithms (such as boosting) that provide an implementation of convex empirical risk minimization over a convex hull or a linear span of a given base class. Assume that condition (3.12) holds for the class $\mathcal{H}$ with some $V > 0$, i.e., $\mathcal{H}$ is a VC-type class. Define

$$\pi_n(M, L, \Lambda; t) := K\biggl[\Lambda M^{V/(V+1)}\Bigl(\frac{L}{\Lambda}\bigvee 1\Bigr)^{(V+2)/(V+1)} n^{-\frac{1}{2}\frac{V+2}{V+1}} + \frac{L^2}{\Lambda}\frac{t}{n}\biggr]$$

with a numerical constant K. The next result is a slightly generalized version of a theorem due to Bartlett, Jordan and McAuliffe [10].

Theorem 5.4 *Under the conditions (5.3) and (5.4),*

$$\mathbb{P}\biggl\{P(\ell \bullet \hat{g}_n) \geq \min_{g\in\mathcal{G}} P(\ell \bullet g) + \pi_n(M, L, \Lambda; t)\biggr\} \leq Ce^{-t}.$$

Proof. To apply Theorem 5.3, it is enough to bound the function ω_n. Since $\mathcal{G} := M\mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a VC-type class of functions from S into $[-1/2, 1/2]$, condition (3.12) holds for $\mathcal{H}$ with envelope $F \equiv 1$ (see Theorem 3.12). Together with (3.15), this gives

$$\omega_n(\delta) \leq C\biggl[\frac{M^\rho}{\sqrt{n}}\delta^{(1-\rho)/2}\bigvee\frac{M^{2\rho/(\rho+1)}}{n^{1/(1+\rho)}}\biggr]$$

with $\rho := \frac{V}{V+2}$. Hence,

$$\omega_n^\sharp(\varepsilon) \leq C\frac{M^{2\rho/(1+\rho)}}{n^{1/(1+\rho)}}\varepsilon^{-2/(1+\rho)}$$

for $\varepsilon \leq 1$. If $\ell(y, \cdot)$ is bounded by 1 in $T \times [-M/2, M/2]$, then Theorem 5.3 yields

$$\mathbb{P}\biggl\{P(\ell \bullet \hat{g}) \geq \min_{g\in\mathcal{G}} P(\ell \bullet g) + \pi_n(M, L, \Lambda; t)\biggr\} \leq Ce^{-t}.$$

To remove the assumption that ℓ is bounded by 1, one should use the same rescaling argument as in the proof of Theorem 5.3.

□

5.3 Binary Classification Problems

Binary classification is a prediction problem with $T = \{-1, 1\}$ and $\ell(y, u) := I(y \neq u),\ \ y, u \in \{-1, 1\}$ (binary loss). It is a simple example of risk minimization with a nonconvex loss function.

Measurable functions $g : S \mapsto \{-1, 1\}$ are called classifiers. The risk of a classifier g with respect to the binary loss

$$L(g) := P(\ell \bullet g) = \mathbb{E}I(Y \neq g(X)) = \mathbb{P}\{Y \neq g(X)\}$$

is called the generalization error. It is well known that the minimum of the generalization error over the set of all classifiers is attained at the classifier

$$g_*(x) = \mathrm{sign}(\eta(x)),$$

where $\eta(x) = \mathbb{E}(Y|X = x)$ is the regression function. The function g_* is called the Bayes classifier. It is also well known that for all classifiers g

$$L(g) - L(g_*) = \int_{\{x: g(x) \neq g_*(x)\}} |\eta(x)| \Pi(dx) \tag{5.5}$$

(see, e.g., [33]).

Suppose there exists $h \in (0, 1]$ such that for all $x \in S$

$$|\eta(x)| \geq h. \tag{5.6}$$

The parameter h characterizes the level of noise in classification problems: for small values of h, $\eta(x)$ can get close to 0 and, in such cases, correct classification is harder to achieve. The following condition provides a more flexible way to describe the level of the noise:

$$\Pi\{x : |\eta(x)| \leq t\} \leq Ct^{\alpha} \tag{5.7}$$

for some $\alpha > 0$. It is often referred to as "Tsybakov's low noise assumption" or "Tsybakov's margin assumption". Classification problems under condition (5.7) have been intensively studied by Mammen and Tsybakov [72] and, especially, by Tsybakov [91]. Condition (5.6) was later suggested by Massart and used in a number of papers (see, e.g., [71]).

Lemma 5.2 *Under condition (5.6),*

$$L(g) - L(g_*) \geq h\Pi(\{x : g(x) \neq g_*(x)\}).$$

Under condition (5.7),

$$L(g) - L(g_*) \geq c\Pi^{\kappa}(\{x : g(x) \neq g_*(x)\}),$$

where $\kappa = \frac{1+\alpha}{\alpha}$ and $c > 0$ is a constant.

Proof. The first bound follows immediately from formula (5.5). To prove the second bound, use the same formula to get

$$L(g) - L(g_*) \geq t\Pi\Bigl\{x : g(x) \neq g_*(x), |\eta(x)| > t\Bigr\} \geq$$

$$t\Pi\Bigl\{x : g(x) \neq g_*(x)\Bigr\} - t\Pi\{x : |\eta(x)| \leq t\} \geq t\Pi\Bigl\{x : g(x) \neq g_*(x)\Bigr\} - Ct^{1+\alpha}.$$

It remains to choose in the last bound t that solves the equation

$$\Pi\Bigl\{x : g(x) \neq g_*(x)\Bigr\} = 2Ct^{\alpha}$$

to get the result.

□

Let $\mathcal{G}$ be a class of binary classifiers. Denote

$$\hat{g} := \mathrm{argmin}_{g\in\mathcal{G}} n^{-1}\sum_{j=1}^{n} I(Y_j \neq g(X_j))$$

a classifier in $\mathcal{G}$ that minimizes the empirical risk with respect to the binary loss (the training error).

First we obtain upper bounds on the excess risk $L(\hat{g})-L(g_*)$ of $\hat{g}$ in terms of random shattering numbers

$$\Delta^{\mathcal{G}}(X_1, \dots, X_n) := \mathrm{card}\Bigl\{(g(X_1), \dots, g(X_n)) : g \in \mathcal{G}\Bigr\}$$

and parameter h involved in condition (5.6).

Theorem 5.5 *Suppose condition (5.6) holds with some $h \in (0, 1]$. If $g_* \in \mathcal{G}$, then*

$$\mathbb{P}\biggl\{L(\hat{g}) - L(g_*) \geq K\biggl(\frac{\mathbb{E}\log\Delta^{\mathcal{G}}(X_1, \dots, X_n)}{nh} + \frac{t}{nh}\biggr)\biggr\} \leq Ce^{-t}$$

with some constants $K, C > 0$. In the general case, when g_* does not necessarily belong to $\mathcal{G}$, the following bound holds for all $\varepsilon \in (0, 1)$:

$$\mathbb{P}\biggl\{L(\hat{g})-L(g_*) \geq (1+\varepsilon)\Bigl(\inf_{g\in\mathcal{G}} L(g)-L(g_*)\Bigr)+K\biggl(\frac{\mathbb{E}\log\Delta^{\mathcal{G}}(X_1, \dots, X_n)}{nh\varepsilon^2}+\frac{t}{nh\varepsilon}\biggr)\biggr\} \leq Ce^{-t}$$

Proof. Note that

$$|(\ell \bullet g)(x,y) - (\ell \bullet g_*)(x,y)| = I(g(x) \neq g_*(x)),$$

which implies

$$\Big\|\ell \bullet g - \ell \bullet g_*\Big\|_{L_2(P)}^2 = P|(\ell \bullet g) - (\ell \bullet g_*)|^2 = \Pi\{x : g(x) \neq g_*(x)\}.$$

As always, denote $\mathcal{F} := \{\ell \bullet g : g \in \mathcal{G}\}$. Under the assumption $g_* \in \mathcal{G}$, the first inequality of Lemma 5.2 implies that

$$\mathcal{F}(\delta) = \Big\{\ell \bullet g : \mathcal{E}(\ell \bullet g) = L(g) - L(g_*) \leq \delta\Big\} \subset \Big\{\ell \bullet g : \Big\|\ell \bullet g - \ell \bullet g_*\Big\|_{L_2(P)} \leq \sqrt{\frac{\delta}{h}}\Big\},$$

so the $L_2(P)$-diameter $D(\delta)$ of the class $\mathcal{F}(\delta)$ satisfies $D(\delta) \leq 2\sqrt{\frac{\delta}{h}}$. Next we have

$$\phi_n(\delta) = \mathbb{E}\|P_n - P\|_{\mathcal{F}'(\delta)} \leq 2\mathbb{E} \sup_{g\in\mathcal{G}, \Pi(\{g \neq g_*\}) \leq \delta/h} |(P_n - P)(\ell \bullet g - \ell \bullet g_*)|.$$

Denote

$$\mathcal{D} := \Big\{\{(x,y) : y \neq g(x)\} : g \in \mathcal{G}\Big\} \text{ and } D_* := \{(x,y) : y \neq g_*(x)\}.$$

It is easy to check that for

$$D_1 := \{(x,y) : y \neq g_1(x)\}, \quad D_2 := \{(x,y) : y \neq g_2(x)\},$$

we have

$$\Pi(\{g_1 \neq g_2\}) = P(D_1 \triangle D_2).$$

From the last bound on $\phi_n(\delta)$, one can obtain that

$$\phi_n(\delta) \leq 2\mathbb{E} \sup_{D\in\mathcal{D}, P(D\triangle D_*) \leq \delta/h} |(P_n - P)(D \setminus D_*)| + 2\mathbb{E} \sup_{D\in\mathcal{D}, P(D\triangle D_*) \leq \delta/h} |(P_n - P)(D_* \setminus D)|.$$

Theorem 3.8 yields

$$\phi_n(\delta) \leq K\left[\sqrt{\frac{\delta}{h}}\sqrt{\frac{\mathbb{E}\log \Delta^{\mathcal{D}}((X_1,Y_1),\ldots,(X_n,Y_n))}{n}} \bigvee \frac{\mathbb{E}\log \Delta^{\mathcal{D}}((X_1,Y_1),\ldots,(X_n,Y_n))}{n}\right]$$

with some constant $K > 0$. Also, it is easy to observe that

$$\Delta^{\mathcal{D}}((X_1,Y_1),\ldots,(X_n,Y_n)) = \Delta^{\mathcal{G}}(X_1,\ldots,X_n)$$

which gives the bound

$$\phi_n(\delta) \le K\left[\sqrt{\frac{\delta}{h}}\sqrt{\frac{\mathbb{E}\log\Delta^{\mathcal{G}}(X_1,\dots,X_n)}{n}} \bigvee \frac{\mathbb{E}\log\Delta^{\mathcal{G}}(X_1,\dots,X_n)}{n}\right].$$

The bounds on $\phi_n(\delta)$ and $D(\delta)$ provide a way to control the quantity σ_n^t involved in Theorem 4.3:

$$\sigma_n^t \le K\left[\frac{\mathbb{E}\log\Delta^{\mathcal{G}}(X_1,\dots,X_n)}{nh} + \frac{t}{nh}\right]$$

with some constant $K > 0$, which implies the first bound of the theorem.

The proof of the second bound follows the same lines and it is based on Lemma 4.1.

□

The next theorem provides bounds on excess risk in terms of shattering numbers under Tsybakov's condition (5.7). We skip the proof which is similar.

Theorem 5.6 *Suppose condition (5.7) holds with some $\alpha > 0$. Let $\kappa := \frac{1+\alpha}{\alpha}$. If $g_* \in \mathcal{G}$, then*

$$\mathbb{P}\left\{L(\hat{g}) - L(g_*) \ge K\left(\left(\frac{\mathbb{E}\log\Delta^{\mathcal{G}}(X_1,\dots,X_n)}{n}\right)^{\kappa/(2\kappa-1)} + \left(\frac{t}{n}\right)^{\kappa/(2\kappa-1)}\right)\right\} \le Ce^{-t}$$

with some constants $K, C > 0$.

We will also mention the following result in spirit of Tsybakov [91].

Theorem 5.7 *Suppose, for some $A > 0, \rho \in (0,1)$*

$$\log N(\mathcal{G}; L_2(P_n); \varepsilon) \le \left(\frac{A}{\varepsilon}\right)^{2\rho} \tag{5.8}$$

and condition (5.7) holds with some $\alpha > 0$. Let $\kappa := \frac{1+\alpha}{\alpha}$. If $g_* \in \mathcal{G}$, then

$$\mathbb{P}\left\{L(\hat{g}) - L(g_*) \ge K\left(\left(\frac{1}{n}\right)^{\kappa/(2\kappa+\rho-1)} + \left(\frac{t}{n}\right)^{\kappa/(2\kappa-1)}\right)\right\} \le Ce^{-t}$$

with some constant $K, C > 0$ depending on A.

The proof is very similar to the proofs of the previous results except that now (3.15) is used to bound the empirical process. One can also use other notions of entropy such as entropy with bracketing and obtain very similar results.

We conclude this section with a theorem by Giné and Koltchinskii [50] that refines an earlier result by Massart and Nedelec [71]. To formulate it, let

$$\mathcal{C} := \Big\{\{g = 1\} : g \in \mathcal{G}\Big\}, \quad C_* := \{g_* = 1\},$$

and define the following local version of Alexander's capacity function of the class $\mathcal{C}$ (see [2]):

$$\tau(\delta) := \frac{\Pi\Big(\bigcup_{C\in\mathcal{C},\Pi(C\triangle C_*)\le\delta}(C\triangle C_*)\Big)}{\delta}.$$

Theorem 5.8 *Suppose condition (5.6) holds with some $h \in (0,1]$. Suppose also that $\mathcal{C}$ is a VC-class of VC-dimension V. If $g_* \in \mathcal{G}$, then*

$$\mathbb{P}\Big\{L(\hat{g}) - L(g_*) \ge K\Big(\frac{V}{nh}\log\tau\Big(\frac{V}{nh^2}\Big) + \frac{t}{nh}\Big)\Big\} \le Ce^{-t}$$

with some constants $K, C > 0$. In the general case, when g_* does not necessarily belong to $\mathcal{G}$, the following bound holds for all $\varepsilon \in (0,1)$:

$$\mathbb{P}\Big\{L(\hat{g}) - L(g_*) \ge (1+\varepsilon)\Big(\inf_{g\in\mathcal{G}} L(g) - L(g_*)\Big) + K\Big(\frac{V}{nh\varepsilon^2}\log\tau\Big(\frac{V}{nh^2\varepsilon^2}\Big) + \frac{t}{nh\varepsilon}\Big)\Big\} \le Ce^{-t}.$$

Proof (sketch). The proof relies on bound (3.13). For instance, to prove the second inequality this bound is used to control

$$\omega_n(\delta) = \mathbb{E}\sup_{g\in\mathcal{G},\|\ell\bullet g-\ell\bullet\bar{g}\|^2_{L_2(P)}\le\delta}|(P_n - P)(\ell\bullet g - \ell\bullet\bar{g})|,$$

where $\bar{g}$ is a minimal point of $P(\ell\bullet g)$ on $\mathcal{G}$. To use (3.13) one has to find the envelope

$$F_\delta(x,y) := \sup_{g\in\mathcal{G},\|\ell\bullet g-\ell\bullet\bar{g}\|^2_{L_2(P)}\le\delta}|\ell\bullet g(x,y) - \ell\bullet\bar{g}(x,y)|.$$

Easy computations show that

$$\|F_\delta\|_{L_2(\Pi)} = 2\sqrt{\delta\tau(\delta)}$$

and an application of (3.13) yields

$$\omega_n(\delta) \le K\Big[\sqrt{\frac{V\delta}{n}\log\tau(\delta)}\bigvee\frac{V}{n}\log\tau(\delta)\Big]$$

with some constant K. This implies that, for all $\varepsilon \in (0,1)$,

$$\omega_n^{\sharp}(\varepsilon) \leq K\frac{V}{n\varepsilon^2}\log\tau\left(\frac{V}{n\varepsilon^2}\right)$$

with some constant $K > 0$. Now we can use Lemma 4.1 to complete the proof of the second bound of the theorem (condition (4.4) of this lemma holds with $D = \frac{1}{h}$).

□

A straightforward upper bound on the capacity function $\tau(\delta) \leq \frac{1}{\delta}$ leads to the result of Massart and Nedelec [71] in which the main part of the error term is $\frac{V}{nh}\log\left(\frac{nh^2}{V}\right)$. However, it is easy to find examples in which the capacity $\tau(\delta)$ is uniformly bounded. For instance, suppose that $S = [0,1]^d$, Π is the Lebesgue measure on S, $\mathcal{C}$ is a VC-class of convex sets, $C_* \in \mathcal{C}$ and $\Pi(C_*) > 0$. Suppose also that with some constant $L > 0$

$$L^{-1}h(C, C_*) \leq \Pi(C \triangle C_*) \leq Lh(C, C_*), C \in \mathcal{C},$$

where h is Hausdorff distance. Then the boundedness of τ easily follows. In such cases, the main part of the error is of the order $\frac{V}{nh}$ (without a logarithmic factor).

6 Penalized Empirical Risk Minimization and Model Selection Problems

Let $\mathcal{F}$ be a class of measurable functions on $(S, \mathcal{A})$ and let $\{\mathcal{F}_k : k \geq 1\}$ be a family of its subclasses $\mathcal{F}_k \subset \mathcal{F}, k \geq 1$. The subclasses $\mathcal{F}_k$ will be used to approximate a solution of the problem of risk minimization (1.1) over a large class $\mathcal{F}$ by a family of solutions of "smaller" empirical risk minimization problems

$$\hat{f}_k := \hat{f}_{n,k} := \mathrm{argmin}_{f \in \mathcal{F}_k} P_n f.$$

For simplicity, we assume that the solutions $\{\hat{f}_{n,k}\}$ exist.

In what follows, we call $\mathcal{E}_P(\mathcal{F}; f) = Pf - \inf_{f \in \mathcal{F}} Pf$ *the global excess risk* of $f \in \mathcal{F}$. Given $k \geq 1$, we call $\mathcal{E}_P(\mathcal{F}; f) = Pf - \inf_{f \in \mathcal{F}} Pf$ *the local excess risk* of $f \in \mathcal{F}_k$.

Usually, the classes $\mathcal{F}_k, k \geq 1$ represent losses associated with different statistical models and the problem is to use the estimators $\{\hat{f}_{n,k}\}$ to construct a function $\hat{f} \in \mathcal{F}$ (for instance, to choose one of the estimators $\hat{f}_{n,k}$) with a small value of the global excess risk $\mathcal{E}_P(\mathcal{F}; \hat{f})$. To be more precise, suppose that there exists an index $k(P)$ such that $\inf_{\mathcal{F}_{k(P)}} Pf = \inf_{\mathcal{F}} Pf$. In other words, the risk minimizer over the whole class $\mathcal{F}$

belongs to a subclass $\mathcal{F}_{k(P)}$. A statistician does not know the distribution P and, hence, the index $k(P)$ of the correct model. Let $\tilde{\delta}_n(k)$ be an upper bound on the local excess risk $\mathcal{E}_P(\mathcal{F}_k; \hat{f}_{n,k})$ of $\hat{f}_{n,k}$ that provides an "optimal", or just a "desirable" accuracy of solution of empirical risk minimization problem on the class $\mathcal{F}_k$. If there were an oracle who could tell the statistician that $k(P) = 100$ is the correct index of the model, then the risk minimization problem could be solved with an accuracy at least $\tilde{\delta}_n(100)$. The *model selection problem* deals with constructing a data dependent index $\hat{k} = \hat{k}(X_1, \dots, X_n)$ of the model such that the excess risk of $\hat{f} := \hat{f}_{n,\hat{k}}$ is within a constant from $\tilde{\delta}_n(k(P))$ with a high probability. More generally, in the case when the global minimum of the risk $Pf, f \in \mathcal{F}$ is not attained in any of the classes $\mathcal{F}_k$, one can still try to show that with a high probability

$$\mathcal{E}_P(\mathcal{F}; \hat{f}) \leq C \inf_k \left[\inf_{\mathcal{F}_k} Pf - Pf_* + \tilde{\pi}_n(k) \right],$$

where

$$f_* := \mathrm{argmin}_{f \in \mathcal{F}} Pf.$$

For simplicity, assume the existence of a function $f_* \in \mathcal{F}$ at which the global minimum of the risk $Pf, f \in \mathcal{F}$ is attained. The quantities $\tilde{\pi}_n(k)$ involved in the above bound are "ideal" distribution dependent complexity penalties associated with risk minimization over $\mathcal{F}_k$ and C is a constant (preferably, $C = 1$ or at least close to 1). The inequalities that express such a property are often called *oracle inequalities.*

Among the most popular approaches to model selection are *penalization methods,* in which $\hat{k}$ is defined as a solution of the following minimization problem

$$\hat{k} := \mathrm{argmin}_{k \geq 1} \left\{ P_n \hat{f}_k + \hat{\pi}_n(k) \right\} \tag{6.1}$$

where $\hat{\pi}_n(k)$ is a *complexity penalty* (generally, data dependent) associated with the class (the model) $\mathcal{F}_k$. In other words, instead of minimizing the empirical risk on the whole class $\mathcal{F}$ we now minimize a penalized empirical risk.

We discuss below penalization strategies with the penalties based on data dependent bounds on excess risk developed in the previous sections. Penalization methods have been widely used in a variety of statistical problems, in particular, in nonparametric regression. At the same time, there are difficulties in extending penalization method of model selection to some other problems, such as nonparametric classification.

To provide some motivation for the approach discussed below, note that ideally one would want to find $\hat{k}$ by minimizing the global excess risk $\mathcal{E}_P(\mathcal{F}; \hat{f}_{n,k})$ of the solutions of ERM problems with respect to k. This is impossible without the help of the oracle.

Instead, data dependent upper confidence bounds on the excess risk have to be developed. The following trivial representation (that plays the role of "bias-variance decomposition")

$$\mathcal{E}_P(\mathcal{F};\hat{f}_{n,k}) = \inf_{\mathcal{F}_k} Pf - Pf_* + \mathcal{E}_P(\mathcal{F}_k;\hat{f}_{n,k})$$

shows that a part of the problem is to come up with data dependent upper bounds on the local excess risk $\mathcal{E}_P(\mathcal{F}_k;\hat{f}_{n,k})$. This was precisely the question studied in the previous sections. Another part of the problem is to bound $\inf_{\mathcal{F}_k} Pf - Pf_*$ in terms of $\inf_{\mathcal{F}_k} P_n f - P_n f_*$, which is what will be done in Lemma 6.3 below. Combining these two bounds provides an upper bound on the global excess risk that can be now minimized with respect to k (the term $P_n f_*$ can be dropped since it does not depend on k).

Suppose that for each class $\mathcal{F}_k$, the function $U_n(\cdot) = U_{n,k}(\cdot)$ is given (it was defined in Section 4.1 in terms of sequences $\{\delta_j\}$ $\{t_j\}$ that, in this case, might also depend on k). In what follows, we will assume that, for each $k \geq 1$, $(\bar{\delta}_n(k), \hat{\delta}_n(k), \tilde{\delta}_n(k))$ is a triple bound on the excess risk for the class $\mathcal{F}_k$ of confidence level $1 - p_k$ (see Definition 4.1). Suppose $p := \sum_{k=1}^{\infty} p_k < 1$. Then, there exists an event E of probability at least $1 - p$ such that on this event the following properties hold for all $k \geq 1$:

(i) $U^{\sharp}_{n,k}\Big(\frac{1}{2}\Big) \leq \bar{\delta}_n(k) \leq \hat{\delta}_n(k) \leq \tilde{\delta}_n(k)$;

(ii) $\mathcal{E}(\mathcal{F}_k, \hat{f}_{n,k}) \leq \bar{\delta}_n(k)$;

(iii) for all $f \in \mathcal{F}_k$,

$$\mathcal{E}_P(\mathcal{F}_k, f) \leq 2\mathcal{E}_{P_n}(\mathcal{F}_k; f) \vee \bar{\delta}_n(k)$$

and

$$\mathcal{E}_{P_n}(\mathcal{F}_k; f) \leq \frac{3}{2}\Big(\mathcal{E}_P(\mathcal{F}_k; f) \vee \bar{\delta}_n(k)\Big);$$

(iv) for all $\delta \geq \bar{\delta}_n(k)$,

$$\|P_n - P\|_{\mathcal{F}'_k(\delta)} \leq U_{n,k}(\delta).$$

In the next sections, we study several special cases of general penalized empirical risk minimization problem in which it will be possible to prove oracle inequalities.

6.1 Penalization in Monotone Families $\mathcal{F}_k$

In this section, we make a simplifying assumption that $\{\mathcal{F}_k\}$ is a monotone family, i.e., $\mathcal{F}_k \subset \mathcal{F}_{k+1},\ k \geq 1$. Let

$$\mathcal{F} := \bigcup_{j\geq 1} \mathcal{F}_j.$$

Define

$$\hat{k} := \mathrm{argmin}_{k\geq 1}\Bigl[\inf_{f\in\mathcal{F}_k} P_n f + 4\hat{\delta}_n(k)\Bigr]$$

and $\hat{f} := \hat{f}_{\hat{k}}$.

The next statement is akin to the result of Bartlett [6].

Theorem 6.1 *The following oracle inequality holds with probability at least* $1-p$:

$$\mathcal{E}_P(\mathcal{F};\hat{f}) \leq \inf_{j\geq 1}\Bigl[\inf_{\mathcal{F}_j} Pf - \inf_{\mathcal{F}} Pf + 9\tilde{\delta}_n(j)\Bigr].$$

Proof. We will consider the event E of probability at least $1-p$ on which properties (i)–(iv) hold. Then, for all $j \geq \hat{k}$,

$$\mathcal{E}_P(\mathcal{F}_j;\hat{f}) \leq 2\mathcal{E}_{P_n}(\mathcal{F}_j;\hat{f}) \vee \bar{\delta}_n(j) \leq 2\Bigl[\inf_{f\in\mathcal{F}_{\hat{k}}} P_n f - \inf_{f\in\mathcal{F}_j} P_n f\Bigr] + \bar{\delta}_n(j) \leq$$

$$2\Bigl[\inf_{f\in\mathcal{F}_{\hat{k}}} P_n f + 4\hat{\delta}_n(\hat{k}) - \inf_{f\in\mathcal{F}_j} P_n f - 4\hat{\delta}_n(j)\Bigr] + 9\hat{\delta}_n(j),$$

which is bounded by $9\tilde{\delta}_n(j)$ since, by the definition of $\hat{k}$, the term in the bracket is nonpositive and $\hat{\delta}_n(j) \leq \tilde{\delta}_n(j)$. This implies

$$P\hat{f} \leq \inf_{f\in\mathcal{F}_j} Pf + 9\tilde{\delta}_n(j).$$

The next case is when $j < \hat{k}$ and $\hat{\delta}_n(j) \geq \hat{\delta}_n(\hat{k})/9$. Then $\mathcal{E}_P(\mathcal{F}_{\hat{k}};\hat{f}_{\hat{k}}) \leq \bar{\delta}_n(\hat{k})$, and, as a consequence,

$$P\hat{f} \leq \inf_{f\in\mathcal{F}_{\hat{k}}} Pf + \bar{\delta}_n(\hat{k}) \leq \inf_{f\in\mathcal{F}_j} Pf + 9\tilde{\delta}_n(j).$$

The last case to consider is when $j < \hat{k}$ and $\hat{\delta}_n(j) < \hat{\delta}_n(\hat{k})/9$. In this case, the definition of $\hat{k}$ implies that

$$\inf_{f\in\mathcal{F}_j} \mathcal{E}_{P_n}(\mathcal{F}_{\hat{k}};f) = \inf_{f\in\mathcal{F}_j} P_n f - \inf_{f\in\mathcal{F}_{\hat{k}}} P_n f \geq 4(\hat{\delta}_n(\hat{k}) - \hat{\delta}_n(j)) \geq 3\hat{\delta}_n(\hat{k}).$$

Hence,

$$\frac{3}{2}\biggl(\inf_{f\in\mathcal{F}_j} \mathcal{E}_P(\mathcal{F}_{\hat{k}};f) \vee \bar{\delta}_n(\hat{k})\biggr) \geq \inf_{f\in\mathcal{F}_j} \mathcal{E}_{P_n}(\mathcal{F}_{\hat{k}};f) \geq 3\hat{\delta}_n(\hat{k}),$$

which yields

$$3\inf_{f\in\mathcal{F}_j} \mathcal{E}_P(\mathcal{F}_{\hat{k}};f) + 3\bar{\delta}_n(\hat{k}) \geq 6\hat{\delta}_n(\hat{k}).$$

Therefore

$$\inf_{f\in\mathcal{F}_j} \mathcal{E}_P(\mathcal{F}_{\hat{k}};f) \geq \hat{\delta}_n(\hat{k}) \geq \mathcal{E}_P(\mathcal{F}_{\hat{k}};\hat{f}).$$

As a consequence,

$$P\hat{f} \leq \inf_{f \in \mathcal{F}_j} Pf \leq \inf_{f \in \mathcal{F}_j} Pf + 9\tilde{\delta}_n(j).$$

This completes the proof.

$\square$

Example. Consider a regression problem with quadratic loss and with a bounded response variable $Y \in [0,1]$ (see Section 5.1). Let $\mathcal{G}_k,\ k \geq 1$ be convex classes of functions g taking values in $[0,1]$ such that $\mathcal{G}_k \subset \mathcal{G}_{k+1},\ k \geq 1$. Moreover, suppose that for all $k \geq 1$ $\mathcal{G}_k \subset L_k$, where L_k is a finite dimensional space of dimension d_k. Let

$$\hat{g}_{n,k} := \mathrm{argmin}_{g \in \mathcal{G}_k} n^{-1} \sum_{j=1}^{n} (Y_j - g(X_j))^2.$$

Take a nondecreasing sequence $\{t_k\}$ of positive numbers such that

$$\sum_{k \geq 1} e^{-t_k} = p \in (0,1).$$

Define

$$\bar{\delta}_n(k) = \hat{\delta}_n(k) = \tilde{\delta}_n(k) = K\biggl(\frac{d_k}{n} + \frac{t_k}{n}\biggr),\ k \geq 1.$$

It is straightforward to see that, for a large enough constant K, $(\bar{\delta}_n(k), \hat{\delta}_n(k), \tilde{\delta}_n(k))$ is a triple bound of level $1 - e^{-t_k}$ (see Example 1, Section 5.1). Hence, if we define

$$\hat{k} := \mathrm{argmin}_{k \geq 1} \biggl[\inf_{g \in \mathcal{G}_k} n^{-1} \sum_{j=1}^{n} (Y_j - g(X_j))^2 + 4K\biggl(\frac{d_k}{n} + \frac{t_k}{n}\biggr)\biggr]$$

with a sufficiently large constant K and set $\hat{g} := \hat{g}_{n,\hat{k}}$, then it follows from Theorem 6.1 that with probability at least $1 - p$

$$\|\hat{g} - g_*\|^2_{L_2(\Pi)} \leq \inf_{k \geq 1} \biggl[\inf_{g \in \mathcal{G}_k} \|g - g_*\|^2_{L_2(\Pi)} + 9K\biggl(\frac{d_k}{n} + \frac{t_k}{n}\biggr)\biggr].$$

Clearly, one can also construct triple bounds and implement this penalization method in more complicated situations (see examples 2-5 in Section 5.1) and for other loss functions (for instance, for convex losses discussed in Section 5.2). Moreover, one can use a general construction of triple bounds in Theorem 4.8 that provides a universal approach to complexity penalization (which, however, is more of theoretical interest).

Despite the fact that it is possible to prove nice and simple oracle inequalities under the monotonicity assumption, this assumption might be restrictive and, in what follows, we explore what can be done without it.

6.2 Penalization by Empirical Risk Minima

In this section, we study a simple penalization technique in spirit of the work of Lugosi and Wegkamp [70] in which the infimum of empirical risk $\inf_{\mathcal{F}_k} P_n f$ is explicitly involved in the penalty. It will be possible to prove rather natural oracle inequalities for this penalization method. However, the drawback of this approach is that, in most of the cases, it yields only suboptimal convergence rates.

Given triple bounds $(\bar{\delta}_n(k), \hat{\delta}_n(k), \tilde{\delta}_n(k))$ of level $1 - p_k$ for classes $\mathcal{F}_k$, define the following penalties:

$$\hat{\pi}(k) := \hat{\pi}_n(k) := \hat{K}\left[\hat{\delta}_n(k) + \sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} P_n f} + \frac{t_k}{n}\right]$$

and

$$\tilde{\pi}(k) := \tilde{\pi}_n(k) := \tilde{K}\left[\tilde{\delta}_n(k) + \sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} P f} + \frac{t_k}{n}\right],$$

where $\hat{K}, \tilde{K}$ are sufficiently large numerical constants. Here $\tilde{\pi}(k)$ represents a "desirable accuracy" of risk minimization on the class $\mathcal{F}_k$.

The index estimate $\hat{k}$ is defined by minimizing the penalized empirical risk

$$\hat{k} := \mathrm{argmin}_{k\geq 1}\Big\{P_n \hat{f}_k + \hat{\pi}(k)\Big\}$$

and, as always, $\hat{f} := \hat{f}_{\hat{k}}$.

The next theorem provides an upper confidence bound on the risk of $\hat{f}$ and an oracle inequality for the global excess risk $\mathcal{E}_P(\mathcal{F}; \hat{f})$.

Theorem 6.2 *There exists a choice of $\hat{K}, \tilde{K}$ such that for any sequence $\{t_k\}$ of positive numbers, the following bounds hold:*

$$\mathbb{P}\Big\{P\hat{f} \geq \inf_{k\geq 1}\Big\{P_n \hat{f}_{n,k} + \hat{\pi}(k)\Big\}\Big\} \leq \sum_{k=1}^{\infty}\Big(p_k + e^{-t_k}\Big)$$

and

$$\mathbb{P}\Big\{\mathcal{E}_P(\mathcal{F}; \hat{f}) \geq \inf_{k\geq 1}\Big\{\inf_{f\in\mathcal{F}_k} Pf - \inf_{f\in\mathcal{F}} Pf + \tilde{\pi}(k)\Big\}\Big\} \leq \sum_{k=1}^{\infty}\Big(p_k + e^{-t_k}\Big).$$

Unless $\inf_{\mathcal{F}_k} Pf = 0$, $\tilde{\pi}(k) = \tilde{\pi}_n(k)$ can not be smaller than const $n^{-1/2}$. In many cases (see Section 5), the excess risk bound $\tilde{\delta}_n(k)$ is smaller than this, and the penalization method of this section is suboptimal.

The following lemma is the main tool used in the proof.

Lemma 6.1 *Let $\mathcal{F}$ be a class measurable functions from S into $[0,1]$. If $\bar{\delta}_n$ is an admissible distribution dependent bound of confidence level $1-p,\ p \in (0,1)$ (see Definition 4.1), then the following inequality holds for all $t > 0$:*

$$\mathbb{P}\biggl\{\Bigl|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf\Bigr| \geq 2\bar{\delta}_n + \sqrt{\frac{2t}{n}\inf_{\mathcal{F}} Pf} + \frac{t}{n}\biggr\} \leq p + e^{-t}.$$

If $(\bar{\delta}_n, \hat{\delta}_n, \tilde{\delta}_n)$ is a triple bound of confidence level $1-p$, then

$$\mathbb{P}\biggl\{\Bigl|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf\Bigr| \geq 4\hat{\delta}_n + 2\sqrt{\frac{2t}{n}\inf_{\mathcal{F}} P_n f} + \frac{8t}{n}\biggr\} \leq p + e^{-t}.$$

Proof. Let E be the event where conditions (i)-(iv) of Definition 4.1 hold. Then $\mathbb{P}(E) \geq 1-p$. On the event E, $\mathcal{E}(\hat{f}_n) \leq \bar{\delta}_n$ and, for all $\varepsilon < \bar{\delta}_n$ and $g \in \mathcal{F}(\varepsilon)$

$$\begin{aligned}&\Bigl|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf\Bigr| = \Bigl|P_n \hat{f}_n - \inf_{\mathcal{F}} Pf\Bigr| \leq \\ &P\hat{f}_n - \inf_{\mathcal{F}} Pf + |(P_n - P)(\hat{f}_n - g)| + |(P_n - P)(g)| \leq \\ &\leq \bar{\delta}_n + \|P_n - P\|_{\mathcal{F}'(\bar{\delta}_n)} + |(P_n - P)(g)|.\end{aligned} \tag{6.2}$$

Also, on the same event E,

$$\|P_n - P\|_{\mathcal{F}'(\bar{\delta}_n)} \leq U_n(\bar{\delta}_n(t)) \leq \bar{V}_n(\bar{\delta}_n)\bar{\delta}_n \leq \bar{\delta}_n. \tag{6.3}$$

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

$$|(P_n - P)(g)| \leq \sqrt{2\frac{t}{n}\mathrm{Var}_P g} + \frac{2t}{3n} \leq \sqrt{2\frac{t}{n}\Bigl(\inf_{\mathcal{F}} Pf + \varepsilon\Bigr)} + \frac{2t}{3n}, \tag{6.4}$$

since g takes values in $[0,1]$, $g \in \mathcal{F}(\varepsilon)$, and $\mathrm{Var}_P g \leq Pg^2 \leq Pg \leq \inf_{\mathcal{F}} Pf + \varepsilon$. It follows from (6.2), (6.3) and (6.4) that, on the event

$$E(\varepsilon) := E \bigcap \biggl\{|(P_n - P)(g)| \leq \sqrt{2\frac{t}{n}\Bigl(\inf_{\mathcal{F}} Pf + \varepsilon\Bigr)} + \frac{2t}{3n}\biggr\}, \tag{6.5}$$

the following inequality holds:

$$\Bigl|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf\Bigr| \leq 2\bar{\delta}_n + \sqrt{2\frac{t}{n}\Bigl(\inf_{\mathcal{F}} Pf + \varepsilon\Bigr)} + \frac{t}{n}. \tag{6.6}$$

Since the events $E(\varepsilon)$ are monotone in ε, let $\varepsilon \to 0$ to get

$$\mathbb{P}(E(0)) \geq 1 - p - e^{-t}.$$

This yields the first bound of the lemma.

For the proof of the second bound, note that on the event $E(0)$,

$$\left|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf\right| \le \sqrt{2\frac{t}{n}|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf| + 2\bar{\delta}_n} + \sqrt{2\frac{t}{n}\inf_{\mathcal{F}} P_n f} + \frac{t}{n}. \tag{6.7}$$

Thus, either

$$|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf| \le \frac{8t}{n}, \quad \text{or} \quad \frac{2t}{n} \le \frac{|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf|}{4}.$$

In the last case (6.7) implies that

$$\left|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf\right| \le 4\bar{\delta}_n + 2\sqrt{2\frac{t}{n}\inf_{\mathcal{F}} P_n f} + \frac{2t}{n}.$$

The condition of the lemma allows us to replace (on the event E) $\bar{\delta}_n$ by $\hat{\delta}_n$ and to get the following bound that holds with probability at least $1 - p - e^{-t}$:

$$\left|\inf_{\mathcal{F}} P_n f - \inf_{\mathcal{F}} Pf\right| \le 4\hat{\delta}_n + 2\sqrt{2\frac{t}{n}\inf_{\mathcal{F}} P_n f} + \frac{8t}{n}.$$

□

Proof of Theorem 6.2. For each class $\mathcal{F}_k$ and $t = t_k$ define the event $E_k(0)$, (with $\varepsilon = 0$) as in (6.5). Clearly,

$$\mathbb{P}(E_k(0)) \ge 1 - p_k - e^{-t_k}.$$

Let

$$F := \bigcap_{k\ge 1} E_k(0).$$

Then

$$\mathbb{P}(F^c) \le \sum_{k=1}^{\infty}\left(p_k + e^{-t_k}\right).$$

We use the following consequence of Lemma 6.1 and the definition of the triple bounds: on the event F for all $k \ge 1$,

$$P\hat{f}_k - \inf_{f\in\mathcal{F}_k} Pf \le \bar{\delta}_n(k) \le \hat{\delta}_n(k) \le \tilde{\delta}_n(k)$$

and

$$\left|\inf_{\mathcal{F}_k} P_n f - \inf_{\mathcal{F}_k} Pf\right| \le 2\bar{\delta}_n(k) + \sqrt{\frac{2t_k}{n}\inf_{\mathcal{F}_k} Pf} + \frac{t_k}{n},$$

$$\Bigl|\inf_{\mathcal{F}_k} P_n f - \inf_{\mathcal{F}_k} Pf\Bigr| \le 4\hat{\delta}_n(k) + 2\sqrt{\frac{2t_k}{n}\inf_{\mathcal{F}_k} P_n f} + \frac{8t_k}{n}.$$

Therefore,

$$Pf̂ = P\hat{f}_{\hat{k}} \le \inf_{\mathcal{F}_{\hat{k}}} Pf + \bar{\delta}_n(\hat{k}) \le \inf_{\mathcal{F}_{\hat{k}}} P_n f + 5\hat{\delta}_n(\hat{k}) + 2\sqrt{\frac{2t_{\hat{k}}}{n}\inf_{\mathcal{F}_{\hat{k}}} P_n f} + \frac{8t_{\hat{k}}}{n} \le$$

$$\le \inf_{\mathcal{F}_{\hat{k}}} P_n f + \hat{\pi}(\hat{k}) = \inf_k \Bigl[\inf_{\mathcal{F}_k} P_n f + \hat{\pi}(k)\Bigr],$$

provided that the constant $\hat{K}$ in the definition of $\hat{\pi}$ was chosen properly. The first bound of the theorem has been proved.

To prove the second bound, note that

$$\sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} P_n f} \le \sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} Pf} + \sqrt{\frac{t_k}{n}|\inf_{\mathcal{F}_k} P_n f - \inf_{\mathcal{F}_k} Pf|} \le$$

$$\sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} Pf} + \frac{t_k}{2n} + \frac{1}{2}|\inf_{\mathcal{F}_k} P_n f - \inf_{\mathcal{F}_k} Pf|.$$

Therefore, on the event F for all k

$$\hat{\pi}(k) = \hat{K}\biggl[\hat{\delta}_n(k) + \sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} P_n f} + \frac{t_k}{n}\biggr] \le \frac{\tilde{K}}{2}\biggl[\tilde{\delta}_n(k) + \sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} Pf} + \frac{t_k}{n}\biggr] = \tilde{\pi}(k)/2$$

and

$$\Bigl|\inf_{\mathcal{F}_k} P_n f - \inf_{\mathcal{F}_k} Pf\Bigr| \le 2\bar{\delta}_n(k) + \sqrt{\frac{2t_k}{n}\inf_{\mathcal{F}_k} Pf} + \frac{t_k}{n} \le \frac{\tilde{K}}{2}\biggl[\tilde{\delta}_n(k) + \sqrt{\frac{t_k}{n}\inf_{\mathcal{F}_k} Pf} + \frac{t_k}{n}\biggr] = \tilde{\pi}(k)/2,$$

provided that the constant $\tilde{K}$ in the definition of $\tilde{\pi}(k)$ is large enough. As a result, on the event F,

$$P\hat{f} \le \inf_k \Bigl[\inf_{\mathcal{F}_k} P_n f + \hat{\pi}(k)\Bigr] \le \inf_k \Bigl[\inf_{\mathcal{F}_k} Pf + \tilde{\pi}(k)\Bigr],$$

proving the second bound. □

Example. As an example, we derive some of the results of Lugosi and Wegkamp [70] (in a slightly modified form). Suppose that $\mathcal{F}$ is a class of measurable functions on S taking values in $\{0,1\}$ (binary functions). As before, let $\Delta^{\mathcal{F}}(X_1,\dots,X_n)$ be the shattering number of the class $\mathcal{F}$ on the sample $(X_1,\dots,X_n)$:

$$\Delta^{\mathcal{F}}(X_1,\dots,X_n) := \mathrm{card}\biggl(\Bigl\{(f(X_1),\dots,f(X_n)) : f \in \mathcal{F}\Bigr\}\biggr).$$

Given a sequence $\{\mathcal{F}_k\}$, $\mathcal{F}_k \subset \mathcal{F}$ of classes of binary functions, define the following complexity penalties

$$\hat{\pi}(k) := \hat{K}\left[\sqrt{\inf_{f\in\mathcal{F}_k} P_n f \frac{\log \Delta^{\mathcal{F}_k}(X_1,\dots,X_n) + t_k}{n}} + \frac{\log \Delta^{\mathcal{F}_k}(X_1,\dots,X_n) + t_k}{n}\right]$$

and

$$\tilde{\pi}(k) := \tilde{K}\left[\sqrt{\inf_{f\in\mathcal{F}_k} Pf \frac{\mathbb{E}\log \Delta^{\mathcal{F}_k}(X_1,\dots,X_n) + t_k}{n}} + \frac{\mathbb{E}\log \Delta^{\mathcal{F}_k}(X_1,\dots,X_n) + t_k}{n}\right],$$

and let $\hat{k}$ be a solution of the following penalized empirical risk minimization problem

$$\hat{k} := \mathrm{argmin}_{k\geq 1}\left[\min_{\mathcal{F}_k} P_n f + \hat{\pi}(k)\right].$$

Denote $\hat{f} := \hat{f}_{n,\hat{k}}$.

Theorem 6.3 *There exists a choice of $\hat{K}, \tilde{K}$ such that for all $t_k > 0$,*

$$\mathbb{P}\left\{\mathcal{E}_P(\mathcal{F};\hat{f}) \geq \inf_{k\geq 1}\left\{\inf_{f\in\mathcal{F}_k} Pf - \inf_{f\in\mathcal{F}} Pf + \tilde{\pi}(k)\right\}\right\} \leq \sum_{k=1}^{\infty} e^{-t_k}.$$

Note that penalization based on random shattering numbers is natural in classification problems and the result of Theorem 6.3 can be easily stated in classification setting. The result follows from Theorem 6.2 and the next lemma that provides a version of triple bound on excess risk for classes of binary functions.

Lemma 6.2 *Given a class of binary functions $\mathcal{F}$ and $t > 0$, define*

$$\bar{\delta}_n := \bar{K}\left[\sqrt{\inf_{f\in\mathcal{F}} Pf \frac{\mathbb{E}\log \Delta^{\mathcal{F}}(X_1,\dots,X_n) + t}{n}} + \frac{\mathbb{E}\log \Delta^{\mathcal{F}}(X_1,\dots,X_n) + t}{n}\right],$$

$$\hat{\delta}_n := \hat{K}\left[\sqrt{\inf_{f\in\mathcal{F}} P_n f \frac{\log \Delta^{\mathcal{F}}(X_1,\dots,X_n) + t}{n}} + \frac{\log \Delta^{\mathcal{F}}(X_1,\dots,X_n) + t}{n}\right]$$

and

$$\tilde{\delta}_n := \tilde{K}\left[\sqrt{\inf_{f\in\mathcal{F}} Pf \frac{\mathbb{E}\log \Delta^{\mathcal{F}}(X_1,\dots,X_n) + t}{n}} + \frac{\mathbb{E}\log \Delta^{\mathcal{F}}(X_1,\dots,X_n) + t}{n}\right].$$

There exists a choice of constants $\bar{K}, \hat{K}, \tilde{K}$ such that $(\bar{\delta}_n, \hat{\delta}_n, \tilde{\delta}_n)$ is a triple bound of level $1 - e^{-t}$ for the class $\mathcal{F}$.

Proof. The following upper bounds on the $L_2(P)$-diameter of the δ-minimal set $\mathcal{F}(\delta)$ and on the function $\phi_n(\delta)$ hold:

$$D^2(\mathcal{F};\delta) = \sup_{f,g\in\mathcal{F}(\delta)} P(f-g)^2 \le \sup_{f,g\in\mathcal{F}(\delta)} (Pf+Pg) \le 2(\inf_{f\in\mathcal{F}} Pf + \delta).$$

By Theorem 3.8,

$$\phi_n(\delta) \le K\left[\sqrt{2\Bigl(\inf_{f\in\mathcal{F}} Pf+\delta\Bigr)\frac{\mathbb{E}\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)}{n}} + \frac{\mathbb{E}\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)}{n}\right].$$

A straightforward computation implies the next bound on the quantity σ_n^t from Theorem 4.3:

$$\sigma_n^t \le \bar{\delta}_n = \bar{K}\left[\sqrt{\inf_{f\in\mathcal{F}} Pf\frac{\mathbb{E}\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)+t}{n}} + \frac{\mathbb{E}\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)+t}{n}\right],$$

provided that the constant $\bar{K}$ is large enough. Moreover, with a proper choice of this constant, $\bar{\delta}_n$ is an admissible bound of level $1-e^{-t}$.

The following deviation inequality for shattering numbers is due to Boucheron, Lugosi and Massart [20]: with probability at least $1-e^{-t}$

$$\log\Delta^{\mathcal{F}}(X_1,\dots,X_n) \le 2\mathbb{E}\log\Delta^{\mathcal{F}}(X_1,\dots,X_n) + 2t$$

and

$$\mathbb{E}\log\Delta^{\mathcal{F}}(X_1,\dots,X_n) \le 2\log\Delta^{\mathcal{F}}(X_1,\dots,X_n) + 2t.$$

Together with the first bound of Lemma 6.1, this easily implies that with probability at least $1-8e^{-t}$, $\bar{\delta}_n \le \hat{\delta}_n \le \tilde{\delta}_n$. First we prove that $\bar{\delta}_n \le \hat{\delta}_n$. To this end, we use the first bound of Lemma 6.1 and the inequality $2ab \le a^2+b^2$ to show that with probability at least $1-2e^{-t}$

$$\inf_{\mathcal{F}} Pf \le \inf_{\mathcal{F}} P_nf + 2\bar{\delta}_n + 2\sqrt{\frac{t}{2n}\inf_{\mathcal{F}} Pf} + \frac{t}{3n} \le \inf_{\mathcal{F}} P_nf + 2\bar{\delta}_n + \frac{\inf_{\mathcal{F}} Pf}{2} + \frac{2t}{n}.$$

Therefore,

$$\inf_{\mathcal{F}} Pf \le 2\inf_{\mathcal{F}} P_nf + 4\bar{\delta}_n + \frac{4t}{n}.$$

We substitute this inequality into the definition of $\bar{\delta}_n$ and replace $\mathbb{E}\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)$ by the upper bound $2\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)+2t$ that holds with probability at least $1-e^{-t}$. It follows that, with some constant K,

$$\bar{\delta}_n \le K\left[\sqrt{\inf_{f\in\mathcal{F}} P_nf\frac{\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)+t}{n}} + \frac{\log\Delta^{\mathcal{F}}(X_1,\dots,X_n)+t}{n}\right]+$$

$$+2\sqrt{\frac{\bar{\delta}_n}{2}\frac{K^2 \log \Delta^{\mathcal{F}}(X_1,\dots,X_n)+t}{2n}},$$

Again, using the inequality $2ab \le a^2 + b^2$, we get the following bound that holds with some constant $\hat{K}$ and with probability at least $1 - 4e^{-t}$:

$$\bar{\delta}_n \le \hat{K}\left[\sqrt{\inf_{f\in\mathcal{F}} P_n f \frac{\log \Delta^{\mathcal{F}}(X_1,\dots,X_n)+t}{n}} + \frac{\log \Delta^{\mathcal{F}}(X_1,\dots,X_n)+t}{n}\right] =: \hat{\delta}_n.$$

The proof of the second inequality $\hat{\delta}_n \le \tilde{\delta}_n$ is similar. By increasing the values of the constants $\bar{K}, \hat{K}, \tilde{K}$, it is easy to eliminate the factor 8 and to obtain a triple bound of level $1 - e^{-t}$, as it was claimed.

□

6.3 Linking Excess Risk and Variance in Penalization

In a variety of regression and classification problems the following assumption plays the crucial role: for all $f \in \mathcal{F}$,

$$Pf - Pf_* \ge \varphi\left(\sqrt{\mathrm{Var}_P(f - f_*)}\right), \tag{6.8}$$

where φ is a convex nondecreasing function on $[0, +\infty)$ with $\varphi(0) = 0$. In section 5, we have already dealt with several examples of this condition. For instance, in the case of regression with quadratic loss $\ell(y,u) = (y-u)^2$ and with bounded response $Y \in [0,1]$, condition (6.8) is satisfied for the loss class $\mathcal{F} = \{\ell \bullet g : g \in \mathcal{G}\}$, where $\mathcal{G}$ is a class of functions from S into $[0,1]$. In this case, one can take $\varphi(u) = u^2/2$, so the function φ does not depend on the unknown distribution P (except that the assumption $Y \in [0,1]$ is already a restriction on the class of distributions P). On the other hand, in classification problems, φ is related to the parameters of the noise such as parameter α in Tsybakov's low noise assumption (5.7) or parameter h in Massart's low noise assumption (5.6). So, in this case, φ does depend on P. The function φ describes the relationship between the excess risk $Pf - P_*$ and the variance $\mathrm{Var}_P(f - f_*)$ of the "excess loss" $f - f_*$. In what follows, we will call φ the *link function.* It happens that the link function is involved in a rather natural way in the construction of complexity penalties that provide optimal convergence rates in many problems. Since the link function is generally distribution dependent, the development of adaptive penalization methods of model selection is a challenge, for instance, in classification setting.

We will assume that with some $\gamma > 0$

$$\varphi(uv) \leq \gamma\varphi(u)\varphi(v), \ u, v \geq 0. \tag{6.9}$$

Denote

$$\varphi^*(v) := \sup_{u \geq 0}[uv - \varphi(u)]$$

the conjugate of φ. Then

$$uv \leq \varphi(u) + \varphi^*(v), \ u, v \geq 0.$$

Let $(\bar{\delta}_n(k), \hat{\delta}_n(k), \tilde{\delta}_n(k))$ be a triple bound of level $1 - p_k$ for the class $\mathcal{F}_k, k \geq 1$. For a fixed $\varepsilon > 0$, define the penalties as follows:

$$\hat{\pi}(k) := A(\varepsilon)\hat{\delta}_n(k) + \varphi^*\Bigl(\sqrt{\frac{2t_k}{\varepsilon n}}\Bigr) + \frac{t_k}{n}$$

and

$$\tilde{\pi}(k) := \frac{A(\varepsilon)}{1 + \gamma\varphi(\sqrt{\varepsilon})}\tilde{\delta}_n(k) + \frac{2}{1 + \gamma\varphi(\sqrt{\varepsilon})}\varphi^*\Bigl(\sqrt{\frac{2t_k}{\varepsilon n}}\Bigr) + \frac{2}{1 + \gamma\varphi(\sqrt{\varepsilon})}\frac{t_k}{n},$$

where

$$A(\varepsilon) := \frac{5}{2} - \gamma\varphi(\sqrt{\varepsilon}).$$

As before, $\hat{k}$ is defined by

$$\hat{k} := \text{argmin}_{k \geq 1}\Bigl\{P_n\hat{f}_k + \hat{\pi}(k)\Bigr\}$$

and $\hat{f} := \hat{f}_{n,\hat{k}}$.

Theorem 6.4 *For any sequence $\{t_k\}$ of positive numbers,*

$$\mathbb{P}\Bigl\{\mathcal{E}_P(\mathcal{F}; \hat{f}) \geq C(\varepsilon) \inf_{k \geq 1}\Bigl\{\inf_{f \in \mathcal{F}_k} Pf - \inf_{f \in \mathcal{F}} Pf + \tilde{\pi}(k)\Bigr\}\Bigr\} \leq \sum_{k=1}^{\infty}\Bigl(p_k + e^{-t_k}\Bigr),$$

where

$$C(\varepsilon) := \frac{1 + \gamma\varphi(\sqrt{\varepsilon})}{1 - \gamma\varphi(\sqrt{\varepsilon})}.$$

The following lemma is needed in the proof.

Lemma 6.3 *Let $\mathcal{G} \subset \mathcal{F}$ and let $(\bar{\delta}_n, \hat{\delta}_n, \tilde{\delta}_n)$ be a triple bound of level $1-p$ for the class $\mathcal{G}$. For all $t > 0$, there exists an event E with probability at least $1 - p - e^{-t}$ such that on this event*

$$\inf_{\mathcal{G}} P_n f - P_n f_* \leq (1 + \gamma\varphi(\sqrt{\varepsilon}))(\inf_{\mathcal{G}} Pf - Pf_*) + \varphi^*\Big(\sqrt{\frac{2t}{\varepsilon n}}\Big) + \frac{t}{n} \tag{6.10}$$

and

$$\inf_{\mathcal{G}} Pf - Pf_* \leq (1 - \gamma\varphi(\sqrt{\varepsilon}))^{-1}\Big[\inf_{\mathcal{G}} P_n f - P_n f_* + \frac{3}{2}\bar{\delta}_n + \varphi^*\Big(\sqrt{\frac{2t}{\varepsilon n}}\Big) + \frac{t}{n}\Big]. \tag{6.11}$$

In addition, if there exists $\bar{\delta}_n^{\varepsilon}$ such that

$$\bar{\delta}_n \leq \varepsilon(\inf_{\mathcal{G}} Pf - Pf_*) + \bar{\delta}_n^{\varepsilon},$$

then

$$\inf_{\mathcal{G}} Pf - Pf_* \leq \Big(1 - \varphi(\sqrt{\varepsilon}) - \frac{3}{2}\varepsilon\Big)^{-1}\Big[\inf_{\mathcal{G}} P_n f - P_n f_* + \frac{3}{2}\bar{\delta}_n^{\varepsilon} + \varphi^*\Big(\sqrt{\frac{2t}{\varepsilon n}}\Big) + \frac{t}{n}\Big]. \tag{6.12}$$

Proof. We assume, for simplicity, that Pf attains its minimum over $\mathcal{G}$ at some $\bar{f} \in \mathcal{G}$ (the proof can be easily modified if the minimum is not attained). Let E' be the event from the Definition 4.1 of the triple bound and let

$$E := \Big\{|(P_n - P)(\bar{f} - f_*)| \leq \sqrt{\frac{2t}{n}\mathrm{Var}_P(\bar{f} - f_*)} + \frac{t}{n}\Big\} \bigcap E'.$$

It follows from Bernstein inequality and the definition of the triple bound that

$$\mathbb{P}(E) \geq 1 - p - e^{-t}.$$

On the event E,

$$|(P_n - P)(\bar{f} - f_*)| \leq \sqrt{\frac{2t}{n}\mathrm{Var}_P(\bar{f} - f_*)} + \frac{t}{n}$$

and

$$\forall f \in \mathcal{G}\ \hat{\mathcal{E}}_n(\mathcal{G}; f) \leq \frac{3}{2}\Big(\mathcal{E}_P(\mathcal{G}; f) \vee \bar{\delta}_n\Big).$$

Also,

$$\mathrm{Var}_P^{1/2}(\bar{f} - f_*) \leq \varphi^{-1}(P\bar{f} - Pf_*)$$

and hence, on the event E,

$$|(P - P_n)(\bar{f} - f_*)| \leq \varphi(\sqrt{\varepsilon}\varphi^{-1}(P\bar{f} - Pf_*)) + \varphi^*\Big(\sqrt{\frac{2t}{\varepsilon n}}\Big) + \frac{t}{n} \leq$$

$$\leq \gamma\varphi(\sqrt{\varepsilon})(P\bar{f} - Pf_*) + \varphi^*\Bigl(\sqrt{\frac{2t}{\varepsilon n}}\Bigr) + \frac{t}{n},$$

implying

$$P_n(\bar{f} - f_*) \leq (1 + \gamma\varphi(\sqrt{\varepsilon}))P(\bar{f} - f_*) + \varphi^*\Bigl(\sqrt{\frac{2t}{\varepsilon n}}\Bigr) + \frac{t}{n} \tag{6.13}$$

and

$$P(\bar{f} - f_*) \leq (1 - \gamma\varphi(\sqrt{\varepsilon}))^{-1}\Bigl[P_n(\bar{f} - f_*) + \varphi^*\Bigl(\sqrt{\frac{2t}{\varepsilon n}}\Bigr) + \frac{t}{n}\Bigr] \tag{6.14}$$

(6.13) immediately yields the first bound of the lemma.

Since, in addition, on the event E

$$P_n(\bar{f} - f_*) = P_n\bar{f} - \inf_{\mathcal{G}} P_n f + \inf_{\mathcal{G}} P_n f - P_n f_* = \hat{\mathcal{E}}_n(\mathcal{G}; \bar{f}) + \inf_{\mathcal{G}} P_n f - P_n f_* \leq$$

$$\leq \inf_{\mathcal{G}} P_n f - P_n f_* + \frac{3}{2}\Bigl(\mathcal{E}_P(\mathcal{G}; \bar{f}) \vee \bar{\delta}_n\Bigr),$$

and since $\mathcal{E}_P(\mathcal{G}; \bar{f}) = 0$, we get

$$P_n(\bar{f} - f_*) \leq \inf_{\mathcal{G}} P_n f - P_n f_* + \frac{3}{2}\bar{\delta}_n.$$

Along with (6.14), this implies

$$\inf_{\mathcal{G}} Pf - Pf_* = P(\bar{f} - f_*) \leq (1 - \gamma\varphi(\sqrt{\varepsilon}))^{-1}\Bigl[\inf_{\mathcal{G}} P_n f - P_n f_* + \frac{3}{2}\bar{\delta}_n + \varphi^*\Bigl(\sqrt{\frac{2t}{\varepsilon n}}\Bigr) + \frac{t}{n}\Bigr],$$

which is the second bound of the lemma.

Finally, to prove the third bound it is enough to substitute the bound on $\bar{\delta}_n$ into (6.11) and to solve the resulting inequality with respect to $\inf_{\mathcal{G}} Pf - Pf_*$. □

Proof of Theorem 6.4. Let E_k be the event defined in Lemma 6.3 for $\mathcal{G} = \mathcal{F}_k$ and $t = t_k$, so that

$$\mathbb{P}(E_k) \geq 1 - p_k - e^{-t_k}.$$

Let

$$E := \bigcap_{k \geq 1} E_k.$$

Then

$$\mathbb{P}(E) \geq 1 - \sum_{k \geq 1}\Bigl(p_k + e^{-t_k}\Bigr).$$

On the event E, for all $k \geq 1$

$$\mathcal{E}_P(\mathcal{F}_k; \hat{f}_k) = P\hat{f}_k - \inf_{\mathcal{F}_k} Pf \leq \bar{\delta}_n(k)$$

and

$$\bar{\delta}_n(k) \leq \hat{\delta}_n(k) \leq \tilde{\delta}_n(k).$$

On the event E, using first bound (6.11) and then bound (6.10) of Lemma 6.3, we get

$$\mathcal{E}_P(\mathcal{F};\hat{f}) = P\hat{f} - \inf_{\mathcal{F}} Pf = P\hat{f}_{\hat{k}} - Pf_* = P\hat{f}_{\hat{k}} - \inf_{\mathcal{F}_{\hat{k}}} Pf + \inf_{\mathcal{F}_{\hat{k}}} Pf - Pf_* \leq$$

$$\leq \bar{\delta}_n(\hat{k}) + \inf_{\mathcal{F}_{\hat{k}}} Pf - Pf_* \leq$$

$$\leq (1-\gamma\varphi(\sqrt{\varepsilon}))^{-1}\Bigl[(1-\gamma\varphi(\sqrt{\varepsilon}))\bar{\delta}_n(\hat{k}) + \inf_{\mathcal{F}_{\hat{k}}} P_n f - P_n f_* + \frac{3}{2}\bar{\delta}_n(\hat{k}) + \varphi^*\Bigl(\sqrt{\frac{2t_{\hat{k}}}{\varepsilon n}}\Bigr) + \frac{t_{\hat{k}}}{n}\Bigr] \leq$$

$$\leq (1-\gamma\varphi(\sqrt{\varepsilon}))^{-1}\Bigl\{\inf_k\Bigl[\inf_{\mathcal{F}_k} P_n f + (5/2-\gamma\varphi(\sqrt{\varepsilon}))\hat{\delta}_n(k) + \varphi^*\Bigl(\sqrt{\frac{2t_k}{\varepsilon n}}\Bigr) + \frac{t_k}{n}\Bigr] - P_n f_*\Bigr\} =$$

$$= (1-\gamma\varphi(\sqrt{\varepsilon}))^{-1}\Bigl\{\inf_k\Bigl[\inf_{\mathcal{F}_k} P_n f + \hat{\pi}(k)\Bigr] - P_n f_*\Bigr\} \leq$$

$$\leq \frac{1+\gamma\varphi(\sqrt{\varepsilon})}{1-\gamma\varphi(\sqrt{\varepsilon})}\inf_k\Bigl[\inf_{\mathcal{F}_k} Pf - \inf_{\mathcal{F}} Pf + \frac{5/2-\gamma\varphi(\sqrt{\varepsilon})}{1+\gamma\varphi(\sqrt{\varepsilon})}\tilde{\delta}_n(k) +$$

$$+\frac{2}{1+\gamma\varphi(\sqrt{\varepsilon})}\varphi^*\Bigl(\sqrt{\frac{2t_k}{\varepsilon n}}\Bigr) + \frac{2}{(1+\gamma\varphi(\sqrt{\varepsilon}))}\frac{t_k}{n}\Bigr] =$$

$$\inf_k \frac{1+\gamma\varphi(\sqrt{\varepsilon})}{1-\gamma\varphi(\sqrt{\varepsilon})}\Bigl[\inf_{\mathcal{F}_k} Pf - \inf_{\mathcal{F}} Pf + \tilde{\pi}(k)\Bigr],$$

and the result follows.

□

Remark. Suppose that, for each k, $\bar{\delta}_n(k)$ is an admissible excess risk bound for the class $\mathcal{F}_k$ on an event E_k with $\mathbb{P}(E_k) \geq 1 - p_k$ (see Definition 4.1). It is easily seen from the proof of Theorem 6.4 that the same oracle inequality holds for arbitrary penalties $\hat{\pi}(k)$ and $\tilde{\pi}(k)$ such that on the event E_k

$$\hat{\pi}(k) \geq A(\varepsilon)\bar{\delta}_n(k) + \varphi^*\Bigl(\sqrt{\frac{2t_k}{\varepsilon n}}\Bigr) + \frac{t_k}{n}$$

and

$$\tilde{\pi}(k) \geq \frac{\hat{\pi}(k)}{1+\gamma\varphi(\sqrt{\varepsilon})} + \frac{\varphi^*\Bigl(\sqrt{\frac{2t_k}{\varepsilon n}}\Bigr)}{1+\gamma\varphi(\sqrt{\varepsilon})} + \frac{t_k}{(1+\gamma\varphi(\sqrt{\varepsilon}))n}.$$

As it has been already mentioned, the dependence of the penalty on the link function φ is the most troubling aspect of this approach since in such problems as classification this function depends on the unknown parameters of distribution P (such as "low noise" constants α in (5.7) and h in (5.6), see Section 5.3). Because of this, it is of importance to know that using Remark 1, it is easy to construct a version of the penalties that *do not depend on φ directly.* Suppose that the number of classes $\mathcal{F}_k$ is finite, say, N. Take

$$t_k := t + \log N,\ \ k = 1, \dots, N.$$

Define

$$\hat{k} := \mathrm{argmin}_{1\le k\le N}\biggl[\min_{f\in\mathcal{F}_k} P_n f + \frac{5}{2}\hat{\delta}_n(k)\biggr]$$

and $\hat{f} := \hat{f}_{\hat{k}}$. Note that we also have

$$\hat{k} := \mathrm{argmin}_{1\le k\le N}\biggl[\min_{f\in\mathcal{F}_k} P_n f + \hat{\pi}(k)\biggr],$$

where

$$\hat{\pi}(k) := \frac{5}{2}\hat{\delta}_n(k) + \varphi^*\Bigl(\sqrt{\frac{2t_k}{\varepsilon n}}\Bigr) + \frac{t_k}{n} = \frac{5}{2}\hat{\delta}_n(k) + \varphi^*\Bigl(\sqrt{\frac{2(t+\log N)}{\varepsilon n}}\Bigr) + \frac{t+\log N}{n},$$

since t_k in the additional two terms of the definition of $\hat{\pi}(k)$ does not depend on k. Denote

$$\tilde{\pi}(k) := \frac{5}{2}\tilde{\delta}_n(k) + 2\varphi^*\Bigl(\sqrt{\frac{2(t+\log N)}{\varepsilon n}}\Bigr) + 2\frac{t+\log N}{n}.$$

Then it follows from Theorem 6.4 and from the Remark that

$$\mathbb{P}\biggl\{\mathcal{E}_P(\mathcal{F};\hat{f}) \ge C(\varepsilon)\inf_{1\le k\le N}\Bigl\{\inf_{f\in\mathcal{F}_k} Pf - \inf_{f\in\mathcal{F}} Pf + \tilde{\pi}(k)\Bigr\}\biggr\} \le e^{-t} + \sum_{k=1}^{N} p_k. \tag{6.15}$$

Example. Consider, for instance, model selection in binary classification problems (see Section 5.3). Suppose that condition (5.6) holds with some $h > 0$ and, as a result, condition (6.8) holds with $\varphi(u) = hu^2$ for any $f = \ell \bullet g$ and $f_* = \ell \bullet g_*$, where g is a binary classifier, g_* is the Bayes classifier and $\ell(y,u) = I(y \neq u)$ is the binary loss.

Let $\{\mathcal{G}_k\}$ be a family of classes of functions from S into $\{-1, 1\}$ (binary classifiers). For any k, define

$$\hat{g}_{n,k} := \mathrm{argmin}_{g\in\mathcal{G}_k} L_n(g) = \mathrm{argmin}_{g\in\mathcal{G}_k} n^{-1}\sum_{j=1}^{n} I(Y_j \neq g(X_j)).$$

Let $\mathcal{F}_k := \{\ell \bullet g : g \in \mathcal{G}_k\}$. Denote $(\bar{\delta}_n(k), \hat{\delta}_n(k), \tilde{\delta}_n(k))$ the standard triple bound of Theorem 4.8 for the class $\mathcal{F}_k$ of level $1 - p_k$. Suppose that $\sum_{k=1}^{N} p_k = p \in (0,1)$. Define

$$\hat{k} := \mathrm{argmin}_{1\le k\le N}\Bigl[\inf_{g\in\mathcal{G}_k} L_n(g) + \frac{5}{2}\hat{\delta}_n(k)\Bigr]$$

and $\hat{g} := \hat{g}_{n,\hat{k}}$. Then it easily follows from bound (6.15) that with probability at least $1 - p - e^{-t}$

$$L(\hat{g}) - L(g_*) \leq C \inf_{1\leq k\leq N}\biggl[\inf_{g\in\mathcal{G}_k} L(g) - L(g_*) + \tilde{\delta}_n(k) + \frac{t + \log N}{nh}\biggr]$$

(we have fixed $\varepsilon > 0$ and the constant C may depend on ε). It is also easy to deduce from the proof of Theorem 5.5 that for the standard choice of $\bar{\delta}_n(k)$

$$\bar{\delta}_n(k) \leq C\biggl[\inf_{g\in\mathcal{G}_k} L(g) - L(g_*) + \frac{\mathbb{E}\log\Delta^{\mathcal{G}_k}(X_1,\dots,X_n)}{nh} + \frac{t_k}{nh}\biggr],$$

which (after some tuning of the constants) leads to the following oracle inequality that holds with probability at least $1 - p - e^{-t}$:

$$L(\hat{g}) - L(g_*) \leq C \inf_{1\leq k\leq N}\biggl[\inf_{g\in\mathcal{G}_k} L(g) - L(g_*) + \frac{\mathbb{E}\log\Delta^{\mathcal{G}_k}(X_1,\dots,X_n)}{nh}\biggr] + C\frac{t + \log N}{nh}.$$

Thus, this penalization method is adaptive to unknown noise parameter h.

We conclude this section with stating a result of Massart [73, 74] that can be derived based on the approach of Theorem 6.4. Suppose that $\{\mathcal{F}_k\}$ is a sequence of function classes such that condition (4.4) holds for each class $\mathcal{F}_k$ with some constant $D_k \geq 1$, i.e.,

$$D_k(Pf - Pf_*) \geq \rho_P^2(f, f_*) \geq \mathrm{Var}_P(f - f_*).$$

We will assume that the sequence $\{D_k\}$ is nondecreasing. Denote

$$\bar{\delta}_n^{\varepsilon}(k) := D_k^{-1}\omega_n^{\sharp}\biggl(\frac{\varepsilon}{KD_k}\biggr) + \frac{KD_kt_k}{n\varepsilon}.$$

If K is large enough, then Lemma 4.1 implies that the following bound holds:

$$\bar{\delta}_n(k) := \sigma_n^{t_k}(\mathcal{F}_k; P) \leq \varepsilon(\inf_{\mathcal{F}_k} Pf - Pf_*) + \bar{\delta}_n^{\varepsilon}(k).$$

Also, it follows from the proof of Theorem 4.3 that $\bar{\delta}_n(k)$ is an admissible excess risk bound of level $1 - C_q e^{-t_k}$.

Suppose that for each k there exist a data dependent bound $\hat{\delta}_n^{\varepsilon}(k)$ and a distribution dependent bound $\tilde{\delta}_n^{\varepsilon}(k)$ such that

$$\mathbb{P}\biggl\{\bar{\delta}_n^{\varepsilon}(k) \leq \hat{\delta}_n^{\varepsilon}(k) \leq \tilde{\delta}_n^{\varepsilon}(k)\biggr\} \geq 1 - p_k,\ k \geq 1.$$

Define the following penalties:

$$\hat{\pi}_n^{\varepsilon}(k) := 3\hat{\delta}_n^{\varepsilon}(k) + \frac{\hat{K}D_kt_k}{\varepsilon n} \text{ and } \tilde{\pi}_n^{\varepsilon}(k) := 3\tilde{\delta}_n^{\varepsilon}(k) + \frac{\tilde{K}D_kt_k}{\varepsilon n}$$

with some numerical constants $\hat{K}, \tilde{K}$. Let

$$\hat{k} := \mathrm{argmin}_{k\geq 1}\biggl[\min_{f\in\mathcal{F}_k} P_n f + \hat{\pi}_n^{\varepsilon}(k)\biggr]$$

and $\hat{f} := \hat{f}_{\hat{k}}$.

Theorem 6.5 *There exist numerical constants $\hat{K}, \tilde{K}, C$ such that for any sequence $\{t_k\}$ of positive numbers,*

$$\mathbb{P}\biggl\{P\hat{f} - Pf_* \geq \frac{1+\varepsilon}{1-\varepsilon}\inf_{k\geq 1}\Bigl\{\inf_{f\in\mathcal{F}_k} Pf - Pf_* + \tilde{\pi}_n^{\varepsilon}(k)\Bigr\}\biggr\} \leq \sum_{k=1}^{\infty}\biggl(p_k + (C+1)e^{-t_k}\biggr).$$

To prove this result one has to extend theorem 6.4 to the case when condition (6.8) holds for each function class $\mathcal{F}_k$ with a different link function φ_k and to use this extension for $\varphi_k(u) = u^2/D_k$ and $\varphi_k^*(v) = D_k v^2/4$.

7 Linear Programming in Sparse Recovery

7.1 Sparse Recovery and Neighborliness of Convex Polytopes

Let $\mathcal{H} := \{h_1, \dots, h_N\}$ be a given finite set of measurable functions from S into $\mathbb{R}$. In what follows, it will be called *a dictionary.* Given $J \subset \{1, \dots, N\}$, we will write $d(J) := \mathrm{card}(J)$. For $\lambda = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}^N$, denote

$$f_\lambda = \sum_{j=1}^{N} \lambda_j h_j, \quad J_\lambda = \mathrm{supp}(\lambda) := \Bigl\{j : \lambda_j \neq 0\Bigr\} \;\text{ and }\; d(\lambda) := d(J_\lambda).$$

Suppose that a function

$$f_* \in \mathrm{l.s.}(\mathcal{H}) = \Bigl\{f_\lambda : \lambda \in \mathbb{R}^N\Bigr\}$$

from the linear span of the dictionary is observed (measured) at points $X_1, \dots, X_n \in S$. For simplicity, we first assume that there is no noise in the observations:

$$Y_j = f_*(X_j),\ j = 1, \dots, n.$$

The goal is to recover a representation of f_* in the dictionary. We are mostly interested in the case when $N > n$ (in fact, N can be much larger than n). Define

$$L := \Bigl\{\lambda \in \mathbb{R}^N : f_\lambda(X_j) = Y_j,\ j = 1, \dots, n\Bigr\}.$$

Then, L is an affine subspace of dimension at least $N-n$, so, the representation of f_* in the dictionary is not unique. In such cases, it is often of interest to find *the sparsest representation,* which means solving the problem

$$\|\lambda\|_{\ell_0} = \sum_{j=1}^N I(\lambda_j \neq 0) \longrightarrow \min, \lambda \in L. \tag{7.1}$$

If we introduce the following $n \times N$ matrix

$$A := \Bigl(h_j(X_i) : 1 \leq i \leq n, 1 \leq j \leq N\Bigr)$$

and denote $\vec{Y}$ the vector with components $Y_1, \dots, Y_n$, then problem (7.1) can be also rewritten as

$$\|\lambda\|_{\ell_0} = \sum_{j=1}^N I(\lambda_j \neq 0) \longrightarrow \min, \quad A\lambda = \vec{Y}. \tag{7.2}$$

When N is large, such problems are computationally intractable since the function to be minimized is non-smooth and non-convex. Essentially, solving (7.2) would require searching through all 2^N coordinate subspaces of $\mathbb{R}^N$. Because of this, the following convex relaxation of the problem is frequently used:

$$\|\lambda\|_{\ell_1} = \sum_{j=1}^N |\lambda_j| \longrightarrow \min, \lambda \in L, \tag{7.3}$$

or, equivalently,

$$\|\lambda\|_{\ell_1} = \sum_{j=1}^N |\lambda_j| \longrightarrow \min, A\lambda = \vec{Y}. \tag{7.4}$$

The last minimization problem is convex and, moreover, it is a linear programming problem. However, the question is whether solving (7.3) has anything to do with solving (7.1). Next result (due to Donoho [37]) gives an answer to this question by reducing it to some interesting problems in the geometry of convex polytopes. To formulate the result, define

$$P := AU_{\ell_1} = \mathrm{conv}\Bigl(\Bigl\{a_1, -a_1, \dots, a_N, -a_N\Bigr\}\Bigr),$$

where

$$U_{\ell_1} := \{\lambda \in \mathbb{R}^N : \|\lambda\|_{\ell_1} \leq 1\}$$

is the unit ball in ℓ_1 and $a_1, \dots, a_N \in \mathbb{R}^n$ are columns of matrix A. (In what follows, U_B denotes the closed unit ball centered at 0 of a Banach space B).

Clearly, P is a centrally symmetric convex polytope in $\mathbb{R}^n$ with at most $2N$ vertices. Such a centrally symmetric polytope is called *d-neighborly* if any set of $d+1$ vertices that does not contain antipodal vertices (such as a_k and $-a_k$) spans a face of P.

Theorem 7.1 *Suppose that $N > n$. The following two statements are equivalent:*
(i) The polytope P has $2N$-vertices and is d-neighborly;
(ii) Any solution λ of the system of linear equations $A\lambda = \vec{Y}$ such that $d(\lambda) \leq d$ is the **unique** *solution of problem (7.4).*

The unit ball U_{ℓ_1} of ℓ_1 is a trivial example of an N-neighborly centrally symmetric polytope. However, it is hard to find nontrivial constructive examples of such polytopes with a "high neighborliness". Their existence is usually proved by a probabilistic method, for instance, by choosing the design matrix A at random and showing that the resulting random polytope P is d-neighborly for sufficiently large d with a high probability. The problem has been studied for several classes of random matrices (projections on an n-dimensional subspace picked at random from the Grassmannian of all n-dimensional subspaces; random matrices with i.i.d. Gaussian or Bernoulli entries, etc) both in the case of centrally symmetric polytopes and without the restriction of central symmetry, see Vershik and Sporyshev [96], Affentranger and Schneider [1] and, in connection with sparse recovery, Donoho [37], Donoho and Tanner [38]. The approach taken in these papers is based on rather subtle geometric analysis of the properties of high-dimensional convex polytopes, in particular, on computation of their internal and external angles. This leads to rather sharp estimates of the largest d for which the neighborliness still holds (in other words, for which the phase transition occurs and the polytope starts losing faces). Here we follow another approach that is close to Rudelson and Vershynin [82] and Mendelson, Pajor and Tomczak-Jaegermann [78]. This approach is more probabilistic, it is much simpler and it addresses the sparse recovery problem more directly. On the other hand, it does not give precise bounds on the maximal d for which sparse recovery is possible (although it still provides correct answers up to constants).

7.2 Geometric Properties of the Dictionary

For $J \subset \{1, \ldots, N\}$ and $b \in [0, +\infty]$, define the following cone consisting of vectors whose "dominant coordinates" are in J :

$$C_{b,J} := \left\{ u \in \mathbb{R}^N : \sum_{j \notin J} |u_j| \leq b \sum_{j \in J} |u_j| \right\}.$$

Clearly, for $b = +\infty$, $C_{b,J} = \mathbb{R}^N$. For $b = 0$, $C_{b,J}$ is the linear subspace $\mathbb{R}^J$ of vectors $u \in \mathbb{R}^N$ with $\mathrm{supp}(u) \subset J$. For $b = 1$, we will write $C_J := C_{1,J}$. Such cones will be called **cones of dominant coordinates** and some norms in $\mathbb{R}^N$ will be compared on these cones.

Some useful geometric properties of the cones of dominant coordinates will be summarized in the following lemma. It includes several well known facts (see Candes and Tao [29], proof of Theorem 1; Ledoux and Talagrand [68], p. 421; Mendelson, Pajor and Tomczak-Jaegermann [78], Lemma 3.3).

With a minor abuse of notations, we will identify in what follows vectors $u \in \mathbb{R}^N$ with $\mathrm{supp} u \subset J$, where $J \subset \{1, \dots, N\}$, with vectors $u = (u_j : j \in J) \in \mathbb{R}^J$.

Lemma 7.1 *Let $J \subset \{1, \dots, N\}$ and let $d := \mathrm{card}(J)$.*
(i) Take $u \in C_{b,J}$ and denote $J_0 := J$. For $s \geq 1$, J_1 will denote the set of s coordinates in $\{1, \dots, N\} \setminus J_0$ for which $|u_j|'s$ are the largest, J_2 will be the set of s coordinates in $\{1, \dots, N\} \setminus (J_0 \cup J_1)$ for which $|u_j|'s$ are the largest, etc. (at the end, there might be fewer than s coordinates left). Denote $u^{(k)} := (u_j : j \in J_k)$. Then $u = \sum_{k \geq 0} u^{(k)}$ and

$$\sum_{k \geq 2} \|u^{(k)}\|_{\ell_2} \leq \frac{b}{\sqrt{s}} \sum_{j \in J} |u_j| \leq b\sqrt{\frac{d}{s}} \biggl(\sum_{j \in J} |u_j|^2\biggr)^{1/2}.$$

In addition,

$$\|u\|_{\ell_2} \leq \biggl(b\sqrt{\frac{d}{s}} + 1\biggr) \biggl(\sum_{j \in J_0 \cup J_1} |u_j|^2\biggr)^{1/2}.$$

(ii) Denote

$$K_J := C_{b,J} \cap U_{B_{\ell_2}}.$$

There exists a set $\mathcal{M}_d \subset U_{\ell_2}$ such that $d(u) \leq d, u \in \mathcal{M}_d$,

$$\mathrm{card}(\mathcal{M}_d) \leq 5^d \binom{N}{\leq d}$$

and

$$K_J \subset 2(2 + b)\mathrm{conv}(\mathcal{M}_d).$$

Proof. To prove (i), note that, for all $j \in J_{k+1}$,

$$|u_j| \leq \frac{1}{s} \sum_{i \in J_k} |u_i|,$$

implying that

$$(\sum_{j\in J_{k+1}} |u_j|^2)^{1/2} \le \frac{1}{\sqrt{s}} \sum_{j\in J_k} |u_j|.$$

Add these inequalities for $k = 1, 2, \dots$ to get

$$\sum_{k\ge 2} \|u^{(k)}\|_{\ell_2} \le \frac{1}{\sqrt{s}} \sum_{j\notin J} |u_j| \le \frac{b}{\sqrt{s}} \sum_{j\in J} |u_j| \le b\sqrt{\frac{d}{s}}\Big(\sum_{j\in J} |u_j|^2\Big)^{1/2} \le b\sqrt{\frac{d}{s}}\Big(\sum_{j\in J\cup J_1} |u_j|^2\Big)^{1/2}.$$

Therefore, for $u \in C_J$,

$$\|u\|_{\ell_2} \le \Big(b\sqrt{\frac{d}{s}} + 1\Big)\Big(\sum_{j\in J_0\cup J_1} |u_j|^2\Big)^{1/2}.$$

To prove (ii) note that

$$K_J \subset (2+b)\ \mathrm{conv}\Big(\bigcup B_I : I \subset \{1,\dots,N\}, d(I) \le d\Big),$$

where

$$B_I := \Big\{(u_i : i \in I) : \sum_{i\in I} |u_i|^2 \le 1\Big\}.$$

Indeed, it is enough to consider $u \in K_J$ and to use statement (i) with $s = d$. Then, we have $u^{(0)} \in B_{J_0}$, $u^{(1)} \in B_{J_1}$ and

$$\sum_{k\ge 2} u^{(k)} \in b\ \mathrm{conv}\Big(\bigcup B_I : I \subset \{1,\dots,N\}, d(I) \le d\Big).$$

It is easy to see that if B is the unit Euclidean ball in $\mathbb{R}^d$ and M is a 1/2-net of this ball, then

$$B \subset 2\ \mathrm{conv}(M).$$

Here is a sketch of the proof of the last claim. For convex sets $C_1, C_2 \subset \mathbb{R}^N$, denote by $C_1 + C_2$ their Minkowski sum

$$C_1 + C_2 = \{x_1 + x_2 : x_1 \in C_1, x_2 \in C_2\}.$$

It follows that

$$B \subset M + \frac{1}{2}B \subset \mathrm{conv}(M) + \frac{1}{2}B \subset \mathrm{conv}(M) + \frac{1}{2}\mathrm{conv}(M) + \frac{1}{4}B \subset \dots$$

$$\mathrm{conv}(M) + \frac{1}{2}\mathrm{conv}(M) + \frac{1}{4}\mathrm{conv}(M) + \dots \subset 2\mathrm{conv}(M).$$

For each I with $d(I) = d$, denote M_I a minimal 1/2-net of B_I. Then,

$$K_J \subset 2(2+b)\ \mathrm{conv}\biggl(\bigcup M_I : I \subset \{1, \dots, N\}, d(I) \le d\biggr) =: 2(2+b)\ \mathrm{conv}(\mathcal{M}_d).$$

By an easy combinatorial argument,

$$\mathrm{card}(\mathcal{M}_d) \le 5^d \binom{N}{\le d},$$

so, the proof is complete.

□

In what follows, we will need several geometric characteristics of the dictionary $\mathcal{H}$.

Given a probability measure Π on S, denote

$$\beta^{(b)}(J;\Pi) := \inf\biggl\{\beta > 0 : \sum_{j\in J} |\lambda_j| \le \beta \biggl\|\sum_{j=1}^N \lambda_j h_j\biggr\|_{L_1(\Pi)},\ \lambda \in C_{b,J}\biggr\}$$

and

$$\beta_2^{(b)}(J;\Pi) := \inf\biggl\{\beta > 0 : \sum_{j\in J} |\lambda_j|^2 \le \beta^2 \biggl\|\sum_{j=1}^N \lambda_j h_j\biggr\|^2_{L_2(\Pi)},\ \lambda \in C_{b,J}\biggr\}.$$

We will denote

$$\beta(J,\Pi) := \beta^{(1)}(J,\Pi), \quad \beta_2(J,\Pi) := \beta_2^{(1)}(J,\Pi).$$

As soon as the distribution Π is fixed, we will often suppress Π in our notations and write $\beta(J), \beta_2(J)$, etc. In the case when $J = \emptyset$, we set $\beta^{(b)}(J) = \beta_2^{(b)}(J) = 0$. Note that if $J \ne \emptyset$ and $h_1, \dots, h_N$ are linearly independent in $L_1(\Pi)$ or in $L_2(\Pi)$, then, for all $b \in (0, +\infty)$, $\beta^{(b)}(J) < +\infty$ or, respectively, $\beta_2^{(b)}(J) < +\infty$. In the case of orthonormal dictionary, $\beta_2^{(b)}(J) = 1$.

We will use several properties of $\beta^{(b)}(J)$ and $\beta_2^{(b)}(J)$ and their relationships with other common characteristics of the dictionary.

Let $\kappa(J)$ denote the minimal eigenvalue of the Gram matrix $\Bigl(\langle h_i, h_j\rangle_{L_2(\Pi)}\Bigr)_{i,j\in J}$. Also denote L_J the linear span of $\{h_j : j \in J\}$ and let

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

Thus, $\rho(J)$ is the largest "correlation coefficient" (or the largest cosine of the angle) between functions in the linear span of a subset $\{h_j : j \in J\}$ of the dictionary and the linear span of its complement (compare $\rho(J)$ with the notion of canonical correlation in multivariate statistical analysis). In fact, we will rather need a somewhat different quantity defined in terms of the cone $C_{b,J}$:

$$\rho^{(b)}(J) := \sup_{\lambda \in C_{b,J}} \frac{\left|\left\langle \sum_{j\in J}\lambda_j h_j, \sum_{j\notin J}\lambda_j h_j\right\rangle_{L_2(\Pi)}\right|}{\left\|\sum_{j\in J}\lambda_j h_j\right\|_{L_2(\Pi)}\left\|\sum_{j\notin J}\lambda_j h_j\right\|_{L_2(\Pi)}}.$$

Clearly, $\rho^{(b)}(J) \le \rho(J)$.

Proposition 7.1 *The following bound holds:*

$$\beta_2(J) \le \frac{1}{\sqrt{\kappa(J)(1-(\rho^{(b)}(J))^2)}}. \tag{7.5}$$

Proof. Indeed, the next inequality is obvious

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

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

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

where α is the angle between f and g. This yields

$$\Bigl(\sum_{j\in J}|\lambda_j|^2\Bigr)^{1/2} \le \frac{1}{\sqrt{\kappa(J)}}\Bigl\|\sum_{j\in J}\lambda_j h_j\Bigr\|_{L_2(\Pi)} \le \frac{1}{\sqrt{\kappa(J)(1-(\rho^{(b)}(J))^2)}}\Bigl\|\sum_{j=1}^N \lambda_j h_j\Bigr\|_{L_2(\Pi)},$$

which implies (7.5).

□

Lemma 7.1 can be used to provide upper bounds on $\beta_2^{(b)}(J)$. To formulate such bounds, we first introduce so called *restricted isometry constants.*

For $d = 1, \dots, N$, let $\delta_d(\Pi)$ be the smallest $\delta > 0$ such that, for all $\lambda \in \mathbb{R}^N$ with $d(\lambda) \le d$,

$$(1-\delta)\|\lambda\|_{\ell_2} \le \Bigl\|\sum_{j=1}^N \lambda_j h_j\Bigr\|_{L_2(\Pi)} \le (1+\delta)\|\lambda\|_{\ell_2}.$$

If $\delta_d(\Pi) < 1$, then d-dimensional subspaces spanned on subsets of the dictionary and equipped with (a) the $L_2(\Pi)$-norm and (b) the ℓ_2-norm on vectors of coefficients are "almost" isometric. For a given dictionary $\{h_1, \dots, h_N\}$, the quantity $\delta_d(\Pi)$ will be called *the restricted isometry constant* of dimension d with respect to measure Π. The dictionary satisfies *a restricted isometry condition* in $L_2(\Pi)$ if $\delta_d(\Pi)$ is sufficiently small for a sufficiently large value of d (that, in sparse recovery, is usually related to the underlying "sparsity" of the problem).

For $I, J \subset \{1, \dots, N\}$, $I \cap J = \emptyset$, denote

$$r(I; J) := \sup_{f \in L_I, g \in L_J, f, g \neq 0} \left| \frac{\langle f, g \rangle_{L_2(\Pi)}}{\|f\|_{L_2(\Pi)} \|g\|_{L_2(\Pi)}} \right|.$$

Note that $\rho(J) = r(J, J^c)$. Let

$$\rho_d := \max\Bigl\{ r(I, J) : I, J \subset \{1, \dots, N\},\ I \cap J = \emptyset,\ \mathrm{card}(I) = 2d,\ \mathrm{card}(J) = d \Bigr\}.$$

This quantity measures the correlation between linear spans of disjoint parts of the dictionary of fixed "small cardinalities", in this case, d and $2d$.

Define

$$m_d := \inf\{\|f_u\|_{L_2(\Pi)} : u \in \mathbb{R}^N, \|u\|_{\ell_2} = 1, d(u) \leq d\}$$

and

$$M_d := \sup\{\|f_u\|_{L_2(\Pi)} : u \in \mathbb{R}^N, \|u\|_{\ell_2} = 1, d(u) \leq d\}.$$

If $m_d \leq 1 \leq M_d \leq 2$, the restricted isometry constant can be written as

$$\delta_d = (M_d - 1) \vee (1 - m_d).$$

Lemma 7.2 *Suppose $J \subset \{1, \dots, N\}$, $d(J) = d$ and $\rho_d < \frac{m_{2d}}{bM_{2d}}$. Then*

$$\beta_2^{(b)}(J) \leq \frac{1}{m_{2d} - b\rho_d M_{2d}}.$$

Proof. Denote P_I the orthogonal projection on $L_I \subset L_2(\Pi)$. Under the notations of Lemma 7.1, for all $u \in C_J$,

$$\Bigl\| \sum_{j=1}^N u_j h_j \Bigr\|_{L_2(\Pi)} \geq \Bigl\| P_{J_0 \cup J_1} \sum_{j=1}^N u_j h_j \Bigr\|_{L_2(\Pi)} \geq \Bigl\| \sum_{j \in J_0 \cup J_1} u_j h_j \Bigr\|_{L_2(\Pi)} - \Bigl\| P_{J_0 \cup J_1} \sum_{j \notin J_0 \cup J_1} u_j h_j \Bigr\|_{L_2(\Pi)} \geq$$

$$\Bigl\| \sum_{j \in J_0 \cup J_1} u_j h_j \Bigr\|_{L_2(\Pi)} - \sum_{k \geq 2} \Bigl\| P_{J_0 \cup J_1} \sum_{j \in J_k} u_j h_j \Bigr\|_{L_2(\Pi)} \geq$$

$$\Bigl\| \sum_{j \in J_0 \cup J_1} u_j h_j \Bigr\|_{L_2(\Pi)} - \rho_d \sum_{k \geq 2} \Bigl\| \sum_{j \in J_k} u_j h_j \Bigr\|_{L_2(\Pi)} \geq$$

$$\Big\|\sum_{j\in J_0\cup J_1} u_j h_j\Big\|_{L_2(\Pi)} - \rho_d M_{2d}\sum_{k\geq 2}\|u^{(k)}\|_{\ell_2} \geq$$

$$\Big\|\sum_{j\in J_0\cup J_1} u_j h_j\Big\|_{L_2(\Pi)} - b\rho_d M_{2d}\Big(\sum_{j\in J\cup J_1}|u_j|^2\Big) \geq$$

$$\Big\|\sum_{j\in J_0\cup J_1} u_j h_j\Big\|_{L_2(\Pi)} - b\rho_d\frac{M_{2d}}{m_{2d}}\Big\|\sum_{j\in J_0\cup J_1} u_j h_j\Big\|_{L_2(\Pi)} = \Big(1 - b\rho_d\frac{M_{2d}}{m_{2d}}\Big)\Big\|\sum_{j\in J_0\cup J_1} u_j h_j\Big\|_{L_2(\Pi)}.$$

On the other hand,

$$\Big(\sum_{j\in J}|u_j|^2\Big)^{1/2} \leq \Big(\sum_{j\in J_0\cup J_1}|u_j|^2\Big)^{1/2} \leq m_{2d}^{-1}\Big\|\sum_{j\in J_0\cup J_1} u_j h_j\Big\|_{L_2(\Pi)},$$

implying that

$$\Big(\sum_{j\in J}|u_j|^2\Big)^{1/2} \leq m_{2d}^{-1}\Big(1 - b\rho_d\frac{M_{2d}}{m_{2d}}\Big)^{-1}\Big\|\sum_{j=1}^{N} u_j h_j\Big\|_{L_2(\Pi)}.$$

Therefore,

$$\beta_2(J) \leq \frac{1}{m_{2d} - b\rho_d M_{2d}}.$$

□

It is easy to check that

$$\rho_d \leq \frac{1}{2}\Big[\Big(\frac{1+\delta_{3d}}{1-\delta_{2d}}\Big)^2 + \Big(\frac{1+\delta_{3d}}{1-\delta_d}\Big)^2 - 2\Big] \bigvee \frac{1}{2}\Big[2 - \Big(\frac{1-\delta_{3d}}{1+\delta_{2d}}\Big)^2 - \Big(\frac{1-\delta_{3d}}{1+\delta_d}\Big)^2\Big].$$

Together with Lemma 7.2 this implies that $\beta_2(J) < +\infty$ for any set J such that $\mathrm{card}(J) \leq d$ provided that $\delta_{3d} \leq \frac{1}{8}$ (a sharper condition is also possible).

We will give a simple modification of Lemma 7.2 in spirit of [14].

Lemma 7.3 *Suppose* $J \subset \{1, \dots, N\}$, $d(J) = d$ *and, for some* $s \geq 1$,

$$\frac{M_s}{m_{d+s}} < \frac{1}{b}\sqrt{\frac{s}{d}}.$$

Then

$$\beta_2^{(b)}(J) \leq \frac{\sqrt{s}}{\sqrt{s}m_{d+s} - b\sqrt{d}M_s}.$$

Proof. For all $u \in C_{b,J}$,

$$\Bigl(\sum_{j\in J}|u_j|^2\Bigr)^{1/2} \le \frac{1}{m_{d+s}}\Bigl\|\sum_{j\in J\cup J_1} u_j h_j\Bigr\|_{L_2(\Pi)} \le \frac{1}{m_{d+s}}\Bigl\|\sum_{j=1}^{N} u_j h_j\Bigr\|_{L_2(\Pi)} + \frac{1}{m_{d+s}}\Bigl\|\sum_{j\notin J\cup J_1} u_j h_j\Bigr\|_{L_2(\Pi)}.$$

To bound the last norm in the right hand side, note that

$$\Bigl\|\sum_{j\notin J\cup J_1} u_j h_j\Bigr\|_{L_2(\Pi)} \le \sum_{k\ge 2}\Bigl\|\sum_{j\in J_k} u_j h_j\Bigr\|_{L_2(\Pi)} \le M_s\sum_{k\ge 2}\|u^{(k)}\|_{\ell_2} \le M_s\sqrt{\frac{d}{s}}\Bigl(\sum_{j\in J}|u_j|^2\Bigr)^{1/2}.$$

This yields the bound

$$\Bigl(\sum_{j\in J}|u_j|^2\Bigr)^{1/2} \le \frac{1}{m_{d+s}}\Bigl\|\sum_{j=1}^{N} u_j h_j\Bigr\|_{L_2(\Pi)} + \frac{M_s}{m_{d+s}}\sqrt{\frac{d}{s}}\Bigl(\sum_{j\in J}|u_j|^2\Bigr)^{1/2},$$

which implies the result.

□

In what follows, we will use several quantities that describe a way in which vectors in $\mathbb{R}^N$, especially, sparse vectors, are "aligned" with the dictionary. We will use the following definitions.

Let $D \subset \mathbb{R}^N$ be a convex set. For $\lambda \in D$, denote

$$T_D(\lambda) := \{v \in \mathbb{R}^N : \exists t > 0 \ \lambda + vt \in D\}.$$

The set $T_D(\lambda)$ will be called the tangent cone of convex set D at point λ (note that in the literature on convex analysis it is more common to refer to the closure of the set $T_D(\lambda)$ as "tangent cone"). Recall that

$$H := \Bigl(\langle h_i, h_j\rangle_{L_2(\Pi)}\Bigr)_{i,j=1,\dots,N}$$

denotes the Gram matrix of the dictionary in the space $L_2(\Pi)$. Whenever it is convenient, H will be viewed as a linear transformation of $\mathbb{R}^N$. For a vector $w \in \mathbb{R}^N$ and $b > 0$, we will denote $C_{b,w} := C_{b,\mathrm{supp}(w)}$, which is a cone of vectors whose "dominant" coordinates are in $\mathrm{supp}(w)$.

Now define

$$a_H^{(b)}(D,\lambda,w) := \sup\Bigl\{\langle w,u\rangle_{\ell_2} : u \in -T_D(\lambda)\cap C_{b,w}, \|f_u\|_{L_2(\Pi)} = 1\Bigr\},\ b \in [0,+\infty].$$

We will call these quantities the *alignment coefficients* of vector w, matrix H and convex set D at point $\lambda \in D$. In applications that follow, we want the alignment coefficient to be either negative, or, if positive, then small enough.

The geometry of the set D could have an impact on the alignment coefficients for some vectors w that are of interest in sparse recovery problems. For instance, if L is a convex function on D and $\lambda \in D$ is its minimal point, then there exists a subgradient $w \in \partial L(\lambda)$ of L at point λ such that, for all $u \in T_D(\lambda)$, $\langle w, u\rangle_{\ell_2} \geq 0$. This implies that $a_H^{(b)}(D, \lambda, w) \leq 0$. If $D = \mathbb{R}^N$, then $T_D(\lambda) = \mathbb{R}^N,\ \lambda \in \mathbb{R}^N$. In this case, we will write

$$a_H^{(b)}(w) := a_H^{(b)}(\mathbb{R}^N, \lambda, w) = \sup\Big\{\langle w, u\rangle_{\ell_2} : u \in C_{b,w}, \|f_u\|_{L_2(\Pi)} = 1\Big\}.$$

Despite the fact that the geometry of set D might be important, often, we are not taking it into account and replace $a_H^{(b)}(D, \lambda, w)$ by its upper bound $a_H^{(b)}(w)$.

Note that

$$\|f_u\|_{L_2(\Pi)}^2 = \langle Hu, u\rangle_{\ell_2} = \langle H^{1/2}u, H^{1/2}u\rangle_{\ell_2}.$$

We will frequently use the following form of alignment coefficient

$$a_H^{(\infty)}(D, \lambda, w) := \sup\Big\{\langle w, u\rangle_{\ell_2} : u \in -T_D(\lambda), \|f_u\|_{L_2(\Pi)} = 1\Big\},$$

or rather a simpler upper bound

$$a_H^{(\infty)}(w) = a_H^{(\infty)}(\mathbb{R}^N, \lambda, w) = \sup\Big\{\langle w, u\rangle_{\ell_2} : \|f_u\|_{L_2(\Pi)} = 1\Big\}.$$

The last quantity is a seminorm in $\mathbb{R}^N$ and, for all b, we have

$$a_H^{(b)}(w) \leq a_H^{(\infty)}(w) = \sup_{\|H^{1/2}u\|_{\ell_2}=1} \langle w, u\rangle_{\ell_2} =: \|w\|_H.$$

If H is nonsingular, we can further write

$$\|w\|_H = \sup_{\|H^{1/2}u\|_{\ell_2}=1} \langle H^{-1/2}w, H^{1/2}u\rangle_{\ell_2} = \|H^{-1/2}w\|_{\ell_2}.$$

Even when H is singular, we still have $\|w\|_H \leq \|H^{-1/2}w\|_{\ell_2}$, where, for $w \in \mathrm{Im}(H^{1/2}) = H^{1/2}\mathbb{R}^N$, one defines

$$\|H^{-1/2}w\|_{\ell_2} := \inf\{\|v\|_{\ell_2} : H^{1/2}v = w\}$$

(which means factorization of the space with respect to $\mathrm{Ker}(H^{1/2})$) and for $w \notin \mathrm{Im}(H^{1/2})$ the norm $\|H^{-1/2}w\|_{\ell_2}$ becomes infinite.

Note also that, for $b = 0$,

$$a_H^{(0)}(w) = a_H^{(0)}(\mathbb{R}^N, \lambda, w) = \sup\Big\{\langle w, u\rangle_{\ell_2} : \|f_u\|_{L_2(\Pi)} = 1, \mathrm{supp}(u) = \mathrm{supp}(w)\Big\}.$$

This also defines seminorms on subspaces of vectors w with a fixed support, say, $\mathrm{supp}(w) = J$. If $H_J := \Big(\langle h_i, h_j\rangle_{L_2(\Pi)}\Big)_{i,j\in J}$ is the corresponding submatrix of the Gram matrix H and H_J is nonsingular, then

$$a_H^{(0)}(w) = \|H_J^{-1/2} w\|_{\ell_2},$$

so, in this case, the alignment coefficient depends only on "small" submatrices of the Gram matrix corresponding to the support of w (which is, usually, sparse).

When $0 < b < +\infty$, the definition of alignment coefficients involves cones of dominant coordinates and their values are between the values in the two extreme cases of $b = 0$ and $b = \infty$.

It is easy to bound the alignment coefficient in terms of geometric characteristics of the dictionary introduced earlier in this section. For instance, if $J = \mathrm{supp}(w)$, then

$$\|w\|_H \le \frac{\|w\|_{\ell_2}}{\sqrt{\kappa(J)(1-\rho^2(J))}} \le \frac{\|w\|_{\ell_\infty}\sqrt{d(J)}}{\sqrt{\kappa(J)(1-\rho^2(J))}},$$

where $\kappa(J)$ is the minimal eigenvalue of the matrix $H_J = \Big(\langle h_i, h_j\rangle_{L_2(\Pi)}\Big)_{i,j\in J}$ and $\rho(J)$ is the "canonical correlation" defined above.

One can also upper bound the alignment coefficient in terms of the quantity

$$\beta_{2,b}(w;\Pi) := \beta_2^{(b)}(\mathrm{supp}(w);\Pi).$$

Namely, the following bound is straightforward:

$$a_H^{(b)}(w) \le \|w\|_{\ell_2}\beta_{2,b}(w;\Pi).$$

These upper bounds show that the size of the alignment coefficient is controlled by the "sparsity" of the vector w as well as by some characteristics of the dictionary (or its Gram matrix H). For orthonormal dictionaries and for dictionaries that are close enough to being orthonormal (so that, for instance, $\kappa(J)$ is bounded away from 0 and $\rho^2(J)$ is bounded away from 1), the alignment coefficient is bounded from above by a quantity of the order $\|w\|_{\ell_\infty}\sqrt{d(J)}$. However, this is only an upper bound and the alignment coefficient itself is a more flexible characteristic of rather complicated geometric relationships between the vector w and the dictionary. Even the quantity $\|H^{-1/2}w\|_{\ell_2}$ (a rough upper bound on the alignment coefficient not taking into account the geometry of the cone of dominant coordinates), depends not only on the sparsity of w, but also on the way in which this vector is aligned with the eigenspaces of H. If w belongs to the linear span of eigenspaces that correspond to large eigenvalues of H, then $\|H^{-1/2}w\|_{\ell_2}$ can be of the order $\|w\|_{\ell_2}$.

Note that the geometry of the problem is the geometry of the Hilbert space $L_2(\Pi)$, so it strongly depends on the unknown distribution Π of the design variable.

7.3 Sparse Recovery in Noiseless Problems

Let Π_n denote the empirical measure based on the points $X_1, \dots, X_n$ (at the moment, not necessarily random).

Proposition 7.2 *Let $\hat\lambda$ be a solution of (7.3). If $\lambda^* \in L$ and $\beta_2(J_{\lambda^*}; \Pi_n) < +\infty$, then $\hat\lambda = \lambda^*$.*

Proof. Since $\hat\lambda \in L$ and $\lambda^* \in L$, we have

$$f_{\hat\lambda}(X_j) = f_{\lambda^*}(X_j),\ j = 1, \dots, n$$

implying that $\|f_{\hat\lambda} - f_{\lambda^*}\|_{L_2(\Pi_n)} = 0$. On the other hand, since $\hat\lambda$ is a solution of (7.3), we have $\|\hat\lambda\|_{\ell_1} \le \|\lambda^*\|_{\ell_1}$ implying that

$$\sum_{j \notin J_{\lambda^*}} |\hat\lambda_j| \le \sum_{j \in J_{\lambda^*}} (|\hat\lambda_j| - |\lambda_j^*|) \le \sum_{j \in J_{\lambda^*}} |\hat\lambda_j - \lambda_j^*|.$$

Therefore, $\hat\lambda - \lambda^* \in C_{J_{\lambda^*}}$ and

$$\|\hat\lambda - \lambda^*\|_{\ell_1} \le 2 \sum_{j \in J_{\lambda^*}} |\hat\lambda_j - \lambda_j^*| \le 2\sqrt{d(\lambda^*)} \biggl(\sum_{j \in J_{\lambda^*}} |\hat\lambda_j - \lambda_j^*|^2 \biggr)^{1/2} \le$$

$$2\beta_2(J_{\lambda^*}; \Pi_n)\sqrt{d(\lambda^*)} \|f_{\hat\lambda} - f_{\lambda^*}\|_{L_2(\Pi_n)} = 0,$$

implying the result. □

In particular, it means that as soon as the restricted isometry condition holds for the empirical distribution Π_n for a sufficiently large d with a sufficiently small δ_d (to be more precise, as soon as $\delta_{3d}(\Pi_n) \le 1/8$), (7.3) provides a solution of the sparse recovery problem for any target vector λ^* such that $f_* = f_{\lambda^*}$ and $d(\lambda^*) \le d$. The restricted isometry condition for Π_n (which can be also viewed as a condition on the design matrix A) has been also referred to as *the uniform uncertainty principle (UUP)* (see, e.g., Candes and Tao [29]). It is computationally hard to check UUP for a given large design matrix

A. Moreover, it is hard to construct $n \times N$-matrices for which UUP holds. The main approach is based on using random matrices of special type and proving that for such matrices UUP holds for a sufficiently large d with a high probability. We will discuss below a slightly different approach in which it is assumed that the design points $X_1, \dots, X_n$ are i.i.d. with common distribution Π. It will be proved directly (without checking UUP for the random matrix A) that under certain conditions (7.3) does provide a solution of sparse recovery problem with a high probability.

In what follows, we frequently use Orlicz norms $\|\cdot\|_\psi$ for random variables, most often, with $\psi = \psi_1$, $\psi_1(x) := e^{|x|} - 1$ or $\psi = \psi_2$, $\psi_2(x) := e^{x^2} - 1$. For any convex nondecreasing function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with $\psi(0) = 0$, it is defined as

$$\|\eta\|_\psi := \inf\biggl\{C > 0 : \mathbb{E}\psi\biggl(\frac{|\eta|}{C}\biggr) \le 1\biggr\}$$

(see Ledoux and Talagrand [68], van der Vaart and Wellner [95], de la Pena and Giné [32]). If we want to emphasize the dependence of the Orlicz norms on the probability measure, we will write $\|\cdot\|_{L_\psi(\mathbb{P})}$ (similarly, $\|\cdot\|_{L_\psi(P)}$, $\|\cdot\|_{L_\psi(\Pi)}$, etc.)

Define

$$\Lambda_S := \biggl\{\lambda \in \mathbb{R}^N : C\beta(J_\lambda; \Pi) \max_{1\le k\le N} \|h_k(X)\|_{\psi_1} \sqrt{\frac{A\log N}{n}} \le 1/4\biggr\}.$$

We will interpret Λ_S as a set of "sparse" vectors. Note that in the case when the dictionary is $L_2(\Pi)$-orthonormal,

$$\beta(J; \Pi) \le \sqrt{\mathrm{card}(J)},$$

so, indeed, Λ_S consists of vectors with a sufficiently small $d(\lambda)$ (or, sparse).

Theorem 7.2 *There exists a constant C in the definition of the set Λ_S such that for all $A \ge 1$ with probability at least $1 - N^{-A}$ $L \cap \Lambda_S = \{\hat\lambda\}$.*

The following lemma is used in the proof.

Lemma 7.4 *There exists a constant $C > 0$ such that for all $A \ge 1$ with probability at least $1 - N^{-A}$*

$$\sup_{\|u\|_{\ell_1}\le 1} \Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr| \le C \max_{1\le k\le N} \|h_k(X)\|_{\psi_1} \biggl(\sqrt{\frac{A\log N}{n}} \bigvee \frac{A\log N}{n}\biggr).$$

Proof. Let $R_n(f)$ be the Rademacher process. We will use symmetrization inequality and then contraction inequality for exponential moments (see sections 2.1, 2.2). For $t > 0$, we get

$$\mathbb{E}\exp\Bigl\{t \sup_{\|u\|_{\ell_1}\le 1}\Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr|\Bigr\} \le \mathbb{E}\exp\Bigl\{2t \sup_{\|u\|_{\ell_1}\le 1}\Bigl|R_n(|f_u|)\Bigr|\Bigr\} \le$$

$$\mathbb{E}\exp\Bigl\{4t \sup_{\|u\|_{\ell_1}\le 1}\Bigl|R_n(f_u)\Bigr|\Bigr\}.$$

Since the mapping $u \mapsto R_n(f_u)$ is linear, the supremum of $R_n(f_u)$ over the set $\{\|u\|_{\ell_1} \le 1\}$ is attained at one of its vertices, and we get

$$\mathbb{E}\exp\Bigl\{t \sup_{\|u\|_{\ell_1}\le 1}\Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr|\Bigr\} \le \mathbb{E}\exp\Bigl\{4t \max_{1\le k\le N}\Bigl|R_n(h_k)\Bigr|\Bigr\} =$$

$$N \max_{1\le k\le N} \mathbb{E}\Bigl[\exp\Bigl\{4tR_n(h_k)\Bigr\} \bigvee \exp\Bigl\{-4tR_n(h_k)\Bigr\}\Bigr] \le$$

$$2N \max_{1\le k\le N} \mathbb{E}\exp\Bigl\{4tR_n(h_k)\Bigr\} \le 2N \max_{1\le k\le N}\Bigl(\mathbb{E}\exp\Bigl\{4\frac{t}{n}\varepsilon h_k(X)\Bigr\}\Bigr)^n.$$

To bound the last expectation and to complete the proof, follow the standard proof of Bernstein's inequality.

□

Proof of Theorem 7.2. Arguing as in the proof of Proposition 7.2, we get that, for all $\lambda \in L$, $\hat\lambda - \lambda \in C_{J_\lambda}$ and $\|f_{\hat\lambda} - f_\lambda\|_{L_2(\Pi_n)} = 0$. Therefore,

$$\|\hat\lambda - \lambda\|_{\ell_1} \le \sum_{j\notin J_\lambda} |\hat\lambda_j| + \sum_{j\in J_\lambda} |\lambda_j - \hat\lambda_j| \le 2\sum_{j\in J_\lambda} |\lambda_j - \hat\lambda_j| \le 2\beta(J_\lambda)\|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi)}. \quad (7.6)$$

We will now upper bound $\|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi)}$ in terms of $\|\hat\lambda - \lambda\|_{\ell_1}$, which will imply the result. First, note that

$$\|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi)} = \|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi_n)} + (\Pi - \Pi_n)(|f_{\hat\lambda} - f_\lambda|) \le$$
$$\sup_{\|u\|_{\ell_1}\le 1}\Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr|\|\hat\lambda - \lambda\|_{\ell_1}. \quad (7.7)$$

By Lemma 7.4, with probability at least $1 - N^{-A}$ (under the assumption $A\log N \le n$)

$$\sup_{\|u\|_{\ell_1}\le 1}\Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr| \le C\max_{1\le k\le N}\|h_k\|_{\psi_1}\sqrt{\frac{A\log N}{n}}.$$

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

$$\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi)} \leq C \max_{1\leq k\leq N} \|h_k\|_{\psi_1}\sqrt{\frac{A\log N}{n}}\|\hat{\lambda} - \lambda\|_{\ell_1}. \tag{7.8}$$

Together with (7.6), this implies

$$\|\hat{\lambda} - \lambda\|_{\ell_1} \leq 2C \max_{1\leq k\leq N} \|h_k\|_{\psi_1}\sqrt{\frac{A\log N}{n}}\beta(J_\lambda)\|\hat{\lambda} - \lambda\|_{\ell_1}.$$

It follows that, for $\lambda \in L \cap \Lambda_S$, with probability at least $1 - N^{-A}$,

$$\|\hat{\lambda} - \lambda\|_{\ell_1} \leq \frac{1}{2}\|\hat{\lambda} - \lambda\|_{\ell_1},$$

and, hence, $\hat{\lambda} = \lambda$.

□

It is of interest to study the problem under the following condition on the dictionary and on the distribution Π : for all $\lambda \in C_J$

$$\|\sum_{j=1}^{N} \lambda_j h_j\|_{L_1(\Pi)} \leq \|\sum_{j=1}^{N} \lambda_j h_j\|_{L_2(\Pi)} \leq B(J)\|\sum_{j=1}^{N} \lambda_j h_j\|_{L_1(\Pi)} \tag{7.9}$$

with some constant $B(J) > 0$. This inequality always holds with some $B(J) > 0$ since any two norms on a finite dimensional space are equivalent. In fact, the first bound is just Cauchy-Schwarz inequality. However, in general, the constant $B(J)$ does depend on J and we are interested in the situation when there is no such dependence (or, at least, $B(J)$ does not grow too fast as $\mathrm{card}(J) \to \infty$). A canonical example in which (7.9) holds, for all $\lambda \in \mathbb{R}^N$ with $B(J) = B$ that does not depend on the dimension J, is when $(h_1(X), \ldots, h_N(X))$ has a normal distribution in $\mathbb{R}^N$, for instance, if $h_1(X), \ldots, h_N(X)$ are i.i.d. standard normal, which is the case of *Gaussian dictionary.* Another example is when $h_1(X), \ldots, h_N(X)$ are i.i.d. Rademacher random variables, i.e., $h_j(X)$ is $+1$ or -1 with probability 1/2 each. Such a dictionary is called *Bernoulli or Rademacher* and, in this case, (7.9) follows from Khinchin inequality. For Gaussian and Bernoulli dictionaries, all L_p norms, $p \geq 1$, and even ψ_1- and ψ_2-norms of $\sum_{j=1}^N \lambda_j h_j$ are equivalent up to numerical constants (see Bobkov and Houdre (1997) for a discussion of Khinchin type inequalities and their connections with isoperimetric constants).

Under the condition (7.9),

$$\beta(J) \leq B(J)\beta_2(J)\sqrt{d(J)}. \tag{7.10}$$

If $\beta_2(J)$ is bounded (as in the case of orthonormal dictionaries), then $\beta(J)$ is "small" for sets J of small cardinality $d(J)$. In this case, the definition of the set of "sparse vectors" Λ_S can be rewritten in terms of β_2.

However, we will give below another version of this result slightly improving the logarithmic factor in the definition of the set of sparse vectors Λ_S and providing bounds on the norms $\|\cdot\|_{L_2(\Pi)}$ and $\|\cdot\|_{\ell_2}$.

Denote

$$\beta_2(d) := \beta_2(d;\Pi) := \max\Big\{\beta_2(J) :\ J \subset \{1,\dots,N\},\ d(J) \le 2d\Big\}.$$

Let

$$B(d) := \max\left\{ B(J) :\ J \subset \{1,\dots,N\},\ d(J) \le d \right\}.$$

Finally, denote $\bar{d}$ the largest d satisfying the conditions $d \le \frac{N}{e} - 1$, $\frac{Ad\log(N/d)}{n} \le 1$, and

$$CB(d)\beta_2(d) \sup_{\|u\|_{\ell_2}\le 1, d(u)\le d} \|f_u\|_{\psi_1}\sqrt{\frac{Ad\log(N/d)}{n}} \le 1/4.$$

We will now use the following definition of the set of "sparse" vectors:

$$\Lambda_{S,2} := \{\lambda \in \mathbb{R}^N : d(\lambda) \le \bar{d}\}.$$

Recall the notation

$$\binom{n}{\le k} := \sum_{j=0}^{k}\binom{n}{j}.$$

Theorem 7.3 *Suppose condition (7.9) holds. There exists a constant C in the definition of $\Lambda_{S,2}$ such that for all $A \ge 1$ with probability at least*

$$1 - 5^{-\bar{d}A}\binom{N}{\le \bar{d}}^{-A}$$

the following equality holds: $L \cap \Lambda_{S,2} = \{\hat{\lambda}\}$.

We will use the following lemma.

Lemma 7.5 *For $J \subset \{1,\dots,N\}$ with $d(J) = d$, recall the following notation of Lemma 7.1 (with $b = 1$):*

$$K_J := C_J \cap \Big\{u \in \mathbb{R}^N :\ \|u\|_{\ell_2} \le 1\Big\}.$$

There exists a constant $C > 0$ such that for all $A \geq 1$ with probability at least

$$1 - 5^{-dA} \binom{N}{\leq d}^{-A}$$

the following bound holds:

$$\sup_{u \in K_J} \Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr| \leq C \sup_{\|u\|_{\ell_2} \leq 1, d(u) \leq d} \|f_u\|_{\psi_1} \biggl(\sqrt{\frac{Ad\log(N/d)}{n}} \bigvee \frac{Ad\log(N/d)}{n}\biggr).$$

Proof. It follows from statement (ii) of Lemma 7.1 with $b = 1$ that

$$K_J \subset 6 \text{ conv}(\mathcal{M}_d)$$

where $\mathcal{M}_d$ is a set of vectors u from the unit ball $\{u \in \mathbb{R}^N : \|u\|_{\ell_2} \leq 1\}$ such that $d(u) \leq d$ and

$$\text{card}(\mathcal{M}_d) \leq 5^d \binom{N}{\leq d}.$$

Now, it is enough to repeat the proof of Lemma 7.4. Bounding of

$$\sup_{u \in K_J} \Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr|$$

is reduced to bounding of

$$\sup_{u \in \mathcal{M}_d} |R_n(f_u)|,$$

$\text{card}(\mathcal{M}_d)$ playing now the role of N. The bound on $\text{card}(\mathcal{M}_d)$ implies that with some $c > 0$

$$\log(\text{card}(\mathcal{M}_d)) \leq cd \log \frac{N}{d}.$$

The proof is now complete.

□

Proof of Theorem 7.3 is a straightforward modification of the proof of Theorem 7.2. Let $\lambda \in L \cap \Lambda_{S,2}$. Instead of (7.7), we use

$$\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi)} = \|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi_n)} + (\Pi - \Pi_n)(|f_{\hat{\lambda}} - f_\lambda|) \leq \sup_{\|u\|_{\ell_2} \leq 1, u \in C_{J_\lambda}} \Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr| \|\hat{\lambda} - \lambda\|_{\ell_2}. \tag{7.11}$$

To bound $\|\hat{\lambda} - \lambda\|_{\ell_2}$ note that, as in the proof of Theorem 7.2, $\hat{\lambda} - \lambda \in C_{J_\lambda}$ and apply Lemma 7.1 to $u = \hat{\lambda} - \lambda$, $J = J_\lambda$:

$$\|\hat{\lambda} - \lambda\|_{\ell_2} \leq 2\biggl(\sum_{j \in J_0 \cup J_1} |\hat{\lambda}_j - \lambda_j|^2\biggr)^{1/2} \leq 2\beta_2(d(\lambda))\|f_{\hat{\lambda}} - f_\lambda\|_{L_2(\Pi)}. \tag{7.12}$$

Use Lemma 7.5 to bound

$$\sup_{\|u\|_{\ell_2}\le 1, u\in C_{J_\lambda}} \Big|(\Pi_n - \Pi)(|f_u|)\Big| \le C \sup_{\|u\|_{\ell_2}\le 1, d(u)\le d(\lambda)} \|f_u\|_{\psi_1}\sqrt{\frac{Ad(\lambda)\log(N/d(\lambda))}{n}}, \quad (7.13)$$

which holds with probability at least $1 - 5^{-d(\lambda)A}\binom{N}{\le d(\lambda)}^{-A}$. Then we use

$$\|\hat{\lambda} - \lambda\|_{\ell_1} \le 2\sum_{j\in J}|\hat{\lambda}_j - \lambda_j| \le 2\sqrt{d(\lambda)}\Big(\sum_{j\in J\cup J_1}|\hat{\lambda}_j - \lambda_j|^2\Big)^{1/2} \le$$
$$2\beta_2(d(\lambda))\sqrt{d(\lambda)}\|f_{\hat{\lambda}} - f_\lambda\|_{L_2(\Pi)}. \quad (7.14)$$

It remains to substitute bounds (7.12), (7.13) and (7.14) in (7.11), to use (7.9) and to solve the resulting inequality with respect to $\|f_{\hat{\lambda}} - f_\lambda\|_{L_2(\Pi)}$. It follows that the last norm is equal to 0. In view of (7.14), this implies that $\hat{\lambda} = \lambda$.

□

7.4 The Dantzig Selector

We now turn to the case when the target function f_* is observed in an additive noise. Moreover, it will not be assumed that f_* belongs to the linear span of the dictionary, this function will be rather approximated in the linear span.

Consider the following regression model with random design

$$Y_j = f_*(X_j) + \xi_j,\ j = 1, \dots, n,$$

where $X, X_1, \dots, X_n$ are i.i.d. random variables in a measurable space $(S, \mathcal{A})$ with distribution Π and $\xi, \xi_1, \dots, \xi_n$ are i.i.d. random variables with $\mathbb{E}\xi = 0$ independent of $(X_1, \dots, X_n)$. Candes and Tao [29] developed a method of sparse recovery based on linear programming suitable in this more general framework. They called it *the Dantzig Selector.*

Given $\varepsilon > 0$, let

$$\hat{\Lambda}_\varepsilon := \Big\{\lambda \in \mathbb{R}^N : \max_{1\le k\le N}\Big|n^{-1}\sum_{j=1}^{n}(f_\lambda(X_j) - Y_j)h_k(X_j)\Big| \le \varepsilon\Big\}$$

and define the Dantzig Selector as

$$\hat{\lambda} := \hat{\lambda}^\varepsilon \in \mathrm{Argmin}_{\lambda\in\hat{\Lambda}_\varepsilon}\|\lambda\|_{\ell_1}.$$

It is easy to reduce the computation of $\hat{\lambda}^{\varepsilon}$ to linear programming. The Dantzig Selector is closely related to the ℓ_1-penalization method (called "LASSO" in statistical literature, see Tibshirani [89]) and defined as a solution of the following penalized empirical risk minimization problem:

$$n^{-1}\sum_{j=1}^{n}(f_\lambda(X_j)-Y_j)^2+2\varepsilon\|\lambda\|_{\ell_1}=:L_n(\lambda)+2\varepsilon\|\lambda\|_{\ell_1}\longrightarrow \min\,. \tag{7.15}$$

The set of constraints of the Dantzig Selector can be written as

$$\hat{\Lambda}_\varepsilon=\Big\{\lambda:\Big\|\nabla L_n(\lambda)\Big\|_{\ell_\infty}\le\varepsilon\Big\}$$

and the condition $\lambda\in\hat{\Lambda}_\varepsilon$ is necessary for λ to be a solution of (7.15).

Candes and Tao studied in [29] the performance of the Dantzig Selector in the case of fixed design regression (nonrandom points $X_1,\dots,X_n$) under the assumption that the design matrix $A=\Big(h_j(X_i)\Big)_{i=1,n;j=1,N}$ satisfies the uniform uncertainty principle (UUP). They stated that UUP holds with a high probability for some random design matrices such as "Gaussian ensemble" (the matrix with i.i.d. standard normal entries) and "Bernoulli or Rademacher ensemble" (the matrix with i.i.d. entries taking values $+1$ and -1 with probability $1/2$), so, their results imply oracle inequalities for special $L_2(\Pi)$-orthonormal dictionaries.

We will prove several "sparsity oracle inequalities" for the Dantzig selector in spirit of recent results of Bunea, Tsybakov and Wegkamp [25], van de Geer [46], Koltchinskii [65] in the case of ℓ_1- or ℓ_p-penalized empirical risk minimization. We follow the paper of Koltchinskii [66] that relies only on elementary empirical and Rademacher processes methods (symmetrization and contraction inequalities for Rademacher processes and Bernstein type exponential bounds), but does not use more advanced techniques, such as concentration of measure and generic chaining. It is also close to the approach of Section 7.3 and to recent papers by Rudelson and Vershynin [82] and Mendelson, Pajor and Tomczak-Jaegermann [78]. As in Section 7.3, the proofs of oracle inequalities in the random design case given are more direct, they are not based on a reduction to the fixed design case and checking UUP for random matrices. The results also cover broader families of design distributions. In particular, the assumption that the dictionary is $L_2(\Pi)$-orthonormal is replaced by the assumption that the dictionary satisfies the restricted isometry condition with respect to Π. In what follows, the values of $\varepsilon>0$, $A>0$ and $C>0$ will be fixed and it will be assumed that $\frac{A\log N}{n}\le 1$. We will need the following set

$$\Lambda := \Lambda_\varepsilon(A) := \biggl\{\lambda \in \mathbb{R}^N : \Bigl|\langle f_\lambda - f_*, h_k\rangle_{L_2(\Pi)}\Bigr| +$$

$$C\Bigl(\|(f_\lambda - f_*)(X)h_k(X)\|_{\psi_1} + \|\xi h_k(X)\|_{\psi_1}\Bigr)\sqrt{\frac{A\log N}{n}} \le \varepsilon,\ k = 1,\dots,N\biggr\},$$

consisting of vectors λ ("oracles") such that f_λ provides a good approximation of f_*. In fact, $\lambda \in \Lambda_\varepsilon(A)$ implies that

$$\max_{1\le k\le N}\Bigl|\langle f_\lambda - f_*, h_k\rangle_{L_2(\Pi)}\Bigr| \le \varepsilon. \tag{7.16}$$

This means that $f_\lambda - f_*$ is “almost orthogonal” to the linear span of the dictionary. Thus, f_λ is close to the projection of f_* on the linear span. Condition (7.16) is necessary for λ to be a minimal point of

$$\lambda \mapsto \|f_\lambda - f_*\|^2_{L_2(\Pi)} + 2\varepsilon\|\lambda\|_{\ell_1},$$

and minimizing the last function is a "population version" of LASSO problem (7.15) ($\lambda \in \hat\Lambda_\varepsilon$ is a necessary condition for (7.15)). Of course, the condition

$$\varepsilon \ge \max_{1\le k\le N}\|\xi h_k(X)\|_{\psi_1}\sqrt{\frac{A\log N}{n}}$$

is necessary for $\Lambda_\varepsilon(A) \neq \emptyset$. It will be contained in the proof of Theorem 7.4 below that $\lambda \in \Lambda_\varepsilon(A)$ implies $\lambda \in \hat\Lambda_\varepsilon$ with a high probability.

The next theorems 7.4 and 7.5 show that if there exists a sufficiently sparse vector λ in the set $\hat\Lambda_\varepsilon$ of constraints of the Dantzig selector, then, with a high probability, the Dantzig selector belongs to a small ball around λ in such norms as $\|\cdot\|_{\ell_1}, \|\cdot\|_{\ell_2}$. At the same time, the function $f_{\hat\lambda}$ belongs to a small ball around f_λ with respect to such norms as $\|\cdot\|_{L_1(\Pi)}$ or $\|\cdot\|_{L_2(\Pi)}$. The radius of this ball is determined by the degree of sparsity of λ and by the properties of the dictionary characterized by such quantities as β or β_2 (see Section 7.2 for the definitions of these quantities and their connection to restricted isometry condition). Essentially, the results show that the Dantzig selector is adaptive to unknown degree of sparsity of the problem, provided that the dictionary is not too far from being orthonormal in $L_2(\Pi)$.

Recall the definition of the set of "sparse" vectors Λ_S in the previous section. Let

$$\tilde\Lambda = \tilde\Lambda_\varepsilon(A) := \Lambda_\varepsilon(A) \cap \Lambda_S.$$

Theorem 7.4 *There exists a constant C in the definitions of $\Lambda_\varepsilon(A), \Lambda_S$ such that for all $A \geq 1$ with probability at least $1 - N^{-A}$ the following bounds hold for all $\lambda \in \hat{\Lambda}_\varepsilon \cap \Lambda_S$*

$$\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi)} \leq 16\beta(J_\lambda)\varepsilon$$

and

$$\|\hat{\lambda} - \lambda\|_{\ell_1} \leq 32\beta^2(J_\lambda)\varepsilon.$$

This implies that

$$\|f_{\hat{\lambda}} - f_*\|_{L_1(\Pi)} \leq \inf_{\lambda \in \tilde{\Lambda}_\varepsilon(A)} \biggl[\|f_\lambda - f_*\|_{L_1(\Pi)} + 16\beta(J_\lambda)\varepsilon\biggr]$$

and, if in addition $f_* = f_{\lambda^*}, \lambda^* \in \mathbb{R}^N$, then also

$$\|\hat{\lambda} - \lambda^*\|_{\ell_1} \leq \inf_{\lambda \in \tilde{\Lambda}_\varepsilon(A)} \biggl[\|\lambda - \lambda^*\|_{\ell_1} + 32\beta^2(J_\lambda)\varepsilon\biggr].$$

We use the following lemma based on Bernstein's inequality (see, e.g., Lemma 2.2.11 in van der Vaart and Wellner [95]).

Lemma 7.6 *Let $\eta^{(k)}, \eta_1^{(k)}, \dots, \eta_n^{(k)}$ be i.i.d. random variables with $\mathbb{E}\eta^{(k)} = 0$ and $\|\eta^{(k)}\|_{\psi_1} < +\infty$, $k = 1, \dots, N$. There exists a numerical constant $C > 0$ such that for all $A \geq 1$ with probability at least $1 - N^{-A}$ for all $k = 1, \dots, N$*

$$\biggl|n^{-1}\sum_{j=1}^n \eta_j^{(k)}\biggr| \leq C\|\eta^{(k)}\|_{\psi_1}\biggl(\sqrt{\frac{A\log N}{n}} \bigvee \frac{A\log N}{n}\biggr).$$

Proof of Theorem 7.4. For $\lambda \in \hat{\Lambda}_\varepsilon \cap \Lambda_S$, we will upper bound the norms $\|\hat{\lambda} - \lambda\|_{\ell_1}$, $\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi)}$ in terms of each other and solve the resulting inequalities, which will yield the first two bounds of the theorem. As in the proof of Proposition 7.2 and theorems 7.2, 7.3, $\lambda \in \hat{\Lambda}_\varepsilon$ and the definition of $\hat{\lambda}$ imply that $\hat{\lambda} - \lambda \in C_{J_\lambda}$ and

$$\|\hat{\lambda} - \lambda\|_{\ell_1} \leq 2\beta(J_\lambda)\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi)}. \tag{7.17}$$

It remains to upper bound $\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi)}$ in terms of $\|\hat{\lambda} - \lambda\|_{\ell_1}$. To this end, note that

$$\begin{aligned} &\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi)} = \|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi_n)} + (\Pi - \Pi_n)(|f_{\hat{\lambda}} - f_\lambda|) \leq \\ &\|f_{\hat{\lambda}} - f_\lambda\|_{L_1(\Pi_n)} + \sup_{\|u\|_{\ell_1} \leq 1}\Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr|\|\hat{\lambda} - \lambda\|_{\ell_1}. \end{aligned} \tag{7.18}$$

The first term in the right hand side can be bounded as follows

$$\|f_{\hat\lambda}-f_\lambda\|^2_{L_1(\Pi_n)}\le \|f_{\hat\lambda}-f_\lambda\|^2_{L_2(\Pi_n)}=\langle f_{\hat\lambda}-f_\lambda, f_{\hat\lambda}-f_\lambda\rangle_{L_2(\Pi_n)}=$$

$$\sum_{k=1}^N(\hat\lambda_k-\lambda_k)\langle f_{\hat\lambda}-f_\lambda, h_k\rangle_{L_2(\Pi_n)}\le \|\hat\lambda-\lambda\|_{\ell_1}\max_{1\le k\le N}\Bigl|\langle f_{\hat\lambda}-f_\lambda, h_k\rangle_{L_2(\Pi_n)}\Bigr|.$$

Both $\hat\lambda\in\hat\Lambda$ and $\lambda\in\hat\Lambda$, implying that

$$\max_{1\le k\le N}\Bigl|\langle f_{\hat\lambda}-f_\lambda, h_k\rangle_{L_2(\Pi_n)}\Bigr|\le$$

$$\max_{1\le k\le N}\Bigl|n^{-1}\sum_{j=1}^n(f_\lambda(X_j)-Y_j)h_k(X_j)\Bigr|+\max_{1\le k\le N}\Bigl|n^{-1}\sum_{j=1}^n(f_{\hat\lambda}(X_j)-Y_j)h_k(X_j)\Bigr|\le 2\varepsilon.$$

Therefore,

$$\|f_{\hat\lambda}-f_\lambda\|_{L_1(\Pi_n)}\le\sqrt{2\varepsilon\|\hat\lambda-\lambda\|_{\ell_1}}.$$

Now we bound the second term in the right hand side of (7.18). Under the assumption $A\log N\le n$, Lemma 7.4 implies that with probability at least $1-N^{-A}$

$$\sup_{\|u\|_{\ell_1}\le 1}\Bigl|(\Pi_n-\Pi)(|f_u|)\Bigr|\le C\max_{1\le k\le N}\|h_k\|_{\psi_1}\sqrt{\frac{A\log N}{n}}.$$

Hence, we conclude from (7.18) that

$$\|f_{\hat\lambda}-f_\lambda\|_{L_1(\Pi)}\le\sqrt{2\varepsilon\|\hat\lambda-\lambda\|_{\ell_1}}+C\max_{1\le k\le N}\|h_k\|_{\psi_1}\sqrt{\frac{A\log N}{n}}\|\hat\lambda-\lambda\|_{\ell_1}. \tag{7.19}$$

Combining this with (7.17) yields

$$\|f_{\hat\lambda}-f_\lambda\|_{L_1(\Pi)}\le\sqrt{4\varepsilon\beta(J_\lambda)\|f_{\hat\lambda}-f_\lambda\|_{L_1(\Pi)}}+2C\max_{1\le k\le N}\|h_k\|_{\psi_1}\sqrt{\frac{A\log N}{n}}\beta(J_\lambda)\|f_{\hat\lambda}-f_\lambda\|_{L_1(\Pi)}.$$

By the definition of Λ_S,

$$2C\max_{1\le k\le N}\|h_k\|_{\psi_1}\sqrt{\frac{A\log N}{n}}\beta(J_\lambda)\le 1/2,$$

so, we end up with

$$\|f_{\hat\lambda}-f_\lambda\|_{L_1(\Pi)}\le 2\sqrt{4\varepsilon\beta(J_\lambda)\|f_{\hat\lambda}-f_\lambda\|_{L_1(\Pi)}},$$

which implies the first bound of the theorem. The second bound holds because of (7.17).

Observe that for all $\lambda \in \Lambda$,

$$\left| n^{-1} \sum_{j=1}^{n} (f_\lambda(X_j) - Y_j) h_k(X_j) \right| \leq \left| \langle f_\lambda - f_*, h_k \rangle_{L_2(\Pi)} \right| +$$

$$\left| n^{-1} \sum_{j=1}^{n} \Big[(f_\lambda(X_j) - f_*(X_j)) h_k(X_j) - \mathbb{E}(f_\lambda(X) - f_*(X)) h_k(X) \Big] \right| + \left| n^{-1} \sum_{j=1}^{n} \xi_j h_k(X_j) \right|.$$

Lemma 7.6 can be used to bound the second and the third terms: with probability at least $1 - 2N^{-A}$

$$\max_{1 \leq k \leq N} \left| n^{-1} \sum_{j=1}^{n} (f_\lambda(X_j) - Y_j) h_k(X_j) \right| \leq \max_{1 \leq k \leq N} \Bigg[\left| \langle f_\lambda - f_*, h_k \rangle_{L_2(\Pi)} \right| +$$

$$C \Big(\|(f_\lambda - f_*)(X) h_k(X)\|_{\psi_1} + \|\xi h_k(X)\|_{\psi_1} \Big) \sqrt{\frac{A \log N}{n}} \Bigg] \leq \varepsilon.$$

This proves that for all $\lambda \in \Lambda$, with probability at least $1 - 2N^{-A}$, we also have $\lambda \in \hat{\Lambda}$.

For each of the remaining bounds, let $\bar{\lambda}$ be the vector for which the infimum in the right hand side of the bound is attained. With probability at least $1 - 2N^{-A}$, $\bar{\lambda} \in \hat{\Lambda}_\varepsilon \cap \Lambda_S$. Hence, it is enough to use the first two bounds of the theorem and the triangle inequality to finish the proof.

□

We will give another result about the Dantzig selector in which the properties of the dictionary are characterized by the quantity β_2 instead of β. Recall the definition of the set of "sparse" vectors $\Lambda_{S,2}$ from the previous section and define

$$\tilde{\Lambda}^2 = \tilde{\Lambda}^2_\varepsilon(A) := \Lambda_\varepsilon(A) \cap \Lambda_{S,2}.$$

Theorem 7.5 *Suppose condition (7.9) holds. There exists a constant C in the definitions of $\Lambda_\varepsilon(A), \Lambda_{S,2}$ such that for all $A \geq 1$ with probability at least*

$$1 - 5^{-\bar{d}A} \binom{N}{\leq \bar{d}}^{-A}$$

the following bounds hold for all $\lambda \in \hat{\Lambda}_\varepsilon \cap \Lambda_{S,2}$

$$\|f_{\hat{\lambda}} - f_\lambda\|_{L_2(\Pi)} \leq 16 B^2 \beta_2(d(\lambda)) \sqrt{d(\lambda)} \varepsilon$$

and

$$\|\hat{\lambda} - \lambda\|_{\ell_2} \leq 32 B^2 \beta_2^2(d(\lambda)) \sqrt{d(\lambda)} \varepsilon.$$

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

$$\|f_{\hat\lambda} - f_*\|_{L_2(\Pi)} \le \inf_{\lambda \in \tilde\Lambda_\varepsilon^2(A)} \Bigl[\|f_\lambda - f_*\|_{L_2(\Pi)} + 16B^2\beta_2(d(\lambda))\sqrt{d(\lambda)}\varepsilon\Bigr]$$

and, if $f_* = f_{\lambda^*}, \lambda^* \in \mathbb{R}^N$, *then*

$$\|\hat\lambda - \lambda^*\|_{\ell_2} \le \inf_{\lambda \in \tilde\Lambda_\varepsilon^2(A)} \Bigl[\|\lambda - \lambda^*\|_{\ell_2} + 32B^2\beta_2^2(d(\lambda))\sqrt{d(\lambda)}\varepsilon\Bigr].$$

Proof. We follow the proof of Theorem 7.4. For $\lambda \in \hat\Lambda_\varepsilon \cap \Lambda_{S,2}$, we use the following bound instead of (7.18):

$$\begin{aligned}&\|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi)} = \|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi_n)} + (\Pi - \Pi_n)(|f_{\hat\lambda} - f_\lambda|) \le \\ &\|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi_n)} + \sup_{\|u\|_{\ell_2}\le 1, u \in C_{J_\lambda}} \Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr| \|\hat\lambda - \lambda\|_{\ell_2}.\end{aligned} \tag{7.20}$$

Again, we have $\hat\lambda - \lambda \in C_{J_\lambda}$, and, using Lemma 7.1, we get for $u = \hat\lambda - \lambda$ and $J = J_\lambda$:

$$\|\hat\lambda - \lambda\|_{\ell_2} \le 2\Bigl(\sum_{j \in J_0 \cup J_1} |\hat\lambda_j - \lambda_j|^2\Bigr)^{1/2} \le 2\beta_2(d(\lambda))\|f_{\hat\lambda} - f_\lambda\|_{L_2(\Pi)}. \tag{7.21}$$

Lemma 7.5 now yields

$$\sup_{\|u\|_{\ell_2}\le 1, u \in C_{J_\lambda}} \Bigl|(\Pi_n - \Pi)(|f_u|)\Bigr| \le C \sup_{\|u\|_{\ell_2}\le 1, d(u) \le d(\lambda)} \|f_u\|_{\psi_1} \sqrt{\frac{Ad(\lambda)\log(N/d(\lambda))}{n}}, \tag{7.22}$$

which holds with probability at least

$$1 - 5^{-d(\lambda)A}\binom{N}{\le d(\lambda)}^{-A}.$$

As in the proof of Theorem 7.4, we bound the first term in the right hand side of (7.20):

$$\|f_{\hat\lambda} - f_\lambda\|_{L_1(\Pi_n)} \le \sqrt{2\varepsilon\|\hat\lambda - \lambda\|_{\ell_1}}. \tag{7.23}$$

In addition,

$$\begin{aligned}&\|\hat\lambda - \lambda\|_{\ell_1} \le 2\sum_{j \in J} |\hat\lambda_j - \lambda_j| \le 2\sqrt{d(\lambda)}\Bigl(\sum_{j \in J \cup J_1} |\hat\lambda_j - \lambda_j|^2\Bigr)^{1/2} \le \\ &2\beta_2(d(\lambda))\sqrt{d(\lambda)}\|f_{\hat\lambda} - f_\lambda\|_{L_2(\Pi)}.\end{aligned} \tag{7.24}$$

Substitute bounds (7.21), (7.22), (7.23) and (7.24) into (7.20), use (7.9) and solve the resulting inequality with respect to $\|f_{\hat{\lambda}} - f_\lambda\|_{L_2(\Pi)}$. This gives the first bound of the theorem.

The second bound follows from (7.21) and the remaining two bounds are proved exactly as in Theorem 7.4.

□

In the fixed design case, the following result holds. Its proof is a simplified version of the proofs of theorems 7.4, 7.5.

Theorem 7.6 *Suppose $X_1, \dots, X_n$ are nonrandom design points in S and let Π_n be the empirical measure based on $X_1, \dots, X_n$. Suppose also $f_* = f_{\lambda^*}$, $\lambda^* \in \mathbb{R}^N$. There exists a constant $C > 0$ such that for all $A \geq 1$ and for all*

$$\varepsilon \geq C\|\xi\|_{\psi_2} \max_{1\leq k\leq N} \|h_k\|_{L_2(\Pi_n)} \sqrt{\frac{A\log N}{n}},$$

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

$$\|f_{\hat{\lambda}} - f_{\lambda^*}\|_{L_2(\Pi_n)} \leq 4\beta_2(J_{\lambda^*}, \Pi_n)\sqrt{d(\lambda^*)}\varepsilon,$$

$$\|\hat{\lambda} - \lambda^*\|_{\ell_1} \leq 8\beta_2^2(J_{\lambda^*}, \Pi_n)d(\lambda^*)\varepsilon$$

and

$$\|\hat{\lambda} - \lambda^*\|_{\ell_2} \leq 8\beta_2^2(d(\lambda^*), \Pi_n)\sqrt{d(\lambda^*)}\varepsilon.$$

Proof As in the proof of Theorem 7.4,

$$\|f_{\hat{\lambda}} - f_{\lambda^*}\|_{L_2(\Pi_n)} \leq \sqrt{2\varepsilon\|\hat{\lambda} - \lambda^*\|_{\ell_1}} \tag{7.25}$$

and

$$\|\hat{\lambda} - \lambda^*\|_{\ell_1} \leq 2\beta_2(J_{\lambda^*}, \Pi_n)\sqrt{d(\lambda^*)}\|f_{\hat{\lambda}} - f_{\lambda^*}\|_{L_2(\Pi_n)}. \tag{7.26}$$

These bounds hold provided that $\lambda^* \in \hat{\Lambda}_\varepsilon$, or

$$\max_{1\leq k\leq N} \left| n^{-1} \sum_{j=1}^{n} \xi_j h_k(X_j) \right| \leq \varepsilon.$$

If $\|\xi\|_{\psi_2} < +\infty$ and

$$\varepsilon \geq C\|\xi\|_{\psi_2} \max_{1\leq k\leq N} \|h_k\|_{L_2(\Pi_n)} \sqrt{\frac{A\log N}{n}},$$

then usual bounds for random variables in Orlicz spaces imply that $\lambda^* \in \hat{\Lambda}_\varepsilon$ with probability at least $1 - N^{-A}$.

Combining (7.25) and (7.26) shows that with probability at least $1 - N^{-A}$

$$\|f_{\hat{\lambda}} - f_{\lambda^*}\|_{L_2(\Pi_n)} \leq 4\beta_2(J_{\lambda^*}, \Pi_n)\sqrt{d(\lambda^*)}\varepsilon$$

and

$$\|\hat{\lambda} - \lambda^*\|_{\ell_1} \leq 8\beta_2^2(J_{\lambda^*}, \Pi_n)d(\lambda^*)\varepsilon.$$

Using Lemma 7.1 and arguing as in the proof of Theorem 7.5, we also get

$$\|\hat{\lambda} - \lambda^*\|_{\ell_2} \leq 8\beta_2^2(d(\lambda^*), \Pi_n)\sqrt{d(\lambda^*)}\varepsilon.$$

□

Bounding $\beta_2(J, \Pi_n)$ in terms of restricted isometry constants (see Lemma 7.2), essentially, allows one to recover Theorem 1 of Candes and Tao [29] that was the first result about the Dantzig selector in the fixed design case. Instead of doing this, we turn again to the case of random design regression and conclude this section with the derivation of the results of Candes and Tao [29] in the random design case.

To simplify the matter, we assume that the dictionary is orthonormal and that the following conditions hold with a numerical constant $B > 0$:

$$\frac{1}{B}\|\lambda\|_{\ell_2} \leq \Big\|\sum_{j=1}^N \lambda_j h_j\Big\|_{L_1(\Pi)} \leq B\|\lambda\|_{\ell_2}$$

and

$$\frac{1}{B}\|\lambda\|_{\ell_2} \leq \Big\|\sum_{j=1}^N \lambda_j h_j\Big\|_{L_{\psi_2}(\Pi)} \leq B\|\lambda\|_{\ell_2}, \quad \lambda \in \mathbb{R}^N.$$

This is the case, for instance, for Gaussian and Rademacher dictionaries. We also assume that the noise $\{\xi_j\}$ is a sequence of i.i.d. normal random variables with mean 0 and variance σ^2. Finally, assume that $f_* = f_{\lambda^*},\ \lambda^* \in \mathbb{R}^N$.

Under the last assumption, $\lambda^* \in \Lambda_\varepsilon(A)$ provided that

$$\varepsilon \geq C \max_{1\leq k\leq N} \|\xi h_k(X)\|_{\psi_1}\sqrt{\frac{A\log N}{n}}. \tag{7.27}$$

Moreover, for a normal random variable ξ with mean 0 and variance σ^2, $\|\xi\|_{\psi_2} = c_1\sigma$ for a numerical constant $c_1 > 0$. In addition, by the assumptions on the dictionary, the norms $\|h_k\|_{\psi_2},\ k=1,\dots,N$ are uniformly bounded by a numerical constant. Therefore, for a numerical constant $c_2>0$,

$$\|\xi h_k(X)\|_{\psi_1} \le \|\xi\|_{\psi_2}\|h_k(X)\|_{\psi_2} \le c_2\sigma,\ k=1,\dots,N,$$

and the condition on ε (7.27) reduces to the following:

$$\varepsilon \ge C\sigma\sqrt{\frac{A\log N}{n}} \tag{7.28}$$

with a proper numerical constant $C>0$.

The conditions on the dictionary also imply that $\beta_2(d)=1$ for all d and that the set $\Lambda_{S,2}$ includes all the vectors $\lambda\in\mathbb{R}^N$ such that, for a proper choice of numerical constant $C>0$, $C\sqrt{\frac{Ad(\lambda)\log N}{n}} \le 1/4$. Hence, if λ^* satisfies the condition

$$C\sqrt{\frac{Ad(\lambda^*)\log N}{n}} \le 1/4, \tag{7.29}$$

then $\lambda^*\in\tilde{\Lambda}_\varepsilon(A)$.

We now can derive the following corollary from the last bound of Theorem 7.5 (using λ^* as an oracle).

Corollary 7.1 *Under the above assumptions on the dictionary and on the noise and also the assumptions (7.28) and (7.29), the following bound holds with probability at least* $1-N^{-A}$:

$$\|\hat{\lambda}-\lambda^*\|_{\ell_2} \le D\sqrt{d(\lambda^*)}\varepsilon,$$

where $D>0$ *is a numerical constant.*

A version of another oracle inequality of Candes and Tao [29] also easily follows from Theorem 7.5.

Corollary 7.2 *Under the assumptions of Corollary 7.1, the following bound holds with probability at least* $1-N^{-A}$ *and with some numerical constant* $D>0$:

$$\|\hat{\lambda}-\lambda^*\|_{\ell_2}^2 \le D\sum_{j=1}^N (|\lambda_j^*|^2\wedge\varepsilon^2) = D\inf_{J\subset\{1,\dots,N\}}\Bigl[\sum_{j\notin J}|\lambda_j^*|^2 + d(J)\varepsilon^2\Bigr].$$

Proof. It is enough to choose $\bar{\lambda}^*$ as follows:

$$\bar{\lambda}^*_j = \lambda^*_j I(|\lambda^*_j| \geq \varepsilon/3),\ \ j = 1, \dots, N.$$

Then

$$\Big|\langle f_{\bar{\lambda}^*} - f_{\lambda^*}, h_k \rangle_{L_2(\Pi)}\Big| = |\lambda^*_k| \leq \varepsilon/3$$

for all $k \in J_{\lambda^*}$ such that $|\lambda^*_k| \leq \varepsilon/3$. Otherwise,

$$\Big|\langle f_{\bar{\lambda}^*} - f_{\lambda^*}, h_k \rangle_{L_2(\Pi)}\Big| = |\lambda^*_k| = 0.$$

In addition, the assumption (7.28) on ε implies that, for some numerical constant C',

$$C' \|\xi h_k(X)\|_{\psi_1} \sqrt{\frac{A \log N}{n}} \leq \varepsilon/3,\ \ k = 1, \dots, N.$$

We also have that with some numerical constant $c > 0$

$$\|(f_{\bar{\lambda}^*} - f_{\lambda^*})(X) h_k(X)\|_{\psi_1} \leq \|(f_{\bar{\lambda}^*} - f_{\lambda^*})(X)\|_{\psi_2} \|h_k(X)\|_{\psi_2}$$

$$\leq c \biggl(\sum_{j : |\lambda^*_j| < \varepsilon/3} |\lambda^*_j|^2 \biggr)^{1/2} \leq c(\varepsilon/3)\sqrt{d(\lambda^*)}.$$

Therefore,

$$C' \|(f_{\bar{\lambda}^*} - f_{\lambda^*})(X) h_k(X)\|_{\psi_1} \sqrt{\frac{A \log N}{n}} \leq \varepsilon/3$$

as soon as $cC' \sqrt{\frac{A d(\lambda^*) \log N}{n}} \leq 1$. The last condition follows from (7.29) with a properly chosen constant. Thus, $\bar{\lambda}^* \in \bar{\Lambda}_\varepsilon(A)$ (again, with a proper choice of constants in the definition of this set). By Theorem 7.5, with probability at least $1 - N^{-A}$

$$\|\hat{\lambda} - \lambda^*\|_{\ell_2} \leq \biggl(\sum_{j : |\lambda^*_j| < \varepsilon/3} |\lambda^*_j|^2 \biggr)^{1/2} + D\sqrt{\mathrm{card}(j : |\lambda^*_j| \geq \varepsilon/3)}\varepsilon,$$

which yields that, for some constant D,

$$\|\hat{\lambda} - \lambda^*\|^2_{\ell_2} \leq D \sum_{j=1}^N (|\lambda^*_j|^2 \wedge \varepsilon^2).$$

□

8 Convex Penalization in Sparse Recovery: ℓ_1-Penalization

8.1 General Aspects of Convex Penalization

In this and in the next section we study an approach to sparse recovery based on penalized empirical risk minimization of the following form

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in D}\Bigl[P_n(\ell \bullet f_\lambda) + \varepsilon \sum_{j=1}^{N} \psi(\lambda_j)\Bigr]. \tag{8.1}$$

We use the notations of Section 1.6. As before, it is assumed that

$$f_\lambda := \sum_{j=1}^{N} \lambda_j h_j,\ \lambda \in \mathbb{R}^N,$$

where

$$\mathcal{H} := \{h_1, \dots, h_N\}$$

is a given finite dictionary of measurable functions from S into $[-1, 1]$. The cardinality of the dictionary is usually very large (often, larger than the sample size n). We will assume in what follows that $N \geq (\log n)^\gamma$ for some $\gamma > 0$ (this is needed only to avoid additional terms of the order $\frac{\log\log n}{n}$ in several inequalities).

We will also assume that ψ is a convex even function and $\varepsilon \geq 0$ is a regularization parameter, and that $D \subset \mathbb{R}^N$ is a convex set.

The excess risk of f is defined as

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

where the infimum is taken over all measurable functions and it is assumed, for simplicity, that it is attained at $f_* \in L_2(\Pi)$ (moreover, it will be assumed in what follows that f_* is uniformly bounded by a constant M).

Definition 8.1 *It will be said that $\ell : T \times \mathbb{R} \mapsto \mathbb{R}_+$ is a loss function of* **quadratic type** *iff the following assumptions are satisfied:*
(i) for all $y \in T$, $\ell(y, \cdot)$ is convex;
(ii) for all $y \in T$, $\ell(y, \cdot)$ is twice differentiable, ℓ''_u is a uniformly bounded function in $T \times \mathbb{R}$ and

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

Moreover, denote

$$\tau(R) := \frac{1}{2} \inf_{y\in T} \inf_{|u|\leq R} \ell''_u(y,u). \tag{8.2}$$

Then

(iii) $\tau(M \vee 1) > 0$.

Recall that M *is a constant such that* $\|f_*\|_\infty \leq M$*). Without loss of generality, it will be also assumed that* $\tau(R) \leq 1, R > 0$ *(otherwise, it can be replaced by a lower bound).*

For losses of quadratic type, the following property is obvious:

$$\tau(\|f\|_\infty \vee M)\|f - f_*\|^2_{L_2(\Pi)} \leq \mathcal{E}(f) \leq C\|f - f_*\|^2_{L_2(\Pi)},$$

where C is a constant depending only on ℓ.

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

Note that the condition that the second derivative ℓ''_u is uniformly bounded in $T \times \mathbb{R}$ can be replaced by its uniform boundedness in $T \times [-M \vee 1, M \vee 1]$. The constants in the theorems below will then depend on the sup-norm of the second derivative (and, as a consequence, on M); otherwise, the results will be the same. This allows one to cover several other choices of the loss function, such as the exponential loss $\ell(y,u) := e^{-yu}$ in binary classification.

Clearly, the conditions that the loss ℓ, the penalty function ψ and the domain D are convex make the optimization problem (8.1) convex and, hence, computationally tractable (at least, in principle).

In the recent literature, there has been considerable attention to the problem of sparse recovery using LASSO type penalties, which is a special case of problem (8.1). In this case, $D = \mathbb{R}^N$, so this is a problem of sparse recovery in the linear span l.s.$(\mathcal{H})$ of the dictionary, and $\psi(u) = u$, which means penalization with ℓ_1-norm. It is also usually assumed that $\ell(y,u) = (y-u)^2$ (the case of regression with quadratic loss). In this setting, it has been shown that sparse recovery is possible not always, but only under some geometric assumptions on the dictionary. They are often expressed in terms of the Gram matrix of the dictionary, which in the case of random design models is the matrix

$$H := \biggl(\langle h_i, h_j\rangle_{L_2(\Pi)}\biggr)_{i,j=1,N},$$

and they take form of various conditions on the entries of this matrix (”coherence coefficients”), or on its submatrices (in spirit of ”uniform uncertainty principle” or ”restricted isometry” conditions, see Section 7.2). The essence of these assumptions is to try to keep the dictionary not too far from being orthonormal in $L_2(\Pi)$ which, in some sense, is an ideal case for sparse recovery (see, e.g., Donoho [35, 36, 37, 39], Candes and Tao [29], Rudelson and Vershynin [82], Mendelson, Pajor and Tomczak-Jaegermann [78], Bunea, Tsybakov and Wegkamp [25], van de Geer [46], Koltchinskii [64, 65, 66], Bickel, Ritov and Tsybakov [14] among many other papers that study both the random design and the fixed design problems).

We will study several special cases of problem (8.1). In the case $D = \mathbb{R}^N$, the most common choice of ψ is $\psi(u) = |u|$ which leads to ℓ_1- or LASSO-penalty. The same penalty can be used in some other cases, for instance, when $D = U_{\ell_1}$ (the unit ball of ℓ_1). This leads to a problem of sparse recovery in the symmetric convex hull

$$\mathrm{conv}_s(\mathcal{H}) := \biggl\{f_\lambda : \lambda \in U_{\ell_1}\biggr\},$$

which can be viewed as a version of convex aggregation problem. More generally, one can consider the case of $D = U_{\ell_p}$ (the unit ball in ℓ_p) with $p \geq 1$ and with $\psi(u) = |u|^p$ (i.e., the penalty becomes $\|\lambda\|_{\ell_p}^p$). It was shown in Koltchinskii [65] that sparse recovery is still possible if p is close enough to 1 (say, of the order $1 + 1/\log N$). Another interesting example is

$$D = \Lambda := \biggl\{\lambda \in \mathbb{R}^N : \lambda_j \geq 0, \sum_{j=1}^N \lambda_j = 1\biggr\},$$

so, D is the simplex of all probability distributions on $\{1,\dots,N\}$. This corresponds to the sparse recovery problem in the convex hull of the dictionary

$$\mathrm{conv}(\mathcal{H}) := \biggl\{f_\lambda : \lambda \in \Lambda\biggr\}.$$

In this case, it is natural to study the penalty

$$-H(\lambda) = \sum_{j=1}^{N} \lambda_j \log \lambda_j.$$

$H(\lambda)$ is the entropy of probability distribution λ; this corresponds to the choice $\psi(u) = u \log u$. Such a problem was studied in Koltchinskii [67].

We will follow the approach of [65, 67]. This approach is based on the analysis of necessary conditions of extremum in problem (8.1). For simplicity, consider the case of $D = \mathbb{R}^N$. In this case, for $\hat{\lambda}^\varepsilon$ to be a solution of (8.1), it is necessary that $0 \in \partial L_{n,\varepsilon}(\hat{\lambda}^\varepsilon)$, where

$$L_{n,\varepsilon}(\lambda) := P_n(\ell \bullet f_\lambda) + \varepsilon \sum_{j=1}^{N} \psi(\lambda_j)$$

and ∂ denotes the subdifferential of convex functions. If ψ is smooth, this leads to the equations

$$P_n(\ell' \bullet f_{\hat{\lambda}^\varepsilon}) h_j + \varepsilon \psi'(\hat{\lambda}_j^\varepsilon) = 0, \ j = 1, \ldots, N. \tag{8.3}$$

Define

$$L_\varepsilon(\lambda) := P(\ell \bullet f_\lambda) + \varepsilon \sum_{j=1}^{N} \psi(\lambda_j)$$

and

$$\nabla L_\varepsilon(\lambda) := \Big(P(\ell' \bullet f_\lambda) h_j + \varepsilon \psi'(\lambda_j)\Big)_{j=1,\ldots,N}.$$

The vector $\nabla L_\varepsilon(\lambda)$ is the gradient and the subgradient of the convex function $L_\varepsilon(\lambda)$ at point λ. It follows from (8.3) that

$$P_n(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_\lambda) + \varepsilon \sum_{j=1}^{N} \psi'(\hat{\lambda}_j^\varepsilon)(\hat{\lambda}_j^\varepsilon - \lambda_j) = 0$$

and we also have

$$P(\ell' \bullet f_\lambda)(f_{\hat{\lambda}^\varepsilon} - f_\lambda) + \varepsilon \sum_{j=1}^{N} \psi'(\lambda_j)(\hat{\lambda}_j^\varepsilon - \lambda_j) = \Big\langle \nabla L_\varepsilon(\lambda), \hat{\lambda}^\varepsilon - \lambda \Big\rangle_{\ell_2}.$$

Subtracting the second equation from the first yields the relationship

$$P(\ell' \bullet f_{\hat{\lambda}^\varepsilon} - \ell' \bullet f_\lambda)(f_{\hat{\lambda}^\varepsilon} - f_\lambda) + \varepsilon \sum_{j=1}^{N} (\psi'(\hat{\lambda}_j^\varepsilon) - \psi'(\lambda_j))(\hat{\lambda}_j^\varepsilon - \lambda_j) =$$

$$\left\langle \nabla L_\varepsilon(\lambda), \lambda - \hat{\lambda}^\varepsilon \right\rangle_{\ell_2} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_\lambda).$$

If ℓ is a loss of quadratic type, then

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

with some constant $c > 0$ depending only on ℓ and the following inequality holds

$$c\|f_{\hat{\lambda}^\varepsilon} - f_\lambda\|^2_{L_2(\Pi)} + \varepsilon \sum_{j=1}^N (\psi'(\hat{\lambda}^\varepsilon_j) - \psi'(\lambda_j))(\hat{\lambda}^\varepsilon_j - \lambda_j) \leq$$
$$\left\langle \nabla L_\varepsilon(\lambda), \lambda - \hat{\lambda}^\varepsilon \right\rangle_{\ell_2} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_\lambda). \tag{8.4}$$

Inequality (8.4) provides some information about "sparsity" of $\hat{\lambda}^\varepsilon$ in terms of "sparsity" of the oracle λ and it also provides tight bounds on $\|f_{\hat{\lambda}^\varepsilon} - f_\lambda\|_{L_2(\Pi)}$. Indeed, if $J = J_\lambda = \mathrm{supp}(\lambda)$ and $\psi'(0) = 0$ (which is the case, for instance, when $\psi(u) = u^p$ for some $p > 1$), then

$$\sum_{j=1}^N (\psi'(\hat{\lambda}^\varepsilon_j) - \psi'(\lambda_j))(\hat{\lambda}^\varepsilon_j - \lambda_j) \geq \sum_{j \notin J} \psi'(\hat{\lambda}^\varepsilon_j)\hat{\lambda}^\varepsilon_j = \sum_{j \notin J} |\psi'(\hat{\lambda}^\varepsilon_j)||\hat{\lambda}^\varepsilon_j|$$

(note that all the terms in the sum in the left hand side are nonnegative since ψ is convex and ψ' is nondecreasing). Thus, the following bound holds

$$c\|f_{\hat{\lambda}^\varepsilon} - f_\lambda\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\psi'(\hat{\lambda}^\varepsilon_j)||\hat{\lambda}^\varepsilon_j| \leq$$
$$\left\langle \nabla L_\varepsilon(\lambda), \lambda - \hat{\lambda}^\varepsilon \right\rangle_{\ell_2} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_\lambda), \tag{8.5}$$

in which the left hand side measures the L_2-distance of $f_{\hat{\lambda}^\varepsilon}$ from the oracle f_λ as well as the degree of sparsity of the empirical solution $\hat{\lambda}^\varepsilon$. This inequality will be applied to sparse vectors λ ("oracles") such that the term $\left\langle \nabla L_\varepsilon(\lambda), \lambda - \hat{\lambda}^\varepsilon \right\rangle_{\ell_2}$ is either negative, or positive, but small enough. This is the case, for instance, when the subgradient $\nabla L_\varepsilon(\lambda)$ is small in certain sense. In such cases, the left hand side is controlled by the empirical process

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

It happens that its size in turn depends on the L_2-distance $\|f_{\hat{\lambda}^\varepsilon} - f_\lambda\|_{L_2(\Pi)}$ and the measure of "sparsity" $\sum_{j \notin J} |\psi'(\hat{\lambda}^\varepsilon_j)||\hat{\lambda}^\varepsilon_j|$, the quantities involved in the left hand side. Writing these bounds precisely yields an inequality on these two quantities which can be solved to derive the explicit bounds. In the case of strictly convex smooth penalty function ψ (such as $\psi(u) = u^p$, $p > 1$ or $\psi(u) = u \log u$), the same approach can be used also in the case of “approximately sparse” oracles λ (since the function ψ' is strictly increasing and smooth). A natural choice of oracle is

$$\lambda^{\varepsilon} := \mathrm{argmin}_{\lambda \in D}\biggl[P(\ell \bullet f_{\lambda}) + \varepsilon \sum_{j=1}^{N} \psi(\lambda_j)\biggr], \tag{8.6}$$

for which in the smooth case

$$\Big\langle \nabla L_{\varepsilon}(\lambda^{\varepsilon}), \lambda^{\varepsilon} - \hat{\lambda}^{\varepsilon} \Big\rangle_{\ell_2} \leq 0$$

(if $D = \mathbb{R}^N$, we even have $\nabla L_{\varepsilon}(\lambda^{\varepsilon}) = 0$). For this oracle, the bounds on $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$ and on the degree of sparsity of $\hat{\lambda}^{\varepsilon}$ do not depend on the properties of the dictionary, but only on ”approximate sparsity” of λ^{ε}. As a consequence, it is also possible to bound the ”random error” $|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})|$ in terms of ”approximate sparsity” of λ^{ε}. It happens that bounding the ”approximation error” $\mathcal{E}(f_{\lambda^{\varepsilon}})$ is a different problem with not entirely the same geometric parameters responsible for the size of the error. The approximation error is much more sensitive to the properties of the dictionary, in particular, its Gram matrix H which depends on unknown design distribution Π.

The case of ℓ_1-penalty is more complicated since the penalty is neither strictly convex, nor smooth. In this case there is no special advantage in using λ^{ε} as an oracle since this vector is not necessarily sparse. It is rather approximately sparse, but bound (8.4) does not provide a way to control the random L_2-error $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$ in terms of approximate sparsity of the oracle (note that in this case $\psi'(\lambda) = \mathrm{sign}(\lambda)$). A possible way to tackle the problem is to study a set of oracles λ for which

$$\Big\langle \nabla L_{\varepsilon}(\lambda), \lambda - \hat{\lambda}^{\varepsilon} \Big\rangle_{\ell_2}$$

is negative, or, if positive, then small enough. This can be expressed in terms of certain quantities that describe a way in which the subgradient $\nabla L_{\varepsilon}(\lambda)$ is **aligned** with the dictionary. Such quantities also emerge rather naturally in attempts to control the approximation error $\mathcal{E}(f_{\lambda^{\varepsilon}})$ in the case of smooth strictly convex penalties.

8.2 ℓ_1-Penalization: Bounding the ℓ_1-Norm of a Solution

In the case when the set D is not bounded, there are some additional technical difficulties involved in the analysis of the problem related to the need to provide bounds on proper norms of the empirical solution $\hat{\lambda}^{\varepsilon}$. The bounds of this type have been developed, for instance, in [65], Theorem 1. Here is a version of this result in the case of ℓ_1-penalization (LASSO) over the whole space $\mathbb{R}^N$.

Denote

$$\hat\lambda^\varepsilon := \mathrm{argmin}_{\lambda\in\mathbb{R}^N}\biggl[P_n(\ell\bullet f_\lambda)+\varepsilon\|\lambda\|_{\ell_1}\biggr] \tag{8.7}$$

and

$$\lambda^\varepsilon := \mathrm{argmin}_{\lambda\in\mathbb{R}^N}\biggl[P(\ell\bullet f_\lambda)+\varepsilon\|\lambda\|_{\ell_1}\biggr]. \tag{8.8}$$

Theorem 8.1 *There exists a constant $D>0$ depending only on ℓ such that for all $A\geq 1$ and for all ε and $\lambda\in\mathbb{R}^N$ satisfying the assumption*

$$\varepsilon\geq D\biggl(\|\ell'\bullet f_\lambda\|_{L_2(P)}\sqrt{\frac{A\log N}{n}}\bigvee\|\ell'\bullet f_\lambda\|_\infty\frac{A\log N}{n}\biggr)\bigvee 4\max_{1\leq k\leq N}|P(\ell'\bullet f_\lambda)h_k|, \tag{8.9}$$

the following inequality holds:

$$\mathbb{P}\Bigl\{\|\hat\lambda^\varepsilon\|_{\ell_1}\geq 3\|\lambda\|_{\ell_1}\Bigr\}\leq N^{-A}.$$

In particular, if

$$\varepsilon\geq D\biggl(\|\ell'\bullet f_{\lambda^{\varepsilon/4}}\|_{L_2(P)}\sqrt{\frac{A\log N}{n}}\bigvee\|\ell'\bullet f_{\lambda^{\varepsilon/4}}\|_\infty\frac{A\log N}{n}\biggr),$$

then

$$\mathbb{P}\Bigl\{\|\hat\lambda^\varepsilon\|_{\ell_1}\geq 3\|\lambda^{\varepsilon/4}\|_{\ell_1}\Bigr\}\leq N^{-A}.$$

Proof. The definition of $\hat\lambda^\varepsilon$ implies that

$$P_n(\ell\bullet f_{\hat\lambda^\varepsilon})+\varepsilon\|\hat\lambda^\varepsilon\|_{\ell_1}\leq P_n(\ell\bullet f_\lambda)+\varepsilon\|\lambda\|_{\ell_1},\ \lambda\in\mathbb{R}^N.$$

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

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

As a result,

$$\varepsilon\|\hat\lambda^\varepsilon\|_{\ell_1}\leq\varepsilon\|\lambda\|_{\ell_1}+P_n(\ell'\bullet f_\lambda)(f_\lambda-f_{\hat\lambda^\varepsilon})\leq$$

$$\varepsilon\|\lambda\|_{\ell_1}+\max_{1\leq k\leq N}|P_n(\ell'\bullet f_\lambda)h_k|\ \|\hat\lambda^\varepsilon-\lambda\|_{\ell_1}.$$

This yields the bound

$$\Bigl(\varepsilon-\max_{1\leq k\leq N}|P_n(\ell'\bullet f_\lambda)h_k|\Bigr)\|\hat\lambda^\varepsilon\|_{\ell_1}\leq\Bigl(\varepsilon+\max_{1\leq k\leq N}|P_n(\ell'\bullet f_\lambda)h_k|\Bigr)\|\lambda\|_{\ell_1}.$$

If

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

then

$$\|\hat{\lambda}^\varepsilon\|_{\ell_1} \le \frac{\varepsilon + \max_{1\le k\le N}|P_n(\ell' \bullet f_\lambda)h_k|}{\varepsilon - \max_{1\le k\le N}|P_n(\ell' \bullet f_\lambda)h_k|}\ \|\lambda\|_{\ell_1}. \tag{8.10}$$

Note that, under the assumption (8.9),

$$\max_{1\le k\le N} |P_n(\ell' \bullet f_\lambda)h_k| \le \max_{1\le k\le N} |P(\ell' \bullet f_\lambda)h_k| +$$

$$\max_{1\le k\le N} |(P_n - P)(\ell' \bullet f_\lambda)h_k| \le \varepsilon/4 + \max_{1\le k\le N} |(P_n - P)(\ell' \bullet f_\lambda)h_k|.$$

The second term is bounded using Bernstein's inequality which yields that with probability at least $1 - N^{-A}$ and with some choice of constant C

$$\max_{1\le k\le N} |(P_n - P)(\ell' \bullet f_\lambda)h_k| \le C\|\ell' \bullet f_\lambda\|_{L_2(P)}\sqrt{\frac{A\log N}{n}} \bigvee \|\ell' \bullet f_\lambda\|_\infty \frac{A\log N}{n}.$$

If the assumption (8.9) on λ is satisfied with $D = 4C$, then with probability at least $1 - N^{-A}$

$$\max_{1\le k\le N} |(P_n - P)(\ell' \bullet f_\lambda)h_k| \le \varepsilon/4$$

and it follows that with the same probability

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

If we use in (8.10) $\lambda := \lambda^{\varepsilon/4}$, then, by the necessary conditions of extremum in the definition of $\lambda^{\varepsilon/4}$,

$$|P(\ell' \bullet f_{\lambda^{\varepsilon/4}})h_k| \le \frac{\varepsilon}{4},\ k = 1, \dots, N,$$

which implies the second statement.

□

It is also possible to show that $\|\lambda^{c\varepsilon}\|_{\ell_1}$ with a large enough constant c provides a lower bound on $\|\hat{\lambda}^\varepsilon\|_{\ell_1}$ (with a high probability). Namely, the following result holds.

Theorem 8.2 *There exist constants $D > 0, c > 0$ depending only on ℓ such that, for all $A \ge 1$ and for all ε satisfying the assumption*

$$\varepsilon \ge D(\|\lambda^{\varepsilon/4}\|_{\ell_1} + 1)\sqrt{\frac{A\log N}{n}}, \tag{8.11}$$

the following inequality holds:

$$\mathbb{P}\Big\{\|\hat{\lambda}^\varepsilon\|_{\ell_1} \le \frac{1}{3}\|\lambda^{c\varepsilon}\|_{\ell_1}\Big\} \le N^{-A}.$$

Proof. By the definition of $\lambda^{c\varepsilon}$,

$$P(\ell \bullet f_{\lambda^{c\varepsilon}}) + c\varepsilon \|\lambda^{c\varepsilon}\|_{\ell_1} \leq P(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) + c\varepsilon \|\hat{\lambda}^{\varepsilon}\|_{\ell_1}.$$

Arguing exactly as at the beginning of the proof of Theorem 8.1, one can show that

$$\|\lambda^{c\varepsilon}\|_{\ell_1} \leq \frac{c\varepsilon + \max_{1\leq k\leq N} |P(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})h_k|}{c\varepsilon - \max_{1\leq k\leq N} |P(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})h_k|} \|\hat{\lambda}^{\varepsilon}\|_{\ell_1} \tag{8.12}$$

as soon as

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

We have

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

and, using necessary conditions of extremum in problem (8.7), the first term can be bounded as follows:

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

To bound the second term, we use the following lemma.

Lemma 8.1 *There exist constants C, L depending only on ℓ such that for all $A \geq 1$ and for all $R > 0$ with probability at least $1 - N^{-A}$*

$$\max_{1\leq k\leq N} \sup_{\|\lambda\|_{\ell_1}\leq R} |(P_n - P)(\ell' \bullet f_{\lambda})h_k| \leq C(1 + LR)\sqrt{\frac{A \log N}{n}}.$$

We use it for $R = \|\lambda^{\varepsilon/4}\|_{\ell_1}$ and combine it with the second bound of Theorem 8.1. This bound can be used since under the assumptions on the loss function ℓ

$$\|\ell' \bullet f_{\lambda}\|_{L_2(P)} \leq \|\ell' \bullet f_{\lambda}\|_{\infty} \leq C_1(1 + \|\lambda\|_{\ell_1})$$

with some constant C_1, which allows one to write down the condition on ε as (8.11). With an adjustment of the constants, it follows that with probability at least $1 - N^{-A}$

$$\max_{1\leq k\leq N} |(P_n - P)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})h_k| \leq C(1 + L\|\lambda^{\varepsilon/4}\|_{\ell_1})\sqrt{\frac{A \log N}{n}}.$$

One can choose the value of c in such a way that the condition

$$c\varepsilon \geq 2C(1 + L\|\lambda^{\varepsilon/4}\|_{\ell_1})\sqrt{\frac{A \log N}{n}} + 2\varepsilon$$

would follow from the assumption on ε (8.11). With such a choice of c, we have that with probability at least $1 - N^{-A}$

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

It is enough to recall (8.12) to get that with probability at least $1 - N^{-A}$

$$\|\lambda^{c\varepsilon}\|_{\ell_1} \le 3\|\hat{\lambda}^{\varepsilon}\|_{\ell_1}.$$

□

Proof of Lemma 8.1. Note that, under the assumptions on ℓ, we have, for all $k = 1, \dots, N$ and all λ satisfying $\|\lambda\|_{\ell_1} \le R$,

$$\|(\ell' \bullet f_\lambda)h_k\|_\infty \le C(1 + LR)$$

with constants $C, L > 0$ depending only on ℓ. We apply the bounded difference inequality to the supremum of the empirical process indexed by the class

$$\mathcal{G} := \Big\{ \frac{(\ell' \bullet f_\lambda)h_k}{C(1 + LR)} : \|\lambda\|_{\ell_1} \le R \Big\}.$$

This yields the following bound that holds with probability at least $1 - e^{-t}$:

$$\sup_{\|\lambda\|_{\ell_1} \le R} |(P_n - P)(\ell' \bullet f_\lambda)h_k| \le \mathbb{E} \sup_{\|\lambda\|_{\ell_1} \le R} |(P_n - P)(\ell' \bullet f_\lambda)h_k| + \frac{C(1 + LR)\sqrt{t}}{\sqrt{n}}.$$

To bound the expectation

$$\mathbb{E} \sup_{\|\lambda\|_{\ell_1} \le R} |(P_n - P)(\ell' \bullet f_\lambda)h_k|,$$

we use the symmetrization inequality followed by the contraction inequality:

$$\mathbb{E} \sup_{\|\lambda\|_{\ell_1} \le R} |(P_n - P)(\ell' \bullet f_\lambda)h_k| \le 2\mathbb{E} \sup_{\|\lambda\|_{\ell_1} \le R} |R_n((\ell' \bullet f_\lambda)h_k)| \le C\mathbb{E} \sup_{\|\lambda\|_{\ell_1} \le R} |R_n(f_\lambda)|.$$

Using Theorem 3.4, we get

$$\mathbb{E} \sup_{\|\lambda\|_{\ell_1} \le R} |R_n(f_\lambda)| \le R\mathbb{E} \max_{1 \le i \le N} |R_n(h_i)| \le CR\sqrt{\frac{\log N}{n}}.$$

It follows from all of the above bounds that with probability at least $1 - e^{-t}$ and with some constants C, L depending only on ℓ,

$$\sup_{\|\lambda\|_{\ell_1} \le R} |(P_n - P)(\ell \bullet f_\lambda)h_k| \le C(1 + LR)\sqrt{\frac{\log N + t}{n}}.$$

We use this bound for all $k = 1, \dots, N$ with $t = A \log N + \log N$ and then apply the union bound. With a proper adjustment of the constants, this completes the proof.

□

The next result provides a simple upper bound on the excess risk of the empirical solution $f_{\hat{\lambda}^\varepsilon}$.

Theorem 8.3 *There exist constants $C, D > 0$ depending only on ℓ such that for all $A \geq 1$ and for all ε and $\lambda \in \mathbb{R}^N$ satisfying the assumption*

$$\varepsilon \geq D(\|\lambda\|_{\ell_1} + 1)\sqrt{\frac{A \log N}{n}} \bigvee 4 \max_{1 \leq k \leq N} |P(\ell' \bullet f_\lambda) h_k|, \tag{8.13}$$

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

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) \leq \mathcal{E}(f_\lambda) + C\|\lambda\|_{\ell_1}\varepsilon.$$

Proof. We will use the following lemma that can be proved quite similarly to Lemma 8.1.

Lemma 8.2 *There exist constants C, L depending only on ℓ such that for all $A \geq 1$ and for all $R > 0$ with probability at least $1 - N^{-A}$*

$$\sup_{\|\lambda\|_{\ell_1} \leq R} |(P_n - P)(\ell \bullet f_\lambda - \ell \bullet f_0)| \leq C(1 + LR)R\sqrt{\frac{A \log N}{n}}.$$

Using the definition of $\hat{\lambda}^\varepsilon$, we get

$$P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_\lambda) \leq P_n(\ell \bullet f_{\hat{\lambda}^\varepsilon}) + \varepsilon\|\hat{\lambda}^\varepsilon\|_{\ell_1} - P_n(\ell \bullet f_\lambda) - \varepsilon\|\lambda\|_{\ell_1} +$$

$$+\varepsilon\|\lambda\|_{\ell_1} + 2 \sup_{\|u\|_{\ell_1} \leq \|\lambda\|_{\ell_1} \vee \|\hat{\lambda}^\varepsilon\|_{\ell_1}} |(P_n - P)(\ell \bullet f_u - \ell \bullet f_0)| \leq$$

$$\varepsilon\|\lambda\|_{\ell_1} + 2 \sup_{\|u\|_{\ell_1} \leq \|\lambda\|_{\ell_1} \vee \|\hat{\lambda}^\varepsilon\|_{\ell_1}} |(P_n - P)(\ell \bullet f_u - \ell \bullet f_0)|.$$

Note that, under the assumptions on the loss function ℓ, (8.13) implies (8.9) (with a proper choice of constants in these assumptions). Then, it follows from Theorem 8.1 that with probability at least $1 - N^{-A}$,

$$\|\hat{\lambda}^\varepsilon\|_{\ell_1} \leq 3\|\lambda\|_{\ell_1}.$$

This can be combined with the bound of Lemma 8.2 to get that with a proper choice of C and with probability at least $1 - N^{-A}$

$$2 \sup_{\|u\|_{\ell_1} \leq \|\lambda\|_{\ell_1} \vee \|\hat{\lambda}^{\varepsilon}\|_{\ell_1}} |(P_n - P)(\ell \bullet f_u - \ell \bullet f_0)| \leq C(1 + L\|\lambda\|_{\ell_1})\|\lambda\|_{\ell_1}\sqrt{\frac{A \log N}{n}}.$$

As a result, we easily get the bound

$$P(\ell \bullet f_{\hat{\lambda}^{\varepsilon}}) - P(\ell \bullet f_\lambda) \leq C\|\lambda\|_{\ell_1}\varepsilon$$

that holds with probability at least $1 - N^{-A}$ with a proper choice of constant $C > 0$. This implies the statement of the theorem.

□

In the following sections, we concentrate on the case when the set D is bounded. However, our method of proof combined with such results as Theorem 8.1 can be easily used to handle the case of unbounded domain (see [65] for some results in this direction).

8.3 ℓ_1-Penalization and Oracle Inequalities

The following penalized empirical risk minimization problem will be studied:

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in U_{\ell_1}}\biggl[P_n(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_1}\biggr], \tag{8.14}$$

where $\varepsilon \geq 0$ is a regularization parameter and

$$\|\lambda\|_{\ell_1} := \sum_{j=1}^{N} |\lambda_j|.$$

Denote

$$L_\varepsilon(\lambda) := P(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_1}.$$

For $\lambda \in \mathbb{R}^N$, let $\nabla L_\varepsilon(\lambda) \in \partial L_\varepsilon(\lambda)$ be the vector with components

$$P(\ell' \bullet f_\lambda)h_j + \varepsilon s_j(\lambda),\ j = 1, \ldots, N$$

where $s_j = s_j(\lambda) = \mathrm{sign}(\lambda_j)$ (assume that $\mathrm{sign}(0) = 0$). The vector $\nabla L_\varepsilon(\lambda)$ is a subgradient of the function L_ε at point λ. Note that $\partial|u| = +1$ for $u > 0$, $\partial|u| = -1$ for $u, 0$ and $\partial|u| = [-1, 1]$ for $u = 0$.

In the case of ℓ_1-penalization, we are going to compare the empirical solution $\hat{\lambda}^{\varepsilon}$ with an oracle $\lambda \in U_{\ell_1}$ that will be characterized by its "sparsity" as well as by a measure of "alignment" of the subgradient $\nabla L_\varepsilon(\lambda) \in \partial L_\varepsilon(\lambda)$.

We will use the following versions of alignment for vectors $\nabla L_\varepsilon(\lambda)$ and $s(\lambda)$:

$$\alpha_+(\varepsilon, \lambda) := a_H^{(\infty)}\Big(U_{\ell_1}, \lambda, \nabla L_\varepsilon(\lambda)\Big) \vee 0$$

and

$$\alpha(\lambda) := a_H^{(2)}\Big(U_{\ell_1}, \lambda, s(\lambda)\Big) \vee 0, \quad \alpha_+(\lambda) := a_H^{(\infty)}\Big(U_{\ell_1}, \lambda, s(\lambda)\Big) \vee 0.$$

Clearly,

$$\alpha(\lambda) \le \alpha_+(\lambda)$$

and it is easy to check that

$$\alpha_+(\varepsilon, \lambda) \le \|\ell' \bullet f_\lambda\|_{L_2(P)} + \varepsilon \alpha_+(\lambda).$$

The last term in the right hand side is the alignment coefficient of vector $s(\lambda)$ that depends on the sparsity of λ as well as on geometry of the dictionary.

Theorem 8.4 *There exist constants $D > 0$ and $C > 0$ depending only on ℓ such that, for all $\bar\lambda \in U_{\ell_1}$, for $J = \mathrm{supp}(\bar\lambda)$ and $d := d(J) = \mathrm{card}(J)$, for all $A \ge 1$ and for all*

$$\varepsilon \ge D\sqrt{\frac{d + A\log N}{n}}, \tag{8.15}$$

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

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\hat\lambda_j^\varepsilon| \le C\biggl[\frac{d + A\log N}{n} \bigvee \alpha_+^2(\varepsilon, \bar\lambda)\biggr].$$

Moreover, with the same probability

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\hat\lambda_j^\varepsilon| \le C\biggl[\frac{d + A\log N}{n} \bigvee \Big\|\ell' \bullet f_{\bar\lambda}\Big\|^2_{L_2(P)} \bigvee \alpha^2(\bar\lambda)\varepsilon^2\biggr].$$

No condition on the dictionary is needed for the bounds of the theorem to be true (except uniform boundedness of functions h_j). On the other hand, the assumption on ε,

$$\varepsilon \ge D\sqrt{\frac{d + A\log N}{n}},$$

essentially, relates the regularization parameter to the unknown sparsity of the problem. To get around this difficulty, we will prove another version of the theorem in which it is only assumed that

$$\varepsilon \ge D\sqrt{\frac{A\log N}{n}},$$

but, on the other hand, there is some dependence on the geometry of the dictionary. At the same time, the error in this result is controlled not by $d = \mathrm{card}(J)$, but rather by the dimension of a linear space L providing a reasonably good approximation of the functions $\{h_j : j \in J\}$ (such a dimension could be much smaller than $\mathrm{card}(J)$). To formulate this result, some further notation will be needed.

Given a linear subspace $L \subset L_2(\Pi)$, denote

$$U(L) := \sup_{f \in L, \|f\|_{L_2(\Pi)}=1} \|f\|_\infty + 1.$$

If $I_L : (L, \|\cdot\|_{L_2(\Pi)}) \mapsto (L, \|\cdot\|_\infty)$ is the identity operator, then $U(L) - 1$ is the norm of the operator I_L. We will use this quantity only for finite dimensional subspaces. In such case, for any $L_2(\Pi)$-orthonormal basis $\phi_1, \dots, \phi_d$ of L,

$$U(L) \leq \max_{1 \leq j \leq d} \|\phi_j\|_\infty \sqrt{d} + 1,$$

where $d := \dim(L)$. In what follows, let P_L be the orthogonal projector onto L and $L^\perp$ be the orthogonal complement of L. We are interested in subspaces L such that

(a) $\dim(L)$ and $U(L)$ are not very large;

(b) functions $\{h_j : j \in J\}$ in the "relevant" part of the dictionary can be approximated well by the functions from L so that the quantity

$$\max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}$$

is small.

Theorem 8.5 *Suppose that*

$$\varepsilon \geq D\sqrt{\frac{A \log N}{n}} \tag{8.16}$$

with a large enough constant $D > 0$ depending only on ℓ. For all $\bar\lambda \in U_{\ell_1}$, for $J = \mathrm{supp}(\bar\lambda)$, for all subspaces L of $L_2(\Pi)$ with $d := \dim(L)$ and for all $A \geq 1$, the following bound holds with probability at least $1 - N^{-A}$ and with a constant $C > 0$ depending only on ℓ :

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\hat\lambda^\varepsilon_j| \leq \tag{8.17}$$

$$C\Biggl[\frac{d + A\log N}{n} \bigvee \max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A \log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \alpha_+^2(\varepsilon; \bar\lambda)\Biggr].$$

Moreover, with the same probability

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j\notin J} |\hat\lambda^\varepsilon_j| \leq \tag{8.18}$$
$$C\biggl[\frac{d + A\log N}{n} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee$$
$$\Bigl\|\ell' \bullet f_{\bar\lambda}\Bigr\|^2_{L_2(P)} \bigvee \alpha^2(\bar\lambda)\varepsilon^2\biggr].$$

The next two corollaries provide bounds on $\|\hat\lambda^\varepsilon - \bar\lambda\|_{\ell_1}$ in terms of the quantity $\beta_{2,2}(\bar\lambda, \Pi)$; they follow in a straightforward way from the proofs of the theorems.

Corollary 8.1 *Under the assumptions and notations of Theorem 8.4, the following bound holds with probability at least* $1 - N^{-A}$:

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon\|\hat\lambda^\varepsilon - \bar\lambda\|_{\ell_1} \leq C\biggl[\frac{d + A\log N}{n} \bigvee \Bigl\|\ell' \bullet f_{\bar\lambda}\Bigr\|^2_{L_2(P)} \bigvee \beta^2_{2,2}(\bar\lambda, \Pi)\varepsilon^2 d\biggr].$$

Corollary 8.2 *Under the assumptions and notations of Theorem 8.5, the following bound holds with probability at least* $1 - N^{-A}$:

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon\|\hat\lambda^\varepsilon - \bar\lambda\|_{\ell_1} \leq \tag{8.19}$$
$$C\biggl[\frac{d + A\log N}{n} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee$$
$$\Bigl\|\ell' \bullet f_{\bar\lambda}\Bigr\|^2_{L_2(P)} \bigvee \beta^2_{2,2}(\bar\lambda, \Pi)\varepsilon^2 d\biggr].$$

Proof of Theorem 8.5. According to the definition of $\hat\lambda^\varepsilon$,

$$\hat\lambda^\varepsilon \in \mathrm{Argmin}_{\|\lambda\|_{\ell_1}\leq 1}\biggl[P_n(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_1}\biggr]. \tag{8.20}$$

Subgradients of convex function

$$\lambda \mapsto P_n(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_1}$$

are the vectors in $\mathbb{R}^N$ with components

$$P_n(\ell' \bullet f_\lambda)h_j + \varepsilon\sigma_j,\ j = 1, \ldots, N$$

where $\sigma_j \in [-1,1]$, $\sigma_j = \mathrm{sign}(\lambda_j)$ if $\lambda_j \neq 0$. It follows from necessary conditions of extremum in problem (8.20) that there exist numbers $\hat{s}_j \in [-1,1]$ such that $\hat{s}_j = \mathrm{sign}(\hat{\lambda}^\varepsilon_j)$ when $\hat{\lambda}^\varepsilon_j \neq 0$ and, for all $u \in T_{U_{\ell_1}}(\hat{\lambda}^\varepsilon)$,

$$\sum_{j=1}^N \Bigl(P_n(\ell' \bullet f_{\hat{\lambda}^\varepsilon})h_j u_j + \varepsilon \hat{s}_j u_j\Bigr) \geq 0.$$

Since $\bar{\lambda} \in U_{\ell_1}$, $\bar{\lambda} - \hat{\lambda}^\varepsilon \in T_{U_{\ell_1}}(\hat{\lambda}^\varepsilon)$, and the next inequality immediately follows:

$$P_n(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}) + \varepsilon \sum_{j=1}^N \hat{s}_j(\hat{\lambda}_j - \bar{\lambda}_j) \leq 0. \tag{8.21}$$

Recalling the definition

$$s_j = s_j(\bar{\lambda}) = \mathrm{sign}(\bar{\lambda}_j)$$

and

$$\nabla L_\varepsilon(\bar{\lambda}) = \Biggl(P(\ell' \bullet f_{\bar{\lambda}})h_j + \varepsilon s_j\Biggr)_{j=1,\dots,N},$$

we also have

$$P(\ell' \bullet f_{\bar{\lambda}})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}) + \varepsilon \sum_{j=1}^N s_j(\hat{\lambda}_j - \bar{\lambda}_j) = \Bigl\langle \nabla L_\varepsilon(\bar{\lambda}), \hat{\lambda}^\varepsilon - \bar{\lambda}\Bigr\rangle_{\ell_2}. \tag{8.22}$$

Subtracting (8.22) from (8.21) yields by a simple algebra

$$P_n(\ell' \bullet f_{\hat{\lambda}^\varepsilon} - \ell' \bullet f_{\bar{\lambda}})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}) + \varepsilon \sum_{j=1}^N (\hat{s}_j - s_j)(\hat{\lambda}_j - \bar{\lambda}_j) \leq$$

$$\Bigl\langle \nabla L_\varepsilon(\bar{\lambda}), \bar{\lambda} - \hat{\lambda}^\varepsilon\Bigr\rangle_{\ell_2} + (P - P_n)(\ell' \bullet f_{\bar{\lambda}})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}) \tag{8.23}$$

and

$$P(\ell' \bullet f_{\hat{\lambda}^\varepsilon} - \ell' \bullet f_{\bar{\lambda}})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}) + \varepsilon \sum_{j=1}^N (\hat{s}_j - s_j)(\hat{\lambda}_j - \bar{\lambda}_j) \leq$$

$$\Bigl\langle \nabla L_\varepsilon(\bar{\lambda}), \bar{\lambda} - \hat{\lambda}^\varepsilon\Bigr\rangle_{\ell_2} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}). \tag{8.24}$$

We use inequalities (8.23) and (8.24) to control the "approximate sparsity" of empirical solution $\hat{\lambda}^\varepsilon$ in terms of "sparsity" of the "oracle" $\bar{\lambda}$ and to obtain bounds on $\|f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}\|_{L_2(\Pi)}$. As always, we use notations $J := J_{\bar{\lambda}} := \mathrm{supp}(\bar{\lambda})$. By the conditions on the loss (namely, the boundedness of its second derivative away from 0), we have

$$P(\ell' \bullet f_{\hat{\lambda}^\varepsilon} - \ell' \bullet f_{\bar{\lambda}})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}) \geq c\|f_{\hat{\lambda}^\varepsilon} - f_\lambda\|^2_{L_2(\Pi)},$$

where $c = \tau(1)$ (note that $\|f_{\bar\lambda}\|_\infty \le 1$ and $\|f_{\hat\lambda^\varepsilon}\|_\infty \le 1$). Observe also that, for all j,

$$(\hat s_j - s_j)(\hat\lambda_j - \bar\lambda_j) \ge 0$$

(by monotonicity of subdifferential of convex function $u \mapsto |u|$). For $j \notin J$, we have $\bar\lambda_j = 0$ and $s_j = 0$. Therefore, (8.24) implies that

$$c\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j\notin J} |\hat\lambda_j| \le$$

$$\Bigl\langle \nabla L_\varepsilon(\bar\lambda), \bar\lambda - \hat\lambda^\varepsilon \Bigr\rangle_{\ell_2} + (P - P_n)(\ell' \bullet f_{\hat\lambda^\varepsilon})(f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}). \tag{8.25}$$

Consider first the case when

$$\Bigl\langle \nabla L_\varepsilon(\bar\lambda), \bar\lambda - \hat\lambda^\varepsilon \Bigr\rangle_{\ell_2} \ge (P - P_n)(\ell' \bullet f_{\hat\lambda^\varepsilon})(f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}). \tag{8.26}$$

In this case, (8.25) implies that

$$c\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j\notin J} |\hat\lambda_j| \le 2\Bigl\langle \nabla L_\varepsilon(\bar\lambda), \bar\lambda - \hat\lambda^\varepsilon \Bigr\rangle_{\ell_2}, \tag{8.27}$$

which, in view of definition of $\alpha_+(\varepsilon, \bar\lambda)$, yields

$$c\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j\notin J} |\hat\lambda_j| \le 2\alpha_+(\varepsilon, \bar\lambda)\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|_{L_2(\Pi)}. \tag{8.28}$$

Therefore,

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|_{L_2(\Pi)} \le \frac{2}{c}\alpha_+(\varepsilon, \bar\lambda),$$

and, as a consequence, with some constant $C > 0$ depending only on ℓ

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j\notin J} |\hat\lambda_j| \le C\alpha_+^2(\varepsilon, \bar\lambda). \tag{8.29}$$

If

$$\Bigl\langle \nabla L_\varepsilon(\bar\lambda), \bar\lambda - \hat\lambda^\varepsilon \Bigr\rangle_{\ell_2} < (P - P_n)(\ell' \bullet f_{\hat\lambda^\varepsilon})(f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}), \tag{8.30}$$

then (8.25) implies that

$$c\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j\notin J} |\hat\lambda_j| \le 2(P - P_n)(\ell' \bullet f_{\hat\lambda^\varepsilon})(f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}). \tag{8.31}$$

Denote

$$\Lambda(\delta; \Delta) := \Bigl\{ \lambda \in U_{\ell_1} : \|f_\lambda - f_{\bar\lambda}\|_{L_2(\Pi)} \le \delta,\ \sum_{j\notin J} |\lambda_j| \le \Delta \Bigr\},$$

$$\alpha_n(\delta; \Delta) := \sup\Bigl\{ |(P_n - P)((\ell' \bullet f_\lambda)(f_\lambda - f_{\bar\lambda}))| : \lambda \in \Lambda(\delta; \Delta) \Bigr\}.$$

To bound $\alpha_n(\delta, \Delta)$, the following lemma will be used.

Lemma 8.3 *Under the assumptions of Theorem 8.5, there exists constant C that depends only on ℓ such that with probability at least $1 - N^{-A}$, for all*

$$n^{-1/2} \leq \delta \leq 1 \quad \text{and} \quad n^{-1/2} \leq \Delta \leq 1 \tag{8.32}$$

the following bounds hold:

$$\alpha_n(\delta; \Delta) \leq \beta_n(\delta; \Delta) := C\Bigg[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \Delta\sqrt{\frac{A\log N}{n}}$$
$$\bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Bigg]. \tag{8.33}$$

Take

$$\delta = \|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \quad \text{and} \quad \Delta = \sum_{j\notin J} \hat{\lambda}_j^\varepsilon. \tag{8.34}$$

If $\delta \geq n^{-1/2}, \Delta \geq n^{-1/2}$, then Lemma 8.3 and (8.31) imply the following bound:

$$c\delta^2 + \varepsilon\Delta \leq 2\beta_n(\delta, \Delta). \tag{8.35}$$

If $\delta < n^{-1/2}$ or $\Delta < n^{-1/2}$, they should be replaced in the expression for $\beta_n(\delta, \Delta)$ by $n^{-1/2}$. With this change, bound (8.35) still holds and the proof goes through with some simplifications. Thus, we will consider only the main case when $\delta \geq n^{-1/2}, \Delta \geq n^{-1/2}$. In this case, the inequality (8.35) has to be solved to complete the proof. It follows from this inequality (with a proper change of constant C) that

$$\varepsilon\Delta \leq C\Delta\sqrt{\frac{A\log N}{n}} + C\Bigg[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee$$
$$\max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Bigg].$$

As soon as D in condition (8.16) is such that $D \geq 2C$, we can write

$$\varepsilon\Delta \leq C\Bigg[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Bigg]$$

(again the value of constant C might have changed). We solve the inequality with respect to Δ separately for each term in the maximum and take the maximum of the solutions, which yields the following bound

$$\Delta \leq C\Bigg[\frac{\delta}{\varepsilon}\sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \frac{1}{\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n\varepsilon} \bigvee \frac{A\log N}{n\varepsilon}\Bigg].$$

Under the assumption (8.16) on ε (assuming also that $D \geq 1$), it is easy to derive that

$$\Delta \leq \Delta(\delta) := C\biggl[\frac{\delta}{\varepsilon}\sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \bigvee \frac{U(L)\log N}{n\varepsilon} \bigvee \sqrt{\frac{A\log N}{n}}\biggr].$$

Note that $\beta_n(\delta, \Delta)$ is nondecreasing in Δ and replace Δ in (8.35) by $\Delta(\delta)$ to get the following bound:

$$\delta^2 \leq C\biggl[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \frac{\delta}{\varepsilon}\sqrt{\frac{d + A\log N}{n}}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee$$

$$\max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\biggr].$$

We skip the second term in the maximum and modify the third term because $\frac{1}{\varepsilon}\sqrt{\frac{A\log N}{n}} \leq 1$. As a result, we get

$$\delta^2 \leq C\biggl[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\biggr].$$

Solving the last inequality with respect to δ yields the following bound on δ^2 :

$$\delta^2 \leq C\biggl[\frac{d + A\log N}{n} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n}\biggr]. \tag{8.36}$$

We substitute the last bound back into the expression for $\Delta(\delta)$ to get:

$$\Delta \leq C\biggl[\frac{d + A\log N}{n\varepsilon} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|^{1/2}_{L_2(\Pi)} \frac{1}{\varepsilon}\Bigl(\frac{A\log N}{n}\Bigr)^{1/4}\sqrt{\frac{d + A\log N}{n}} \bigvee$$

$$\sqrt{\frac{U(L)\log N}{n\varepsilon}}\sqrt{\frac{d + A\log N}{n\varepsilon}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \bigvee \sqrt{\frac{A\log N}{n}}\biggr].$$

Using the inequality $ab \leq (a^2 + b^2)/2$ and the condition $\frac{1}{\varepsilon}\sqrt{\frac{A\log N}{n}} \leq 1$, we can simplify the resulting bound as follows

$$\Delta \leq C\biggl[\frac{d + A\log N}{n\varepsilon} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \bigvee \frac{U(L)\log N}{n\varepsilon} \bigvee \sqrt{\frac{A\log N}{n}}\biggr] \tag{8.37}$$

with a proper change of C that depends only on ℓ. Finally, bounds (8.36) and (8.37) can be substituted in the expression for $\beta_n(\delta, \Delta)$. By a simple computation and in view of Lemma 8.3, we get the following bound on $\alpha_n(\delta, \Delta)$ that holds for δ, Δ defined by (8.34) with probability at least $1 - N^{-A}$:

$$\alpha_n(\delta, \Delta) \leq C\biggl[\frac{d + A\log N}{n} + \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} + \frac{U(L)\log N}{n}\biggr].$$

Combining this with (8.31) yields

$$c\|f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}\|^2_{L_2(\Pi)} + \varepsilon\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j| \le$$

$$C\biggl[\frac{d + A\log N}{n} + \max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} + \frac{U(L)\log N}{n}\biggr], \qquad (8.38)$$

which holds under condition (8.30).

Together with bound (8.29), that is true under the alternative condition (8.26), this gives (8.17).

To prove bound (8.18), we again use (8.25), but this time we control the term

$$\Bigl\langle \nabla L_\varepsilon(\bar{\lambda}), \bar{\lambda} - \hat{\lambda}^\varepsilon\Bigr\rangle$$

somewhat differently. First note that

$$\Bigl\langle \nabla L_\varepsilon(\bar{\lambda}), \bar{\lambda} - \hat{\lambda}^\varepsilon\Bigr\rangle_{\ell_2} = \Bigl\langle \ell' \bullet f_{\bar{\lambda}}, f_{\bar{\lambda}} - f_{\hat{\lambda}^\varepsilon}\Bigr\rangle_{L_2(P)} + \varepsilon\langle s(\bar{\lambda}), \bar{\lambda} - \hat{\lambda}^\varepsilon\rangle_{\ell_2}.$$

This implies that

$$\Bigl\langle \nabla L_\varepsilon(\bar{\lambda}), \bar{\lambda} - \hat{\lambda}^\varepsilon\Bigr\rangle_{\ell_2} \le \Bigl\|\ell' \bullet f_{\bar{\lambda}}\Bigr\|_{L_2(P)}\|f_{\bar{\lambda}} - f_{\hat{\lambda}^\varepsilon}\|_{L_2(\Pi)} + \varepsilon\sum_{j\in J}s_j(\bar{\lambda}_j - \hat{\lambda}^\varepsilon_j) \le$$

$$\frac{1}{2c}\Bigl\|\ell' \bullet f_{\bar{\lambda}}\Bigr\|^2_{L_2(P)} + \frac{c}{2}\|f_{\bar{\lambda}} - f_{\hat{\lambda}^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon\sum_{j\in J}s_j(\bar{\lambda}_j - \hat{\lambda}^\varepsilon_j).$$

Combining this with bound (8.25) yields the following inequality

$$\frac{c}{2}\|f_{\bar{\lambda}} - f_{\hat{\lambda}^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j| \le$$

$$\varepsilon\sum_{j\in J}s_j(\bar{\lambda}_j - \hat{\lambda}^\varepsilon_j) + \frac{1}{2c}\Bigl\|\ell' \bullet f_{\bar{\lambda}}\Bigr\|^2_{L_2(P)} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}).$$

If

$$\varepsilon\sum_{j\in J}s_j(\bar{\lambda}_j - \hat{\lambda}^\varepsilon_j) \ge \frac{1}{2c}\Bigl\|\ell' \bullet f_{\bar{\lambda}}\Bigr\|^2_{L_2(P)} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}),$$

then

$$\frac{c}{2}\|f_{\bar{\lambda}} - f_{\hat{\lambda}^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j| \le 2\varepsilon\sum_{j\in J}s_j(\bar{\lambda}_j - \hat{\lambda}^\varepsilon_j),$$

which implies

$$\sum_{j\notin J}|\hat{\lambda}^\varepsilon_j| \le 2\sum_{j\in J}|\bar{\lambda}_j - \hat{\lambda}^\varepsilon_j|,$$

or $\hat{\lambda}^{\varepsilon} - \bar{\lambda} \in C_{2,\bar{\lambda}}$. The definition of $\alpha(\bar{\lambda})$ then implies the bound

$$\frac{c}{2}\|f_{\bar{\lambda}} - f_{\hat{\lambda}^{\varepsilon}}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\hat{\lambda}^{\varepsilon}_j| \leq 2\varepsilon\alpha(\bar{\lambda})\|f_{\bar{\lambda}} - f_{\hat{\lambda}^{\varepsilon}}\|_{L_2(\Pi)}.$$

Solving this inequality with respect to $\|f_{\bar{\lambda}} - f_{\hat{\lambda}^{\varepsilon}}\|_{L_2(\Pi)}$ proves (8.18) in this case.

If

$$\varepsilon \sum_{j \in J} s_j(\bar{\lambda}_j - \hat{\lambda}^{\varepsilon}_j) \leq \frac{1}{2c}\Big\|\ell' \bullet f_{\bar{\lambda}}\Big\|^2_{L_2(P)} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\bar{\lambda}})$$

and

$$\frac{1}{2c}\Big\|\ell' \bullet f_{\bar{\lambda}}\Big\|^2_{L_2(P)} \geq (P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\bar{\lambda}}),$$

we get

$$\frac{c}{2}\|f_{\bar{\lambda}} - f_{\hat{\lambda}^{\varepsilon}}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\hat{\lambda}^{\varepsilon}_j| \leq \frac{2}{c}\Big\|\ell' \bullet f_{\bar{\lambda}}\Big\|^2_{L_2(P)},$$

which also implies (8.18) with a proper choice of constant C in the bound.

Thus, it remains to consider the case when

$$\varepsilon \sum_{j \in J} s_j(\bar{\lambda}_j - \hat{\lambda}^{\varepsilon}_j) \leq \frac{1}{2c}\Big\|\ell' \bullet f_{\bar{\lambda}}\Big\|^2_{L_2(P)} + (P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\bar{\lambda}})$$

and

$$\frac{1}{2c}\Big\|\ell' \bullet f_{\bar{\lambda}}\Big\|^2_{L_2(P)} \leq (P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\bar{\lambda}}),$$

which implies

$$\frac{c}{2}\|f_{\bar{\lambda}} - f_{\hat{\lambda}^{\varepsilon}}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\hat{\lambda}^{\varepsilon}_j| \leq 4(P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\bar{\lambda}}).$$

In this case, we repeat the argument based on Lemma 8.3 to show that with probability at least $1 - N^{-A}$

$$\frac{c}{2}\|f_{\hat{\lambda}^{\varepsilon}} - f_{\bar{\lambda}}\|^2_{L_2(\Pi)} + \varepsilon \sum_{j \notin J} |\hat{\lambda}^{\varepsilon}_j| \leq$$

$$C\Bigg[\frac{d + A\log N}{n} + \max_{j \in J}\|P_{L^{\perp}}h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} + \frac{U(L)\log N}{n}\Bigg],$$

which again implies (8.18).

This completes the proof.

□

The proof of Theorem 8.4 is quite similar. The following lemma is used instead of Lemma 8.3.

Lemma 8.4 *Under the assumptions of Theorem 8.4, there exists constant C that depends only on ℓ such that with probability at least $1 - N^{-A}$, for all*

$$n^{-1/2} \leq \delta \leq 1 \ \text{ and } \ n^{-1/2} \leq \Delta \leq 1$$

the following bounds hold:

$$\begin{aligned} &\alpha_n(\delta;\Delta) \leq \beta_n(\delta;\Delta) := \\ &C\biggl[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \Delta\sqrt{\frac{d + A\log N}{n}} \bigvee \frac{A\log N}{n}\biggr]. \end{aligned} \tag{8.39}$$

Proof of Lemma 8.3. First we use Talagrand's concentration inequality to get with probability at least $1 - e^{-t}$

$$\alpha_n(\delta;\Delta) \leq 2\biggl[\mathbb{E}\alpha_n(\delta;\Delta) + C\delta\sqrt{\frac{t}{n}} + \frac{Ct}{n}\biggr]. \tag{8.40}$$

Next, symmetrization inequality followed by contraction inequality for Rademacher sums yield:

$$\begin{aligned} &\mathbb{E}\alpha_n(\delta;\Delta) \leq 2\mathbb{E}\sup\Bigl\{|R_n((\ell' \bullet f_\lambda)(f_\lambda - f_{\bar\lambda}))| : \lambda \in \Lambda(\delta;\Delta)\Bigr\} \leq \\ &C\mathbb{E}\sup\Bigl\{|R_n(f_\lambda - f_{\bar\lambda})| : \lambda \in \Lambda(\delta;\Delta)\Bigr\} \end{aligned} \tag{8.41}$$

with a constant C depending only on ℓ. In contraction inequality part, we write

$$\ell'(f_\lambda(\cdot))(f_\lambda(\cdot) - f_{\bar\lambda}(\cdot)) = \ell'(f_{\bar\lambda}(\cdot) + u)u\Big|_{u = f_\lambda(\cdot) - f_{\bar\lambda}(\cdot)}$$

and use the fact that the function

$$[-1,1] \ni u \mapsto \ell'(f_{\bar\lambda}(\cdot) + u)u$$

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

The following representation is straightforward:

$$f_\lambda - f_{\bar\lambda} = P_L(f_\lambda - f_{\bar\lambda}) + \sum_{j\in J}(\lambda_j - \bar\lambda_j)P_{L^\perp}h_j + \sum_{j\notin J}\lambda_j P_{L^\perp}h_j. \tag{8.42}$$

For all $\lambda \in \Lambda(\delta,\Delta)$,

$$\|P_L(f_\lambda - f_{\bar\lambda})\|_{L_2(\Pi)} \leq \|f_\lambda - f_{\bar\lambda}\|_{L_2(\Pi)} \leq \delta$$

and $P_L(f_\lambda - f_{\bar\lambda}) \in L$. Since L is a d-dimensional subspace,

$$\mathbb{E}\sup\Bigl\{|R_n(P_L(f_\lambda - f_{\bar\lambda}))| : \lambda \in \Lambda(\delta;\Delta)\Bigr\} \le C\delta\sqrt{\frac{d}{n}}$$

(see Proposition 3.2). On the other hand, $\lambda, \bar\lambda \in U_{\ell_1}$, so, we have $\sum_{j\in J}|\lambda_j - \bar\lambda_j| \le 2$. Hence,

$$\mathbb{E}\sup\Bigl\{\Bigl|R_n\Bigl(\sum_{j\in J}(\lambda_j - \bar\lambda_j)P_{L^\perp}h_j\Bigr)\Bigr| : \lambda \in \Lambda(\delta;\Delta)\Bigr\} \le 2\mathbb{E}\max_{j\in J}|R_n(P_{L^\perp}h_j)|.$$

Note also that

$$\|P_{L^\perp}h_j\|_\infty \le \|P_L h_j\|_\infty + \|h_j\|_\infty \le (U(L)-1)\|P_L h_j\|_{L_2(\Pi)} + 1$$
$$\le (U(L)-1)\|h_j\|_{L_2(\Pi)} + 1 \le U(L),$$

and Theorem 3.4 yields

$$\mathbb{E}\max_{j\in J}|R_n(P_{L^\perp}h_j)| \le C\Bigl[\max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{\log N}{n}} + U(L)\frac{\log N}{n}\Bigr].$$

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

$$\sum_{j\notin J}|\lambda_j| \le \Delta$$

and

$$\mathbb{E}\sup\Bigl\{\Bigl|R_n\Bigl(\sum_{j\notin J}\lambda_j P_{L^\perp}h_j\Bigr)\Bigr| : \lambda \in \Lambda(\delta;\Delta)\Bigr\} \le \Delta\mathbb{E}\max_{j\notin J}|R_n(P_{L^\perp}h_j)|.$$

Another application of Theorem 3.4, together with the fact that

$$\|P_{L^\perp}h_j\|_{L_2(\Pi)} \le \|h_j\|_{L_2(\Pi)} \le 1,$$

results in the bound

$$\mathbb{E}\max_{j\notin J}|R_n(P_{L^\perp}h_j)| \le C\Bigl[\sqrt{\frac{\log N}{n}} + U(L)\frac{\log N}{n}\Bigr],$$

Now we use representation (8.42) and bound (8.41). It follows that

$$\mathbb{E}\alpha_n(\delta,\Delta) \le C\Bigl[\delta\sqrt{\frac{d}{n}}\bigvee\Delta\sqrt{\frac{\log N}{n}}\bigvee\max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{\log N}{n}}\bigvee$$
$$\Delta U(L)\frac{\log N}{n}\bigvee U(L)\frac{\log N}{n}\Bigr]. \tag{8.43}$$

The right hand side can be bounded further as follows

$$\mathbb{E}\alpha_n(\delta,\Delta) \le C\left[\delta\sqrt{\frac{d}{n}}\bigvee \Delta\sqrt{\frac{\log N}{n}}\bigvee \max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{\log N}{n}}\bigvee \frac{U(L)\log N}{n}\right]. \tag{8.44}$$

Substituting this bound into (8.40) shows with probability $1-e^{-t}$

$$\begin{aligned}\alpha_n(\delta,\Delta) \le \tilde{\beta}_n(\delta,\Delta,t) := C\Bigg[\delta\sqrt{\frac{d}{n}}\bigvee \Delta\sqrt{\frac{\log N}{n}}\bigvee \\ \max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{\log N}{n}}\bigvee \frac{U(L)\log N}{n}\bigvee \delta\sqrt{\frac{t}{n}}\bigvee \frac{t}{n}\Bigg]\end{aligned} \tag{8.45}$$

with a constant $C>0$ depending only on ℓ.

It remains to prove that the above bound holds uniformly in δ,Δ satisfying (8.32) with a high probability. Let

$$\delta_j := 2^{-j} \text{ and } \Delta_j := 2^{-j}.$$

We will replace t by $t+2\log(j+1)+2\log(k+1)$. By the union bound, with probability at least

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

for all δ and Δ satisfying (8.32), and for j,k such that

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

the following bound holds:

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

Using the fact that

$$2\log j \le 2\log\log_2\Big(\frac{1}{\delta_j}\Big) \le 2\log\log_2\Big(\frac{2}{\delta}\Big)$$

and

$$2\log k \le 2\log\log_2\Big(\frac{2}{\Delta}\Big),$$

we get

$$\tilde{\beta}_n\Big(\delta_j,\Delta_k,t+2\log j+2\log k\Big) \le$$

$$\tilde{\beta}_n\Big(2\delta,2\Delta,t+2\log\log_2\Big(\frac{2}{\delta}\Big)+2\log\log_2\Big(\frac{2}{\Delta}\Big)\Big) =: \bar{\beta}_n(\delta;\Delta;t).$$

As a result, with probability at least $1-4e^{-t}$, for all δ and Δ satisfying (8.32),

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

Take now $t = A\log N + \log 4$ so that $4e^{-t} = N^{-A}$. With some constant C that depends only on ℓ,

$$\bar{\beta}_n(\delta;\Delta;t) \le C\Biggl[\delta\sqrt{\frac{d}{n}} \bigvee \delta\sqrt{\frac{A\log N}{n}} \bigvee \delta\sqrt{\frac{2\log\log_2\left(\frac{2}{\delta}\right)}{n}} \bigvee \delta\sqrt{\frac{2\log\log_2\left(\frac{2}{\Delta}\right)}{n}} \bigvee$$
$$\Delta\sqrt{\frac{\log N}{n}} \bigvee \max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee$$
$$\frac{2\log\log_2\left(\frac{2}{\delta}\right)}{n} \bigvee \frac{2\log\log_2\left(\frac{2}{\Delta}\right)}{n} \bigvee \frac{A\log N}{n}\Biggr].$$

For all δ and Δ satisfying (8.32),

$$\frac{2\log\log_2\left(\frac{2}{\delta}\right)}{n} \le C\frac{\log\log n}{n}$$

and

$$\frac{2\log\log_2\left(\frac{2}{\Delta}\right)}{n} \le C\frac{\log\log n}{n}.$$

Assumptions on N, n, imply that $A\log N \ge \gamma\log\log n$. Thus, for δ and Δ satisfying (8.32),

$$\alpha_n(\delta,\Delta) \le \bar{\beta}_n(\delta;\Delta;t) \le C\Biggl[\delta\sqrt{\frac{d}{n}} \bigvee \delta\sqrt{\frac{A\log N}{n}} \bigvee \Delta\sqrt{\frac{\log N}{n}} \bigvee$$
$$\max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Biggr]. \tag{8.46}$$

The last bound holds with probability at least $1-N^{-A}$ proving the lemma.

□

In theorems 8.4 and 8.5, we used a special version of subgradient $\nabla L_\varepsilon(\bar{\lambda})$. More generally, one can consider an arbitrary couple $(\bar{\lambda}, \nabla L_\varepsilon(\bar{\lambda}))$ where $\bar{\lambda}\in U_{\ell_1}$ and $\nabla L_\varepsilon(\bar{\lambda}) \in \partial L_\varepsilon(\bar{\lambda})$. This couple can be viewed as "an oracle" in our problem. As before,

$$\nabla L_\varepsilon(\bar{\lambda}) = \Bigl((P(\ell'\bullet f_{\bar{\lambda}}))h_j + \varepsilon s_j\Bigr)_{j=1,\dots,N},$$

but now $s_j = s_j(\bar\lambda)$ are arbitrary numbers from $[-1,1]$ satisfying the condition

$$s_j = \mathrm{sign}(\bar\lambda_j), \quad \bar\lambda_j \neq 0.$$

Denote

$$\alpha^{(b)}(\lambda) := a_H^{(b)}\Bigl(U_{\ell_1}, \lambda, s(\lambda)\Bigr) \vee 0$$

with some $b > 0$.

Theorem 8.6 *There exist constants $D > 0$ and $C > 0$ depending only on ℓ with the following property. Let $\bar\lambda \in U_{\ell_1}$ and*

$$\nabla L_\varepsilon(\bar\lambda) = \Bigl((P(\ell' \bullet f_{\bar\lambda}))h_j + \varepsilon s_j\Bigr)_{j=1,\dots,N} \in \partial L_\varepsilon(\bar\lambda).$$

Let $J \subset \{1,\dots,N\}$ with $d := d(J) = \mathrm{card}(J)$. Suppose that, for some $\gamma \in (0,1)$,

$$|s_j| \leq 1-\gamma,\ j \notin J.$$

Then, for all $A \geq 1$ and for all

$$\varepsilon \geq D\sqrt{\frac{d + A\log N}{n}}, \tag{8.47}$$

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

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon\gamma\sum_{j\notin J}|\hat\lambda^\varepsilon_j| \leq C\biggl[\frac{d + A\log N}{n}\bigvee \alpha_+^2(\varepsilon,\bar\lambda)\biggr].$$

Moreover, with the same probability,

$$\|f_{\hat\lambda^\varepsilon} - f_{\bar\lambda}\|^2_{L_2(\Pi)} + \varepsilon\gamma\sum_{j\notin J}|\hat\lambda^\varepsilon_j| \leq C\biggl[\frac{d + A\log N}{n}\bigvee\Bigl\|\ell' \bullet f_{\bar\lambda}\Bigr\|^2_{L_2(P)}\bigvee\Bigl(\alpha^{(2/\gamma)}(\bar\lambda)\Bigr)^2\varepsilon^2\biggr].$$

Theorem 8.7 *Suppose that*

$$\varepsilon \geq D\sqrt{\frac{A\log N}{n}} \tag{8.48}$$

with a large enough constant $D > 0$ depending only on ℓ. Let $\bar\lambda \in U_{\ell_1}$ and

$$\nabla L_\varepsilon(\bar\lambda) = \Bigl((P(\ell' \bullet f_{\bar\lambda}))h_j + \varepsilon s_j\Bigr)_{j=1,\dots,N} \in \partial L_\varepsilon(\bar\lambda).$$

Let $J \subset \{1,\dots,N\}$. Suppose that, for some $\gamma \in (0,1)$,

$$|s_j| \leq 1-\gamma,\ j \notin J.$$

Then, for all subspaces L of $L_2(\Pi)$ with $d := \dim(L)$ and for all $A \geq 1$, the following bound holds with probability at least $1 - N^{-A}$ and with a constant $C > 0$ depending only on ℓ :

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}\|^2_{L_2(\Pi)} + \varepsilon\gamma \sum_{j \notin J} |\hat{\lambda}^\varepsilon_j| \leq \tag{8.49}$$

$$C\biggl[\frac{d + A\log N}{n} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \alpha^2_+(\varepsilon; \bar{\lambda})\biggr].$$

Moreover, with the same probability

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\bar{\lambda}}\|^2_{L_2(\Pi)} + \varepsilon\gamma \sum_{j \notin J} |\hat{\lambda}^\varepsilon_j| \leq \tag{8.50}$$

$$C\biggl[\frac{d + A\log N}{n} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n}$$

$$\bigvee \Bigl\|\ell' \bullet f_{\bar{\lambda}}\Bigr\|^2_{L_2(P)} \bigvee \Bigl(\alpha^{(2/\gamma)}(\bar{\lambda})\Bigr)^2 \varepsilon^2\biggr].$$

For some choices of vector $\bar{\lambda}$ and of subgradient $\nabla L_\varepsilon(\bar{\lambda})$, the alignment coefficient might be smaller than for the choice we used in theorems 8.4 and 8.5 resulting in tighter bounds. An appealing choice would be $\bar{\lambda} = \lambda^\varepsilon$,

$$\lambda^\varepsilon = \text{argmin}_{\lambda \in U_{\ell_1}} \biggl[P(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|_{\ell_1}\biggr],$$

since in this case it is possible to take $\nabla L_\varepsilon(\lambda^\varepsilon) \in \partial L_\varepsilon(\lambda^\varepsilon)$ such that

$$a_H^{(b)}(U_{\ell_1}, \lambda^\varepsilon, \nabla L_\varepsilon(\lambda^\varepsilon)) \leq 0$$

(this follows from the necessary conditions of extremum). Therefore, with this choice, $\alpha_+(\varepsilon, \lambda^\varepsilon) = 0$, implying the following corollaries.

Corollary 8.3 *There exist constants $D > 0$ and $C > 0$ depending only on ℓ with the following property. Let*

$$\nabla L_\varepsilon(\lambda^\varepsilon) = \Bigl((P(\ell' \bullet f_{\lambda^\varepsilon}))h_j + \varepsilon s_j\Bigr)_{j=1,\ldots,N} \in \partial L_\varepsilon(\lambda^\varepsilon)$$

be such that, for all $u \in T_{U_{\ell_1}}(\lambda^\varepsilon)$,

$$\langle \nabla L_\varepsilon(\lambda^\varepsilon), u\rangle_{\ell_2} \geq 0.$$

Let $J \subset \{1, \dots, N\}$ *with* $d := d(J) = \mathrm{card}(J)$. *Suppose that, for some* $\gamma \in (0,1)$,

$$|s_j| \leq 1 - \gamma,\ j \notin J.$$

Then, for all $A \geq 1$ *and for all*

$$\varepsilon \geq D\sqrt{\frac{d + A\log N}{n}}, \tag{8.51}$$

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

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon\gamma\sum_{j \notin J}|\hat{\lambda}^\varepsilon_j| \leq C\frac{d + A\log N}{n}.$$

Corollary 8.4 *Suppose that*

$$\varepsilon \geq D\sqrt{\frac{A\log N}{n}} \tag{8.52}$$

with a large enough constant $D > 0$ *depending only on* ℓ. *Let*

$$\nabla L_\varepsilon(\lambda^\varepsilon) = \Bigl((P(\ell' \bullet f_{\lambda^\varepsilon}))h_j + \varepsilon s_j\Bigr)_{j=1,\dots,N} \in \partial L_\varepsilon(\lambda^\varepsilon)$$

be such that, for all $u \in T_{U_{\ell_1}}(\lambda^\varepsilon)$,

$$\langle \nabla L_\varepsilon(\lambda^\varepsilon), u\rangle_{\ell_2} \geq 0.$$

Let $J \subset \{1, \dots, N\}$. *Suppose that, for some* $\gamma \in (0,1)$,

$$|s_j| \leq 1 - \gamma,\ j \notin J.$$

Then, for all subspaces L *of* $L_2(\Pi)$ *with* $d := \dim(L)$ *and for all* $A \geq 1$, *the following bound holds with probability at least* $1 - N^{-A}$ *and with a constant* $C > 0$ *depending only on* ℓ :

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon\gamma\sum_{j \notin J}|\hat{\lambda}^\varepsilon_j| \leq \tag{8.53}$$

$$C\biggl[\frac{d + A\log N}{n}\bigvee \max_{j \in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}}\bigvee\frac{U(L)\log N}{n}\biggr].$$

9 Strictly Convex Penalization in Sparse Recovery

In this section, we study two examples of problem (8.1) with *strictly convex and smooth* penalty function ψ. The first example (that will be discussed in more detail) deals with sparse recovery in convex hulls with negative entropy penalization, i.e., $\psi(u) = u \log u$. In the second example, we consider sparse recovery in the ℓ_p-ball for $p > 1$, the penalty being the p-th power of the ℓ_p-norm, i.e., $\psi(u) = u^p$. More details on these problems are given in [65, 67]. It happens that strict convexity and smoothness of the penalty give some advantages in the analysis of the problem. In particular, it is possible in such cases to study the random error $|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})|$ completely separately from the approximation error $\mathcal{E}(f_{\lambda^\varepsilon})$. If the solution λ^ε of the true penalized problem (8.6) is approximately sparse, there is a way to control the size of the random error in terms of its sparsity without any restrictive assumptions on the dictionary. However, the control of approximation error still requires some assumption on the Gram matrix of the dictionary that can be expressed, for instance, in terms of alignment coefficients introduced and used in the previous sections.

9.1 Entropy Penalization and Sparse Recovery in Convex Hulls: Random Error Bounds

As before, it will be assumed that ℓ is a loss function of quadratic type (see Definition 8.1).

Denote

$$\Lambda := \{(\lambda_1, \dots, \lambda_N) : \ \lambda_j \geq 0, \ j = 1, \dots, N, \ \sum_{j=1}^{N} \lambda_j = 1\}.$$

The following penalized empirical risk minimization problem will be studied:

$$\hat{\lambda}^\varepsilon := \mathrm{argmin}_{\lambda \in \Lambda} \Big[P_n(\ell \bullet f_\lambda) - \varepsilon H(\lambda) \Big] = \mathrm{argmin}_{\lambda \in \Lambda} \Big[P_n(\ell \bullet f_\lambda) + \varepsilon \sum_{j=1}^{N} \lambda_j \log \lambda_j \Big], \quad (9.1)$$

where $\varepsilon \geq 0$ is a regularization parameter and

$$H(\lambda) = -\sum_{j=1}^{N} \lambda_j \log \lambda_j$$

is the entropy of λ. Since, for all y, $\ell(y, \cdot)$ is convex, the empirical risk $P_n(\ell \bullet f_\lambda)$ is a convex function of λ. Since also the set Λ is convex and so is the function $\lambda \mapsto -H(\lambda)$, the problem (9.1) a convex optimization problem.

It is natural to compare this problem with its distribution dependent version

$$\lambda^{\varepsilon} := \operatorname{argmin}_{\lambda\in\Lambda}\biggl[P(\ell\bullet f_{\lambda}) - \varepsilon H(\lambda)\biggr] = \operatorname{argmin}_{\lambda\in\Lambda}\biggl[P(\ell\bullet f_{\lambda}) + \varepsilon\sum_{j=1}^{N}\lambda_j\log\lambda_j\biggr]. \quad (9.2)$$

There has been a considerable amount of work on entropy penalization in information theory and statistics, for instance, in problems of aggregation of statistical estimators using exponential weighting and in PAC-Bayesian methods of learning theory (see, e.g., McAllester [75], Catoni [31], Audibert [4], Zhang [99, 100, 101] and references therein). Dalalyan and Tsybakov [34] studied PAC-Bayesian method with special priors in sparse recovery problems. However, the minimum of the penalty $-H(\lambda)$ is attained at the uniform distribution $\lambda_j = N^{-1}, j = 1,\dots,N$. Because of this, at the first glance, $-H(\lambda)$ penalizes for "sparsity" rather than for "non-sparsity".

We will show that if λ^{ε} is *"approximately sparse",* then $\hat{\lambda}^{\varepsilon}$ has a similar property with a high probability. Moreover, the approximate sparsity of λ^{ε} will allow us to control $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$ and $K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon})$, where

$$K(\lambda,\nu) := K(\lambda|\nu) + K(\nu|\lambda)$$

is the symmetrized Kullback-Leibler distance between λ and ν,

$$K(\lambda|\nu) := \sum_{j=1}^{N}\lambda_j\log\biggl(\frac{\lambda_j}{\nu_j}\biggr)$$

being the Kullback-Leibler divergence between λ, ν.

In particular, it will follow from our results that for any set $J\subset\{1,\dots,N\}$ with $\mathrm{card}(J) = d$ and such that

$$\sum_{j\notin J}\lambda_j^{\varepsilon} \le \sqrt{\frac{\log N}{n}},$$

with a high probability,

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

This easily implies upper bounds on "the random error" $|\mathcal{E}(f_{\hat{\lambda}^{\varepsilon}}) - \mathcal{E}(f_{\lambda^{\varepsilon}})|$ in terms of "approximate sparsity" of λ^{ε}.

Some further geometric parameters (such as "the alignment coefficient" introduced in Section 7.2) provide a way to control "the approximation error" $\mathcal{E}(f_{\lambda^{\varepsilon}})$. As a result, if there exists a "sparse" vector $\lambda\in\Lambda$ for which the excess risk $\mathcal{E}(f_{\lambda})$ is small and λ is properly "aligned" with the dictionary, then λ^ε is approximately sparse and its excess risk $\mathcal{E}(f_{\lambda^\varepsilon})$ is controlled by sparsity of λ and its "alignment" with the dictionary. Together with sparsity bounds on the random error this yields oracle inequalities on the excess risk $\mathcal{E}(f_{\hat\lambda^\varepsilon})$ showing that this estimation method provides certain degree of adaptation to the unknown "sparsity" of the problem.

The first result in this direction is the following theorem that provides the bounds on approximate sparsity of $\hat\lambda^\varepsilon$ in terms of approximate sparsity of λ^ε as well as the bounds on the L_2-error of approximation of f_{λ^ε} by $f_{\hat\lambda^\varepsilon}$ and the Kullback-Leibler error of approximation of λ^ε by $\hat\lambda^\varepsilon$.

Theorem 9.1 *There exist constants $D > 0$ and $C > 0$ depending only on ℓ such that, for all $J \subset \{1, \dots, N\}$ with $d := d(J) = \mathrm{card}(J)$, for all $A \ge 1$ and for all*

$$\varepsilon \ge D\sqrt{\frac{d + A\log N}{n}}, \tag{9.3}$$

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

$$\sum_{j \notin J} \hat\lambda_j^\varepsilon \le C\biggl[\sum_{j \notin J} \lambda_j^\varepsilon + \sqrt{\frac{d + A\log N}{n}}\biggr],$$

$$\sum_{j \notin J} \lambda_j^\varepsilon \le C\biggl[\sum_{j \notin J} \hat\lambda_j^\varepsilon + \sqrt{\frac{d + A\log N}{n}}\biggr]$$

and

$$\|f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}^2 + \varepsilon K(\hat\lambda^\varepsilon, \lambda^\varepsilon) \le C\biggl[\frac{d + A\log N}{n} \bigvee \sum_{j \notin J} \lambda_j^\varepsilon \sqrt{\frac{d + A\log N}{n}}\biggr].$$

Similarly to what was done in Section 8.2, we will also establish another version of these bounds that hold for smaller values of ε (the quantity $U(L)$ introduced in Section 8.2 will be involved in these bounds).

Theorem 9.2 *Suppose that*

$$\varepsilon \ge D\sqrt{\frac{A\log N}{n}} \tag{9.4}$$

with a large enough constant $D > 0$ depending only on ℓ. For all $J \subset \{1, \dots, N\}$, for all subspaces L of $L_2(\Pi)$ with $d := \dim(L)$ and for all $A \ge 1$, the following bounds hold with probability at least $1 - N^{-A}$ and with a constant $C > 0$ depending only on ℓ :

$$\sum_{j \notin J} \hat\lambda_j^\varepsilon \le C\biggl[\sum_{j \notin J} \lambda_j^\varepsilon + \frac{d + A\log N}{n\varepsilon} + \max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} + \frac{U(L)\log N}{n\varepsilon}\biggr], \tag{9.5}$$

$$\sum_{j\notin J} \lambda_j^\varepsilon \le C\Bigl[\sum_{j\notin J} \hat\lambda_j^\varepsilon + \frac{d + A\log N}{n\varepsilon} + \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} + \frac{U(L)\log N}{n\varepsilon}\Bigr] \tag{9.6}$$

and

$$\|f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}^2 + \varepsilon K(\hat\lambda^\varepsilon, \lambda^\varepsilon) \le C\biggl[\frac{d + A\log N}{n} \bigvee \sum_{j\notin J} \lambda_j^\varepsilon \sqrt{\frac{A\log N}{n}} \bigvee$$

$$\max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n}\biggr]. \tag{9.7}$$

If, for some J,

$$\sum_{j\notin J} \lambda_j^\varepsilon \le \sqrt{\frac{A\log N}{n}}$$

and, for some L with $U(L) \le d$, $h_j \in L$, $j \in J$, then the bound (9.7) simplifies and becomes

$$\|f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}^2 + \varepsilon K(\hat\lambda^\varepsilon, \lambda^\varepsilon) \le C\frac{Ad\log N}{n}.$$

In particular, it means that the size of the random errors $\|f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}^2$ and $K(\hat\lambda^\varepsilon, \lambda^\varepsilon)$ is controlled by the dimension d of the linear span L of the "relevant part" of the dictionary $\{h_j : j \in J\}$. Note that d can be much smaller than card(J) in the case when the functions in the dictionary are not linearly independent (so, the lack of "orthogonality" of the dictionary might help to reduce the random error).

The proofs of theorems 9.1 and 9.2 are quite similar. We give only the proof of Theorem 9.2.

Proof of Theorem 9.2. We use the method described in Section 8.1. In the current case, necessary conditions of minima in minimization problems defining λ^ε and $\hat\lambda^\varepsilon$ can be written as follows:

$$P(\ell' \bullet f_{\lambda^\varepsilon})(f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}) + \varepsilon \sum_{j=1}^N (\log \lambda_j^\varepsilon + 1)(\hat\lambda_j^\varepsilon - \lambda_j^\varepsilon) \ge 0 \tag{9.8}$$

and

$$P_n(\ell' \bullet f_{\hat\lambda^\varepsilon})(f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}) + \varepsilon \sum_{j=1}^N (\log \hat\lambda_j^\varepsilon + 1)(\hat\lambda_j^\varepsilon - \lambda_j^\varepsilon) \le 0. \tag{9.9}$$

The inequality (9.8) follows from the fact that the directional derivative of the penalized risk function (note that it is smooth and convex)

$$\Lambda \ni \lambda \mapsto P(\ell \bullet f_\lambda) + \varepsilon \sum_{j=1}^N \lambda_j \log \lambda_j$$

at the point of its minimum λ^{ε} is nonnegative in the direction of any point of the convex set Λ, in particular, in the direction of $\hat{\lambda}^{\varepsilon}$. The same observation in the case of penalized empirical risk lead to inequality (9.9). Subtract (9.8) from (9.9) and replace P by P_n in (9.9) to get

$$P\Big((\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}) - (\ell' \bullet f_{\lambda^{\varepsilon}})\Big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}) + \varepsilon \sum_{j=1}^{N} \Big(\log \hat{\lambda}_j^{\varepsilon} - \log \lambda_j^{\varepsilon}\Big)(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon})$$
$$\leq (P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}). \tag{9.10}$$

It is easy to see that

$$\sum_{j=1}^{N} \Big(\log \hat{\lambda}_j^{\varepsilon} - \log \lambda_j^{\varepsilon}\Big)(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon}) = \sum_{j=1}^{N} \left(\log \frac{\hat{\lambda}_j^{\varepsilon}}{\lambda_j^{\varepsilon}}\right)(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon}) = K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon})$$

and rewrite bound (9.10) as

$$P\Big((\ell' \bullet f_{\hat{\lambda}^{\varepsilon}}) - (\ell' \bullet f_{\lambda^{\varepsilon}})\Big)(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}) + \varepsilon K(\hat{\lambda}^{\varepsilon}; \lambda^{\varepsilon})$$
$$\leq (P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}). \tag{9.11}$$

We use the following simple inequality

$$K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon}) = \sum_{j=1}^{N} \left(\log \frac{\hat{\lambda}_j^{\varepsilon}}{\lambda_j^{\varepsilon}}\right)(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon}) \geq$$
$$\frac{\log 2}{2} \sum_{j:\hat{\lambda}_j^{\varepsilon} \geq 2\lambda_j^{\varepsilon}} \hat{\lambda}_j^{\varepsilon} + \frac{\log 2}{2} \sum_{j:\lambda_j^{\varepsilon} \geq 2\hat{\lambda}_j^{\varepsilon}} \lambda_j^{\varepsilon}, \tag{9.12}$$

which implies that for all $J \subset \{1, \dots, N\}$

$$\sum_{j \notin J} \hat{\lambda}_j^{\varepsilon} \leq 2 \sum_{j \notin J} \lambda_j^{\varepsilon} + \frac{2}{\log 2} K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon}) \tag{9.13}$$

and

$$\sum_{j \notin J} \lambda_j^{\varepsilon} \leq 2 \sum_{j \notin J} \hat{\lambda}_j^{\varepsilon} + \frac{2}{\log 2} K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon}). \tag{9.14}$$

If $K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon})$ is small, the last bounds show that "sparsity patterns" of vectors $\hat{\lambda}^{\varepsilon}$ and λ^{ε} are closely related. Then, it follows from (9.11) that

$$\varepsilon \sum_{j \notin J} \hat{\lambda}_j^{\varepsilon} \leq 2\varepsilon \sum_{j \notin J} \lambda_j^{\varepsilon} + \frac{2}{\log 2} (P - P_n)(\ell' \bullet f_{\hat{\lambda}^{\varepsilon}})(f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}). \tag{9.15}$$

As in the previous section, for the loss functions of quadratic type,

$$P\Big((\ell' \bullet f_{\hat{\lambda}^\varepsilon}) - (\ell' \bullet f_{\lambda^\varepsilon})\Big)(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}) \geq c\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2,$$

where $c = \tau(1)$. Note that $\|f_{\lambda^\varepsilon}\|_\infty \leq 1$ and $\|f_{\hat{\lambda}^\varepsilon}\|_\infty \leq 1$. Bound (9.11) then yields

$$c\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2 + \varepsilon K(\hat{\lambda}^\varepsilon, \lambda^\varepsilon) \leq (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}). \tag{9.16}$$

Following the methodology of Section 8.1, we have now to control the empirical process $(P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon})$. To this end, let

$$\Lambda(\delta; \Delta) := \Big\{\lambda \in \Lambda : \|f_\lambda - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \leq \delta,\ \sum_{j \notin J} \lambda_j \leq \Delta\Big\}$$

and

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

The following two lemmas are similar to lemmas 8.4 and 8.3 of the previous section. Their proof is also similar and we skip it.

Lemma 9.1 *Under the assumptions of Theorem 9.1, there exists constant C that depends only on ℓ such that with probability at least $1 - N^{-A}$, for all*

$$n^{-1/2} \leq \delta \leq 1 \quad \text{and} \quad n^{-1/2} \leq \Delta \leq 1$$

the following bounds hold:

$$\alpha_n(\delta; \Delta) \leq \beta_n(\delta; \Delta) := C\Bigg[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \Delta\sqrt{\frac{d + A\log N}{n}} \\ \bigvee \sum_{j \notin J} \lambda_j^\varepsilon \sqrt{\frac{d + A\log N}{n}} \bigvee \frac{A\log N}{n}\Bigg]. \tag{9.17}$$

Lemma 9.2 *Under the assumptions of Theorem 9.2, there exists constant C that depends only on ℓ such that with probability at least $1 - N^{-A}$, for all*

$$n^{-1/2} \leq \delta \leq 1 \quad \text{and} \quad n^{-1/2} \leq \Delta \leq 1 \tag{9.18}$$

the following bounds hold:

$$\alpha_n(\delta; \Delta) \leq \beta_n(\delta; \Delta) := C\Bigg[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \Delta\sqrt{\frac{A\log N}{n}} \\ \bigvee \sum_{j \notin J} \lambda_j^\varepsilon \sqrt{\frac{A\log N}{n}} \bigvee \max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \\ \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Bigg]. \tag{9.19}$$

We now proceed exactly as in the proof of Theorem 8.5. Let

$$\delta = \|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)} \quad \text{and} \quad \Delta = \sum_{j \notin J} \hat{\lambda}_j^{\varepsilon}, \tag{9.20}$$

and suppose $\delta \geq n^{-1/2}, \Delta \geq n^{-1/2}$. Then, by Lemma 9.2 and bounds (9.16), (9.15), the following bounds hold with probability at least $1 - N^{-A}$:

$$c\delta^2 \leq \beta_n(\delta, \Delta) \tag{9.21}$$

and

$$\varepsilon\Delta \leq 2\varepsilon \sum_{j \notin J} \lambda_j^{\varepsilon} + \frac{2}{\log 2}\beta_n(\delta, \Delta), \tag{9.22}$$

where $\beta_n(\delta, \Delta)$ is defined in (9.19) (as in the proof of Theorem 8.5, the case $\delta < n^{-1/2}$ or $\Delta < n^{-1/2}$ is even simpler). Thus, it remains to solve the inequalities (9.21), (9.22) to complete the proof. First, rewrite (9.22) (with a possible change of constant C) as

$$\varepsilon\Delta \leq C\Delta\sqrt{\frac{A\log N}{n}} + C\Bigg[\varepsilon\sum_{j\notin J}\lambda_j^{\varepsilon} \bigvee \delta\sqrt{\frac{d + A\log N}{n}} \bigvee$$

$$\sum_{j\notin J}\lambda_j^{\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee \max_{j\in J}\|P_{L^{\perp}}h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Bigg].$$

If the constant D in condition (9.4) satisfies $D \geq 2C \vee 1$, then the term $\sum_{j\notin J}\lambda_j^{\varepsilon}\sqrt{\frac{A\log N}{n}}$ in the maximum can be dropped since it smaller than the first term $\varepsilon\sum_{j\notin J}\lambda_j^{\varepsilon}$, and the bound can be written as follows

$$\varepsilon\Delta \leq C\Bigg[\varepsilon\sum_{j\notin J}\lambda_j^{\varepsilon} \bigvee \delta\sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J}\|P_{L^{\perp}}h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Bigg]$$

We solve the inequality separately for each term in the maximum and take the maximum of the solutions. This yields

$$\Delta \leq C\Bigg[\sum_{j\notin J}\lambda_j^{\varepsilon} \bigvee \frac{\delta}{\varepsilon}\sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J}\|P_{L^{\perp}}h_j\|_{L_2(\Pi)}\frac{1}{\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n\varepsilon} \bigvee \frac{A\log N}{n\varepsilon}\Bigg],$$

which, under the assumption (9.4) with $D \geq 1$, can be upper bounded as follows

$$\Delta \leq \Delta(\delta) := C\Bigg[\sum_{j\notin J}\lambda_j^{\varepsilon} \bigvee \frac{\delta}{\varepsilon}\sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J}\|P_{L^{\perp}}h_j\|_{L_2(\Pi)} \bigvee \frac{U(L)\log N}{n\varepsilon} \bigvee \sqrt{\frac{A\log N}{n}}\Bigg].$$

Using the fact that $\beta_n(\delta, \Delta)$ is nondecreasing in Δ, substituting $\Delta(\delta)$ instead of Δ in (9.21) and dropping the smallest terms, we get

$$\delta^2 \le C\Big[\delta\sqrt{\frac{d + A\log N}{n}} \bigvee \sum_{j\notin J} \lambda_j^{\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee$$

$$\max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n} \bigvee \frac{A\log N}{n}\Big].$$

Solving the inequality yields the following bound on δ^2 :

$$\delta^2 \le C\Big[\frac{d + A\log N}{n} \bigvee \sum_{j\notin J} \lambda_j^{\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n}\Big]. \tag{9.23}$$

We substitute this into the expression for $\Delta(\delta)$ which results in the following bound on Δ :

$$\Delta \le C\Big[\sum_{j\notin J} \lambda_j^{\varepsilon} \bigvee \frac{d + A\log N}{n\varepsilon} \bigvee \Big(\sum_{j\notin J} \lambda_j^{\varepsilon}\Big)^{1/2} \frac{1}{\varepsilon}\Big(\frac{A\log N}{n}\Big)^{1/4} \sqrt{\frac{d + A\log N}{n}} \bigvee$$

$$\sqrt{\frac{U(L)\log N}{n\varepsilon}} \sqrt{\frac{d + A\log N}{n}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}^{1/2} \frac{1}{\varepsilon}\Big(\frac{A\log N}{n}\Big)^{1/4} \sqrt{\frac{d + A\log N}{n}} \bigvee$$

$$\max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \bigvee \sqrt{\frac{A\log N}{n}}\Big],$$

The inequality $ab \le (a^2 + b^2)/2$ and the condition $\frac{1}{\varepsilon}\sqrt{\frac{A\log N}{n}} \le 1$, allows us to simplify the last bound and to get

$$\Delta \le C\Big[\sum_{j\notin J} \lambda_j^{\varepsilon} \bigvee \frac{d + A\log N}{n\varepsilon} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \bigvee \frac{U(L)\log N}{n\varepsilon} \bigvee \sqrt{\frac{A\log N}{n}}\Big] \tag{9.24}$$

with a constant C depending only on ℓ. Substitute bounds (9.23) and (9.24) in the expression for $\beta_n(\delta, \Delta)$. With a little further work and using Lemma 9.2, we get the following bound on $\alpha_n(\delta, \Delta)$ that holds for δ, Δ defined by (9.20) with probability at least $1 - N^{-A}$:

$$\alpha_n(\delta, \Delta) \le C\Big[\frac{d + A\log N}{n} + \sum_{j\notin J} \lambda_j^{\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{A\log N}{n}}$$

$$\bigvee \frac{U(L)\log N}{n}\Big].$$

This bound and (9.16) implies that

$$c\|f_{\hat\lambda^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon K(\hat\lambda^\varepsilon, \lambda^\varepsilon) \le C\biggl[\frac{d + A\log N}{n} + \sum_{j\notin J}\lambda_j^\varepsilon\sqrt{\frac{A\log N}{n}}\bigvee$$

$$\max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}}\bigvee\frac{U(L)\log N}{n}\biggr], \tag{9.25}$$

and (9.7) follows. Bound (9.5) is an immediate consequence of (9.24); bound (9.6) follows from (9.14) and (9.25).

□

From theorems 9.1, 9.2 and the properties of the loss function, we will easily deduce the following result.

Let $\mathcal{L}$ be the linear span of the dictionary $\{h_1, \dots, h_N\}$ and let $P_{\mathcal{L}}$ be the orthogonal projector on $\mathcal{L} \subset L_2(P)$. Define

$$g_\varepsilon := P_{\mathcal{L}}(\ell' \bullet f_{\lambda^\varepsilon}).$$

Theorem 9.3 *Under the conditions of Theorem 9.1, the following bound holds with probability at least* $1 - N^{-A}$, *with a constant* $C > 0$ *depending only on* ℓ *and with* $d = \mathrm{card}(J)$:

$$\Bigl|P(\ell\bullet f_{\hat\lambda^\varepsilon}) - P(\ell\bullet f_{\lambda^\varepsilon})\Bigr| \le C\biggl[\frac{d + A\log N}{n}\bigvee\sum_{j\notin J}\lambda_j^\varepsilon\sqrt{\frac{d + A\log N}{n}}\biggr]\bigvee$$

$$C^{1/2}\|g_\varepsilon\|_{L_2(\Pi)}\biggl[\frac{d + A\log N}{n}\bigvee\sum_{j\notin J}\lambda_j^\varepsilon\sqrt{\frac{d + A\log N}{n}}\biggr]^{1/2}. \tag{9.26}$$

Similarly, under the conditions of Theorem 9.2, with probability at least $1 - N^{-A}$ *and with* $d = \dim(L)$

$$\Bigl|P(\ell\bullet f_{\hat\lambda^\varepsilon}) - P(\ell\bullet f_{\lambda^\varepsilon})\Bigr| \le$$

$$C\biggl[\frac{d + A\log N}{n}\bigvee\Bigl(\sum_{j\notin J}\lambda_j^\varepsilon\bigvee\max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\Bigr)\sqrt{\frac{A\log N}{n}}\bigvee\frac{U(L)\log N}{n}\biggr]\bigvee$$

$$C^{1/2}\|g_\varepsilon\|_{L_2(\Pi)}\biggl[\frac{d + A\log N}{n}\bigvee\Bigl(\sum_{j\notin J}\lambda_j^\varepsilon\bigvee\max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\Bigr)\sqrt{\frac{A\log N}{n}}$$

$$\bigvee\frac{U(L)\log N}{n}\biggr]^{1/2}. \tag{9.27}$$

Proof of Theorem 9.3. For the losses of quadratic type,

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

where

$$|R(x,y)| \leq C(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon})^2(x).$$

Integrate with respect to P and get

$$\Bigl|P(\ell \bullet f_{\hat{\lambda}^\varepsilon}) - P(\ell \bullet f_{\lambda^\varepsilon}) - P(\ell' \bullet f_{\lambda^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon})\Bigr| \leq C\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)}.$$

Since

$$\Bigl|P(\ell' \bullet f_{\lambda^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon})\Bigr| = \Bigl|\Bigl\langle \ell' \bullet f_{\lambda^\varepsilon}, f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\Bigr\rangle_{L_2(P)}\Bigr| = \Bigl|\Bigl\langle P_{\mathcal{L}}(\ell' \bullet f_{\lambda^\varepsilon}), f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\Bigr\rangle_{L_2(P)}\Bigr| \leq$$

$$\|g_\varepsilon\|_{L_2(P)}\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}$$

theorems 9.1 and 9.2 imply the result.

□

Recall that f_* is a function that minimizes the risk $P(\ell \bullet f)$ and that f_* is uniformly bounded by a constant M. It follows from necessary conditions of minimum that

$$P(\ell' \bullet f_*)h_j = 0, \quad j = 1, \dots, N,$$

or $\ell' \bullet f_* \in \mathcal{L}^\perp$. For any function $\bar{f}$ uniformly bounded by M and such that $\ell' \bullet \bar{f} \in \mathcal{L}^\perp$ (for instance, for f_*), the following bounds hold

$$\|g_\varepsilon\|_{L_2(\Pi)} = \|P_{\mathcal{L}}(\ell' \bullet f_{\lambda^\varepsilon})\|_{L_2(P)} = \|P_{\mathcal{L}}(\ell' \bullet f_{\lambda^\varepsilon} - \ell' \bullet \bar{f})\|_{L_2(P)} \leq$$

$$\|(\ell' \bullet f_{\lambda^\varepsilon} - \ell' \bullet \bar{f})\|_{L_2(P)} \leq C\|f_{\lambda^\varepsilon} - \bar{f}\|_{L_2(\Pi)}$$

since ℓ' is Lipschitz with respect to the second variable.

Since ℓ is the loss of quadratic type, we have, for all $\lambda \in \Lambda$,

$$\mathcal{E}(f_\lambda) \geq \frac{1}{2}\tau(\|f_*\|_\infty \vee 1)\|f_\lambda - f_*\|^2_{L_2(\Pi)} =: \tau\|f_\lambda - f_*\|^2_{L_2(\Pi)}. \tag{9.28}$$

Theorem 9.3 implies the following bound on the random error

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

under the conditions of Theorem 9.1, with probability at least $1 - N^{-A}$

$$\left|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\right| \leq C\left[\frac{d + A\log N}{n} \bigvee \sum_{j \notin J} \lambda_j^\varepsilon \sqrt{\frac{d + A\log N}{n}}\right] \bigvee$$

$$C^{1/2}\sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau}}\left[\frac{d + A\log N}{n} \bigvee \sum_{j \notin J} \lambda_j^\varepsilon \sqrt{\frac{d + A\log N}{n}}\right]^{1/2}, \tag{9.29}$$

where $d = d(J)$, and under the conditions of Theorem 9.2, with probability at least $1 - N^{-A}$

$$\left|\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) - \mathcal{E}(f_{\lambda^\varepsilon})\right| \leq$$

$$C\left[\frac{d + A\log N}{n} \bigvee \Bigl(\sum_{j \notin J} \lambda_j^\varepsilon \bigvee \max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}\Bigr) \sqrt{\frac{A\log N}{n}}\right.$$

$$\left. \bigvee \frac{U(L)\log N}{n}\right] \bigvee$$

$$C^{1/2}\sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau}}\left[\frac{d + A\log N}{n} \bigvee \Bigl(\sum_{j \notin J} \lambda_j^\varepsilon \bigvee \max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}\Bigr) \sqrt{\frac{A\log N}{n}} \bigvee\right.$$

$$\left.\frac{U(L)\log N}{n}\right]^{1/2}, \tag{9.30}$$

where $d = \dim(L)$.

9.2 Approximation Error Bounds, Alignment and Oracle Inequalities

To consider the approximation error, we will use the definitions of alignment coefficients from Section 7.2.

For $\lambda \in \mathbb{R}^N$, let $s_j^N(\lambda) := \log(eN^2\lambda_j),\ j \in \mathrm{supp}(\lambda)$ and $s_j^N(\lambda) := 0,\ j \notin \mathrm{supp}(\lambda)$. Note that $\log\lambda_j + 1$ is the derivative of the function $\lambda\log\lambda$ involved in the definition of the penalty and, for $j \in \mathrm{supp}(\lambda)$, $s_j^N(\lambda) = \log\lambda_j + 1 + 2\log N$. Introduce the following vector

$$s^N(\lambda) := (s_1^N(\lambda), \ldots, s_N^N(\lambda)).$$

We will show that both the approximation error $\mathcal{E}(f_{\lambda^\varepsilon})$ and the "approximate sparsity" of λ^ε can be controlled in terms of the alignment coefficient of the vector $s_N(\lambda)$ for an arbitrary $\lambda \in \Lambda$. We will use the following version of the alignment coefficient:

$$\alpha_N^+(\lambda) := a_H^{(b)}(\Lambda, \lambda, s^N(\lambda)) \vee 0,$$

where

$$b := b(\lambda) := 2\|s_N(\lambda)\|_{\ell_\infty}.$$

Theorem 9.4 *There exists a constant $C > 0$ that depends only on ℓ and on the constant M such that $\|f_*\|_\infty \le M$ with the following property. For all $\varepsilon > 0$ and all $\lambda \in \Lambda$,*

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin \mathrm{supp}(\lambda)} \lambda_j^\varepsilon \le 3\mathcal{E}(f_\lambda) + C\biggl((\alpha_N^+(\lambda))^2\varepsilon^2 + \frac{\varepsilon}{N}\biggr). \tag{9.31}$$

Proof of Theorem 9.4. The definition of λ^ε implies that, for all $\lambda \in \Lambda$,

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j=1}^N \lambda_j^\varepsilon \log(N^2\lambda_j^\varepsilon) \le \mathcal{E}(f_\lambda) + \varepsilon \sum_{j=1}^N \lambda_j \log(N^2\lambda_j)$$

By convexity of the function $u \mapsto u\log(N^2 u)$ and the fact that its derivative is $\log(eN^2u)$,

$$\begin{aligned} &\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \log(N^2\lambda_j^\varepsilon) \le \\ &\mathcal{E}(f_\lambda) + \varepsilon \sum_{j \in J_\lambda} \Bigl(\lambda_j \log(N^2\lambda_j) - \lambda_j^\varepsilon \log(N^2\lambda_j^\varepsilon)\Bigr) \le \\ &\mathcal{E}(f_\lambda) + \varepsilon \sum_{j \in J_\lambda} \log(eN^2\lambda_j)(\lambda_j - \lambda_j^\varepsilon). \end{aligned} \tag{9.32}$$

Note that

$$\varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon = \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \log(N^2\lambda_j^\varepsilon) +$$

$$\varepsilon \sum_{j \notin J_\lambda, \lambda_j^\varepsilon \le eN^{-2}} \lambda_j^\varepsilon \Bigl(1 - \log(N^2\lambda_j^\varepsilon)\Bigr) + \varepsilon \sum_{j \notin J_\lambda, \lambda_j^\varepsilon > eN^{-2}} \lambda_j^\varepsilon \Bigl(1 - \log(N^2\lambda_j^\varepsilon)\Bigr).$$

We have

$$\varepsilon \sum_{j \notin J_\lambda, \lambda_j^\varepsilon > eN^{-2}} \lambda_j^\varepsilon \Bigl(1 - \log(N^2\lambda_j^\varepsilon)\Bigr) \le 0$$

and

$$\varepsilon \sum_{j \notin J_\lambda, \lambda_j^\varepsilon \le eN^{-2}} \lambda_j^\varepsilon \Bigl(1 - \log(N^2\lambda_j^\varepsilon)\Bigr) \le \varepsilon \sum_{j \notin J_\lambda, \lambda_j^\varepsilon \le eN^{-2}} \lambda_j^\varepsilon \le \varepsilon e N^{-1}.$$

Therefore,

$$\varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \le \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \log(N^2\lambda_j^\varepsilon) + \varepsilon e N^{-1}.$$

Recalling (9.32), we get

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \le \mathcal{E}(f_\lambda) + \varepsilon \sum_{j \in J_\lambda} \log(eN^2\lambda_j)(\lambda_j - \lambda_j^\varepsilon) + \varepsilon e N^{-1}.$$

If

$$\mathcal{E}(f_\lambda) + \varepsilon e N^{-1} \ge \varepsilon \sum_{j \in J_\lambda} \log(eN^2\lambda_j)(\lambda_j - \lambda_j^\varepsilon),$$

then

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \le 2\mathcal{E}(f_\lambda) + 2\varepsilon e N^{-1},$$

and the bound of the theorem follows. Otherwise, we have

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \le 2\varepsilon \sum_{j \in J_\lambda} \log(eN^2\lambda_j)(\lambda_j - \lambda_j^\varepsilon),$$

which, in particular, implies that

$$\sum_{j \notin J_\lambda} \lambda_j^\varepsilon \le 2\|s_N(\lambda)\|_{\ell_\infty} \sum_{j \in J_\lambda} |\lambda_j - \lambda_j^\varepsilon|.$$

This means that $\lambda - \lambda^\varepsilon \in C_{b,\lambda}$. The definition of $\alpha_N^+(\lambda)$ implies in this case that

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \le 2\varepsilon \sum_{j \in J_\lambda} \log(eN^2\lambda_j)(\lambda_j - \lambda_j^\varepsilon) \le 2\varepsilon \alpha_N^+(\lambda) \|f_\lambda - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}.$$

Since ℓ is the loss of quadratic type, we have

$$\|f_\lambda - f_{\lambda^\varepsilon}\|_{L_2(\Pi)} \le \|f_\lambda - f_*\|_{L_2(\Pi)} + \|f_{\lambda^\varepsilon} - f_*\|_{L_2(\Pi)} \le \sqrt{\frac{\mathcal{E}(f_\lambda)}{\tau}} + \sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau}}$$

(see (9.28)). This yields

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \le \mathcal{E}(f_\lambda) + 2\varepsilon \alpha_N^+(\lambda) \biggl(\sqrt{\frac{\mathcal{E}(f_\lambda)}{\tau}} + \sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau}} \biggr).$$

Using the fact that

$$2\varepsilon \alpha_N^+(\lambda) \sqrt{\frac{\mathcal{E}(f_{\lambda^\varepsilon})}{\tau}} \le 2\frac{(\alpha_N^+(\lambda))^2 \varepsilon^2}{\tau} + \frac{1}{2}\mathcal{E}(f_{\lambda^\varepsilon})$$

and

$$2\varepsilon \alpha_N^+(\lambda) \sqrt{\frac{\mathcal{E}(f_\lambda)}{\tau}} \le 2\frac{(\alpha_N^+(\lambda))^2 \varepsilon^2}{\tau} + \frac{1}{2}\mathcal{E}(f_\lambda),$$

we get

$$\frac{1}{2}\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin J_\lambda} \lambda_j^\varepsilon \leq \frac{3}{2}\mathcal{E}(f_\lambda) + 2\frac{(\alpha_N^+(\lambda))^2\varepsilon^2}{\tau},$$

which completes the proof.

□

Theorem 9.4 and random error bounds (9.29), (9.30) imply oracle inequalities for the excess risk $\mathcal{E}(f_{\hat{\lambda}^\varepsilon})$. The next corollary is based on (9.30).

Corollary 9.1 *Under the conditions of Theorem 9.2, for all* $\lambda \in \Lambda$ *with* $J = \mathrm{supp}(\lambda)$ *and for all subspaces* L *of* $L_2(\Pi)$ *with* $d := \dim(L)$, *the following bound holds with probability at least* $1 - N^{-A}$ *and with a constant* C *depending on* ℓ *and on* M :

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) \leq 4\mathcal{E}(f_\lambda) + C\biggl(\frac{d + A\log N}{n} + \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} + \frac{U(L)\log N}{n} + (\alpha_N^+(\lambda))^2\varepsilon^2 + \frac{\varepsilon}{N}\biggr).$$

9.3 Density Estimation and Sparse Mixtures Recovery

Let $X_1, \dots, X_n$ be i.i.d. observations in S with common distribution P that has density f_* with respect to a σ-finite measure μ in $(S, \mathcal{A})$. Suppose f_* is uniformly bounded by a constant M and $h_1, \dots, h_N$ be a dictionary of probability densities with respect to μ uniformly bounded by 1 (if they are uniformly bounded by an arbitrary constant, the results hold with a proper change of constants in the theorems). The goal is to construct an estimator of the density f_* in the family of all mixtures $\{f_\lambda : \lambda \in \Lambda\}$ and in the case when the dictionary is large, but there exists a "sparse" mixture that provides a good approximation of the unknown density. We study an estimator based on minimizing the entropy penalized empirical risk with respect to quadratic loss:

$$\hat{\lambda}^\varepsilon := \mathrm{argmin}_{\lambda\in\Lambda}\biggl[\|f_\lambda\|^2_{L_2(\mu)} - 2P_n f_\lambda + \varepsilon\sum_{j=1}^N \lambda_j \log\lambda_j\biggr]. \tag{9.33}$$

We compare $\hat{\lambda}^\varepsilon$ with the solution of penalized true risk minimization problem:

$$\lambda^\varepsilon := \mathrm{argmin}_{\lambda\in\Lambda}\biggl[\|f_\lambda - f_*\|^2_{L_2(\mu)} - \varepsilon H(\lambda)\biggr] = \mathrm{argmin}_{\lambda\in\Lambda}\biggl[\|f_\lambda\|^2_{L_2(\mu)} - 2Pf_\lambda + \varepsilon\sum_{j=1}^N \lambda_j\log\lambda_j\biggr]. \tag{9.34}$$

Bunea, Tsybakov and Wegkamp [26] studied a density estimation problem with ℓ_1-penalized empirical risk with respect to quadratic loss (in the case of linear aggregation rather than convex aggregation).

The results are quite similar to what was done in the previous sections in the case of prediction problems. We formulate them without proofs.

Theorem 9.5 *There exist numerical constants $D > 0$ and $C > 0$ such that, for all $J \subset \{1, \dots, N\}$ with $d := d(J) = \mathrm{card}(J)$, for all $A \geq 1$ and for all*

$$\varepsilon \geq D\sqrt{\frac{d + A\log N}{n}},$$

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

$$\sum_{j \notin J} \hat{\lambda}_j^{\varepsilon} \leq C\biggl[\sum_{j \notin J} \lambda_j^{\varepsilon} + M^2\sqrt{\frac{d + A\log N}{n}}\biggr],$$

$$\sum_{j \notin J} \lambda_j^{\varepsilon} \leq C\biggl[\sum_{j \notin J} \hat{\lambda}_j^{\varepsilon} + M^2\sqrt{\frac{d + A\log N}{n}}\biggr]$$

and

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\mu)}^2 + \varepsilon K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon}) \leq C\biggl[M^2\frac{d + A\log N}{n} \bigvee \sum_{j \notin J} \lambda_j^{\varepsilon}\sqrt{\frac{d + A\log N}{n}}\biggr].$$

Theorem 9.6 *Suppose that*

$$\varepsilon \geq D\sqrt{\frac{A\log N}{n}}$$

with a large enough numerical constant $D > 0$. For all $J \subset \{1, \dots, N\}$, for all subspaces L of $L_2(P)$ with $d := \dim(L)$ and for all $A \geq 1$, the following bounds hold with probability at least $1 - N^{-A}$ and with a numerical constant $C > 0$:

$$\sum_{j \notin J} \hat{\lambda}_j^{\varepsilon} \leq C\biggl[\sum_{j \notin J} \lambda_j^{\varepsilon} + M^2\frac{d + A\log N}{n\varepsilon} + \max_{j \in J}\|P_{L^{\perp}}h_j\|_{L_2(P)} + \frac{U(L)\log N}{n\varepsilon}\biggr], \tag{9.35}$$

$$\sum_{j \notin J} \lambda_j^{\varepsilon} \leq C\biggl[\sum_{j \notin J} \hat{\lambda}_j^{\varepsilon} + M^2\frac{d + A\log N}{n\varepsilon} + \max_{j \in J}\|P_{L^{\perp}}h_j\|_{L_2(P)} + \frac{U(L)\log N}{n\varepsilon}\biggr] \tag{9.36}$$

and

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\mu)}^2 + \varepsilon K(\hat{\lambda}^{\varepsilon}, \lambda^{\varepsilon}) \leq C\biggl[M^2\frac{d + A\log N}{n} \bigvee \sum_{j \notin J} \lambda_j^{\varepsilon}\sqrt{\frac{A\log N}{n}} \bigvee$$

$$\max_{j \in J}\|P_{L^{\perp}}h_j\|_{L_2(P)}\sqrt{\frac{A\log N}{n}} \bigvee \frac{U(L)\log N}{n}\biggr]. \tag{9.37}$$

In the case of density estimation, the alignment coefficient should be defined in terms of measure μ :

$$a_H^{(b)}(\Lambda, \lambda, w) := \sup\Bigl\{\langle w, u\rangle_{\ell_2} : u \in -T_\Lambda(\lambda) \cap C_{b,\lambda}, \|f_u\|_{L_2(\mu)} = 1\Bigr\},$$

$$\alpha_N^+(\lambda) := a_H^{(b)}(\Lambda, \lambda, s^N(\lambda)) \vee 0$$

with

$$b := b(\lambda) := 2\|s_N(\lambda)\|_{\ell_\infty}.$$

The next two statements provide an approximation error bound and an oracle inequality on the risk $L_2(\mu)$-risk of the estimator.

Theorem 9.7 *There exists a numerical constant $C > 0$ such that, for all $\varepsilon > 0$ and all $\lambda \in \Lambda$*

$$\|f_{\lambda^\varepsilon} - f_*\|^2_{L_2(\mu)} + \varepsilon \sum_{j \notin \mathrm{supp}(\lambda)} \lambda_j^\varepsilon \le 3\|f_\lambda - f_*\|^2_{L_2(\mu)} + C\biggl(\varepsilon^2(\alpha_N^+(\lambda))^2 + \frac{\varepsilon}{N}\biggr). \tag{9.38}$$

Corollary 9.2 *Under the conditions of Theorem 9.6, for all $\lambda \in \Lambda$ with $J = \mathrm{supp}(\lambda)$ and for all subspaces L of $L_2(\Pi)$ with $d := \dim(L)$, the following bound holds with probability at least $1 - N^{-A}$ and with a numerical constant C :*

$$\|f_{\hat\lambda^\varepsilon} - f_*\|^2_{L_2(\mu)} \le 4\|f_\lambda - f_*\|^2_{L_2(\mu)} +$$

$$C\biggl(M^2\frac{d + A\log N}{n} + \max_{j\in J}\|P_{L^\perp}h_j\|_{L_2(\Pi)}\sqrt{\frac{A\log N}{n}} + \frac{U(L)\log N}{n} + \varepsilon^2(\alpha_N^+(\lambda))^2 + \frac{\varepsilon}{N}\biggr).$$

9.4 ℓ_p-Penalization in Sparse Recovery

In this section, we will briefly discuss another example of strictly convex penalization, penalization with the p-th power of ℓ_p-norm. Specifically, we will study the following penalized empirical risk minimization problem

$$\hat\lambda^\varepsilon := \mathrm{argmin}_{\lambda \in U_{\ell_p}}\Bigl[P_n(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|^p_{\ell_p}\Bigr] \tag{9.39}$$

that will be compared with its true version

$$\lambda^\varepsilon := \mathrm{argmin}_{\lambda \in U_{\ell_p}}\Bigl[P(\ell \bullet f_\lambda) + \varepsilon\|\lambda\|^p_{\ell_p}\Bigr]. \tag{9.40}$$

Here $p \in [1,2]$ and we will denote

$$q := \frac{p}{p-1}.$$

We will be also using the notations of the previous sections such as, for instance, the quantity $U(L)$.

The following theorems provide control of the size of random $L_2(\Pi)$-error $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}$ in terms of sparsity or approximate sparsity of λ^{ε}. They also show that approximate sparsity of λ^{ε} implies that $\hat{\lambda}^{\varepsilon}$ possesses a similar property. We will start with the case when the vector λ^{ε} is sparse, i.e., there exists a small set $J \subset \{1, \dots, N\}$ such that $\lambda_j^{\varepsilon} = 0, j \notin J$.

Theorem 9.8 *There exist constants $D_0 > 0$ and $C > 0$ depending only on ℓ with the following property. Suppose that $J \subset \{1, \dots, N\}$ with $d := d(J) = \mathrm{card}(J)$ and*

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

For all $A \geq 1$, for all $D \geq D_0$ and for

$$\varepsilon = DN^{1/q}\biggl(\sqrt{\frac{d + A\log N}{n}} + \sqrt{\frac{q-1}{n}}\biggr), \tag{9.41}$$

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

$$\sum_{j \notin J} |\hat{\lambda}_j^{\varepsilon}|^p \leq C\biggl[\frac{d + A\log N}{n\varepsilon} \bigvee \frac{1}{D^q}\biggr]$$

and

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}^2 \leq C\biggl[\frac{d + A\log N}{n} \bigvee \frac{(p-1)\varepsilon}{D^q}\biggr].$$

Another version of the result allows one to use smaller values of regularization parameter ε than in condition (9.41).

Theorem 9.9 *There exist constants $D_0 > 0$ and $C > 0$ depending only on ℓ with the following property. Suppose that $J \subset \{1, \dots, N\}$ is such that*

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

Let L be a subspace of $L_2(\Pi)$ with $d := \dim(L) < +\infty$. For all $A \geq 1$, for all $D \geq D_0$ and for

$$\varepsilon = DN^{1/q}\sqrt{\frac{q-1}{n}}, \tag{9.42}$$

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

$$\sum_{j \notin J} |\hat{\lambda}_j^{\varepsilon}|^p \leq C \biggl[\frac{d + A \log N}{n\varepsilon} \bigvee \max_{j \in J} \|P_{L^{\perp}} h_j\|_{L_2(\Pi)} \bigvee U(L) \frac{N^{2/q}(q-1)}{n\varepsilon} \bigvee \frac{1}{D^q} \biggr]$$

and

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}^2 \leq C \biggl[\frac{d + A \log N}{n} \bigvee \max_{j \in J} \|P_{L^{\perp}} h_j\|_{L_2(\Pi)} \frac{N^{1/q}(q-1)^{1/2}}{n^{1/2}} \bigvee$$

$$U(L) \frac{N^{2/q}(q-1)}{n} \bigvee \frac{(p-1)\varepsilon}{D^q} \biggr].$$

Note that if $p - 1 \asymp \frac{1}{\log N}$ (this was the case of p close to 1 studied in [65]), then

$$N^{1/q} \sqrt{\frac{q-1}{n}} \asymp \sqrt{\frac{\log N}{n}}$$

and the error terms in the bounds of the theorems start resembling the error terms in the case of ℓ_1-penalization (see Section 8.2). Also, in the case of p close to 1 the term $\frac{1}{D^q}$ becomes of the order N^{-B} for some $B > 0$, so, it is small. The size of this term can be also controlled by the choice of D (which could be an arbitrary number larger than D_0 for some constant D_0 depending only on ℓ).

We turn now to the case when λ^{ε} is approximately sparse.

Theorem 9.10 *There exist constants* $D_0 > 0$ *and* $C > 0$ *depending only on* ℓ *such that, for all* $J \subset \{1, \ldots, N\}$ *with* $d := d(J) = \mathrm{card}(J)$, *for all* $A \geq 1$, *for all* $D \geq D_0$ *and for*

$$\varepsilon = D(q-1)N^{1/q} \biggl(\sqrt{\frac{d + A \log N}{n}} + \sqrt{\frac{q-1}{n}} \biggr) \tag{9.43}$$

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

$$\sum_{j \notin J} |\hat{\lambda}_j^{\varepsilon}|^p \leq C \biggl[\sum_{j \notin J} |\lambda_j^{\varepsilon}|^p \bigvee (q-1) \frac{d + A \log N}{n\varepsilon} \bigvee \frac{1}{D^q} \biggr],$$

$$\sum_{j \notin J} |\lambda_j^{\varepsilon}|^p \leq C \biggl[\sum_{j \notin J} |\hat{\lambda}_j^{\varepsilon}|^p \bigvee (q-1) \frac{d + A \log N}{n\varepsilon} \bigvee \frac{1}{D^q} \biggr]$$

and

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda^{\varepsilon}}\|_{L_2(\Pi)}^2 \leq C \biggl[\frac{d + A \log N}{n} \bigvee \biggl(\sum_{j \notin J} |\lambda_j^{\varepsilon}|^p \biggr)^{1/p} \biggl(\sqrt{\frac{d + A \log N}{n}}$$

$$\bigvee \frac{N^{1/q}(q-1)^{1/2}}{n^{1/2}} \biggr) \bigvee \frac{(p-1)\varepsilon}{D^q} \biggr].$$

Similarly to Theorem 9.9, we will also establish another version of these bounds that hold for smaller values of ε.

Theorem 9.11 *There exist constants $D_0 > 0$ and $C > 0$ depending only on ℓ with the following property. For all $J \subset \{1, \dots, N\}$, for all subspaces $L \subset L_2(\Pi)$ with $d := \dim(L)$, for all $A \geq 1$, for all $D \geq D_0$ and for*

$$\varepsilon = D(q-1)N^{1/q}\sqrt{\frac{q-1}{n}} \tag{9.44}$$

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

$$\sum_{j \notin J} |\hat{\lambda}_j^{\varepsilon}|^p \leq$$

$$C\Biggl[\sum_{j \notin J} |\lambda_j^{\varepsilon}|^p \bigvee (q-1)\frac{d + A\log N}{n\varepsilon} \bigvee \max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \bigvee U(L)\frac{N^{2/q}(q-1)}{n\varepsilon} \bigvee \frac{1}{D^q}\Biggr],$$

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

$$C\Biggl[\sum_{j \notin J} |\hat{\lambda}_j^{\varepsilon}|^p \bigvee (q-1)\frac{d + A\log N}{n\varepsilon} \bigvee \max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \bigvee U(L)\frac{N^{2/q}(q-1)}{n\varepsilon} \bigvee \frac{1}{D^q}\Biggr],$$

and

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}^2 \leq C\Biggl[\frac{d + A\log N}{n} \bigvee \Bigl(\sum_{j \notin J} |\lambda_j^{\varepsilon}|^p\Bigr)^{1/p} \frac{N^{1/q}(q-1)^{1/2}}{n^{1/2}} \bigvee$$
$$\max_{j \in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \frac{N^{1/q}(q-1)^{1/2}}{n^{1/2}} \bigvee U(L)\frac{N^{2/q}(q-1)}{n} \bigvee \frac{(p-1)\varepsilon}{D^q}\Biggr].$$

We are not going to give the proofs of these results. They are based on general approach outlined in Section 8.1 and the details of the arguments are close to the proofs of theorems 8.5 and 9.2. Theorem 3.5 is being used to control the ℓ_q-norms of Rademacher processes indexed by finite classes which is needed in the proofs. It is worth mentioning that, in this case, inequality (8.4) takes the following form (for $\lambda = \lambda^\varepsilon$):

$$c\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|_{L_2(\Pi)}^2 + \varepsilon p \sum_{j=1}^{N} \Bigl(|\hat{\lambda}_j^{\varepsilon}|^{p-1}\mathrm{sign}(\hat{\lambda}_j^{\varepsilon}) - |\lambda_j^{\varepsilon}|^{p-1}\mathrm{sign}(\hat{\lambda}_j^{\varepsilon})\Bigr)(\hat{\lambda}_j^{\varepsilon} - \lambda_j^{\varepsilon})$$
$$\leq (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}). \tag{9.45}$$

In the case when λ^ε is sparse, i.e., there exists $J \subset \{1, \dots, N\}$ such that $\lambda_j^\varepsilon = 0,\ j \notin J$, this yields the following bound

$$c\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon p \sum_{j \notin J} |\hat{\lambda}^\varepsilon_j|^p \le (P - P_n)(\ell' \bullet f_{\hat{\lambda}^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}). \tag{9.46}$$

This provides a way to control the sparsity of the empirical solution $\hat{\lambda}^\varepsilon$ in terms of the empirical process in the right hand side which is used to complete the proofs of theorems 9.8 and 9.9.

In the case when the true solution λ^ε is only approximately sparse, the corresponding bounds become more complicated: for an arbitrary set J,

$$\frac{p \log 2}{4}(p-1)\varepsilon \sum_{j \notin J} |\hat{\lambda}^\varepsilon_j|^p \le \frac{p 2^p \log 2}{4}(p-1)\varepsilon \sum_{j \notin J} |\lambda^\varepsilon_j|^p + (P-P_n)(\ell' \bullet f_{\lambda^\varepsilon})(f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}) \tag{9.47}$$

and

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

However, it is still possible to use these inequalities and implement the program outlined in Section 8.1. This leads to theorems 9.10 and 9.11. Note that, in this case, there is an additional factor $p-1$ in front of the expression

$$\varepsilon \sum_{j \notin J} |\hat{\lambda}^\varepsilon_j|^p$$

characterizing the sparsity of $\hat{\lambda}^\varepsilon$. Essentially, it comes from the second derivative of the penalty function $\psi(u) = |u|^p$. In the case when p is close to 1 this factor is small and it leads to a large extra factor $q-1$ in the lower bound on ε in theorems 9.10 and 9.11. This is not needed in the sparse case of theorems 9.8 and 9.9.

Finally, we will discuss a version of approximation error bounds and oracle inequalities in the case of ℓ_p-penalization. This can be done by repeating the arguments of Section 9.2 for entropy penalization.

For $\lambda \in \mathbb{R}^N$, let

$$s_j(\lambda) := p|\lambda_j|^{p-1}\mathrm{sign}(\lambda_j),\ j = 1, \dots, N.$$

Clearly,

$$s_j(\lambda) = \psi'(\lambda_j),$$

where $\psi(u) = |u|^p$ is the penalty function. The vector

$$s(\lambda) := (s_1(\lambda), \dots, s_N(\lambda)) = \nabla \sum_{j=1}^{N} |\lambda_j|^p$$

is the gradient of the penalty.

Define the following version of the alignment coefficient:

$$\alpha_+(\lambda) := a_H^{(b)}(U_{\ell_p}, \lambda, s(\lambda)) \vee 0$$

with

$$b := b(\lambda) := 2\|s(\lambda)\|_{\ell_\infty}.$$

The next result shows that the approximation error $\mathcal{E}(f_{\lambda^\varepsilon})$ and the "approximate sparsity" of λ^ε can be controlled in terms of $\alpha_+(\lambda)$.

Theorem 9.12 *There exists a constant $C > 0$ that depends only on ℓ and on the constant M such that $\|f_*\|_\infty \le M$ with the following property. For all $\varepsilon > 0$ and all $\lambda \in U_{\ell_1}$,*

$$\mathcal{E}(f_{\lambda^\varepsilon}) + \varepsilon \sum_{j \notin \mathrm{supp}(\lambda)} |\lambda_j^\varepsilon|^p \le 3\mathcal{E}(f_\lambda) + C(\alpha_+(\lambda))^2\varepsilon^2. \tag{9.48}$$

Together with random error bounds, Theorem 9.12 easily implies oracle inequalities. For instance, the next result follows from Theorem 9.11.

Corollary 9.3 *Under the conditions of Theorem 9.11, for all $\lambda \in U_{\ell_1}$ with $J = \mathrm{supp}(\lambda)$ and for all subspaces L of $L_2(\Pi)$ with $d := \dim(L)$, the following bound holds with probability at least $1 - N^{-A}$ and with a constant C depending on ℓ and on M :*

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) \le 4\mathcal{E}(f_\lambda) + C\biggl(\frac{d + A\log N}{n} \bigvee \max_{j\in J} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \frac{N^{1/q}(q-1)^{1/2}}{n^{1/2}} \bigvee$$

$$U(L)\frac{N^{2/q}(q-1)}{n} \bigvee \frac{(p-1)\varepsilon}{D^q} \bigvee (\alpha_+(\lambda))^2\varepsilon^2\biggr).$$

10 Appendix: Properties of $\sharp$- and $\flat$-Transforms

In this appendix, we provide some properties of $\sharp$- and $\flat$-transforms introduced in Section 4.1 and used in the construction of excess risk bounds. The proofs of these properties are rather elementary. We are mainly interested in $\sharp$-transform.

1. If $\psi(u) = o(u)$ as $u \to \infty$, then the function $\psi^\sharp$ is defined on $(0, +\infty)$ and is a nonincreasing function on this interval.

2. If $\psi_1 \leq \psi_2$, then $\psi_1^\sharp \leq \psi_2^\sharp$. Moreover, if $\psi_1(\delta) \leq \psi_2(\delta)$ either for all $\delta \geq \psi_2^\sharp(\varepsilon)$, or for all $\delta \geq \psi_1^\sharp(\varepsilon) - \tau$ with an arbitrary $\tau > 0$, then also $\psi_1^\sharp(\varepsilon) \leq \psi_2^\sharp(\varepsilon)$.

3. For all $a > 0$,

$$(a\psi)^\sharp(\varepsilon) = \psi^\sharp(\varepsilon/a).$$

4. If $\varepsilon = \varepsilon_1 + \cdots + \varepsilon_m$, then

$$\psi_1^\sharp(\varepsilon) \bigvee \cdots \bigvee \psi_m^\sharp(\varepsilon) \leq (\psi_1 + \cdots + \psi_m)^\sharp(\varepsilon) \leq \psi_1^\sharp(\varepsilon_1) \bigvee \cdots \bigvee \psi_m^\sharp(\varepsilon_m).$$

5. If $\psi(u) \equiv c$, then

$$\psi^\sharp(\varepsilon) = c/\varepsilon.$$

6. If $\psi(u) := u^\alpha$ with $\alpha \leq 1$, then

$$\psi^\sharp(\varepsilon) := \varepsilon^{-1/(1-\alpha)}.$$

7. For $c > 0$, denote $\psi_c(\delta) := \psi(c\delta)$. Then

$$\psi_c^\sharp(\varepsilon) = \frac{1}{c}\psi^\sharp(\varepsilon/c).$$

If ψ is nondecreasing and $c \geq 1$, then

$$c\psi^\sharp(u) \leq \psi^\sharp(u/c).$$

8. For $c > 0$, denote $\psi_c(\delta) := \psi(\delta + c)$. Then for all $u > 0, \varepsilon \in (0, 1]$

$$\psi_c^\sharp(u) \leq \psi^\sharp(\varepsilon u/2) - c \vee c\varepsilon.$$

Recall the definitions of functions of concave type and strictly concave type from Section 4.1.

9. If ψ is of concave type, then $\psi^\sharp$ is the inverse of the function

$$\delta \mapsto \frac{\psi(\delta)}{\delta}.$$

In this case,

$$\psi^{\sharp}(cu) \geq \psi^{\sharp}(u)/c$$

for $c \leq 1$ and

$$\psi^{\sharp}(cu) \leq \psi^{\sharp}(u)/c$$

for $c \geq 1$.

10. If ψ is of strictly concave type with exponent γ, then for $c \leq 1$

$$\psi^{\sharp}(cu) \leq \psi^{\sharp}(u)c^{-\frac{1}{1-\gamma}}.$$

References

[1] Affentranger, F. and Schneider, R. (1992) Random Projections of Regular Simplices. *Discrete Comput. Geom.*, 7(3), 219–226.

[2] Alexander, K.S. (1987) Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields,* 75, 379–423.

[3] Anthony, M. and Bartlett, P. (1999) Neural Network Learning: Theoretical Foundations. Cambridge University Press.

[4] Audibert, J.-Y. (2004) Une approche PAC-bayésienne de la théorie statistique de l'apprentissage. PhD Thesis, University of Paris 6.

[5] Barron, A., Birgé, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields*, 113, 301–413.

[6] Bartlett, P. (2008) Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory,* 24(2), to appear.

[7] Bartlett, P., Bousquet, O. and Mendelson, S. (2005) Local Rademacher Complexities, *Annals of Statistics,* 33, 4, 1497-1537.

[8] Bartlett, P., Boucheron, S. and Lugosi, G. (2002) Model selection and error estimation. *Machine Learning,* 48, 85–113.

[9] Bartlett, P. and Mendelson, S. (2006) Empirical Risk Minimization. *Probability Theory and Related Fields,* 135(3), 311–334.

[10] Bartlett, P., Jordan, M. and McAuliffe, J. (2006) Convexity, Classification and Risk Bounds. *Journal of American Statistical Association,* 101(473), 138–156.

[11] Bartlett, P. and Williamson, R.C. (1996) Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6), 2118–2132.

[12] Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM: Probability and Statistics,* 6, 127–146.

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

[14] Bickel, P., Ritov, Y. and Tsybakov, A. (2007) Simultaneous Analysis of LASSO and Dantzig Selector. Preprint.

[15] Birgé, L. and Massart, P. (1997) From Model Selection to Adaptive Estimation. In: Festschrift for L. Le Cam. Research Papers in Probability and Statistics. D. Pollard, E. Torgersen and G. Yang (Eds.), 55-87. Springer, New York.

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

[17] Blanchard, G., Lugosi, G. and Vayatis, N. (2003) On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research,* 4, 861–894.

[18] Bobkov, S. and Houdré, C. (1997) Isoperimetric constants for product probability measures, *Ann. Probab.,* v. 25, no. 1, 184–205.

[19] Boucheron, S., Bousquet, O. and Lugosi, G. (2005) Theory of Classification. A Survey of Some Recent Advances. *ESAIM Probability & Statistics,* 9, 323–371.

[20] Boucheron, S., Lugosi, G. and Massart, P. (2000) A sharp concentration inequality with applications. *Random Structures and Algorithms,* 16, 277–292.

[21] Boucheron, S., Bousquet, O., Lugosi, G. and Massart, P. (2005) Moment inequalities for functions of independent random variables. *Annals of Probability,* 33, 2, 514–560.

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

[23] Bousquet, O., Koltchinskii, V. and Panchenko, D. (2002) Some local measures of complexity of convex hulls and generalization bounds. In: COLT2002, Lecture Notes in Artificial Intelligence, 2375, Springer, 59 - 73.

[24] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Aggregation for Gaussian regression, *Annals of Statistics,* 35,4, 1674–1697.

[25] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the LASSO. *Electronic Journal of Statistics,* 1, 169–194.

[26] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparse density estimation with ℓ_1 penalties. In: *Proc. 20th Annual Conference on Learning Theory (COLT 2007),* Lecture Notes in Artificial Intelligence, Springer, v. 4539, pp. 530–543.

[27] Candes, E., Romberg, J. and Tao, T. (2006) Stable Signal Recovery from Incomplete and Inaccurate Measurements. *Communications on Pure and Applied Mathematics,* 59, 1207–1223.

[28] Candes, E., Rudelson, M., Tao, T. and Vershynin, R. (2005) Error Correction via Linear Programming. *Proc. 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS05),* IEEE, 295–308.

[29] Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics,* 35,6, 2313–2351.

[30] Candes, E. and Plan, Y. (2007) Near-ideal model selection by ℓ_1-minimization. Preprint.

[31] Catoni, O. (2004) Statistical Learning Theory and Stochastic Optimization. *Ecole d'Eté de Probabilités de Saint-Flour XXXI -2001,* Lecture Notes in Mathematics, **1851**, Springer, New York.

[32] de la Pena, V. and Giné, E. (1998) Decoupling: From Dependence to Independence, Springer, New York.

[33] Devroye, L., Györfi, G. and Lugosi, G. (1996) A Probabilistic Theory of Pattern Recognition. Springer, New York.

[34] Dalalyan, A. and Tsybakov, A. (2007) Aggregation by exponential weighting and sharp oracle inequalities. In: *Proc. 20th Annual Conference on Learning Theory (COLT 2007),* Lecture Notes in Artificial Intelligence, Springer, v. 4539, pp. 97–111.

[35] Donoho, D.L. (2004) For Most Large Underdetermined Systems of Equations the Minimal ℓ^1-norm Near-Solution Approximates the Sparsest Near-Solution. Preprint.

[36] Donoho, D.L. (2006) For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ^1-norm Solution is also the Sparsest Solution. *Communications on Pure and Applied Mathematics,* 59, 797–829.

[37] Donoho, D.L. (2006) Neighborly Polytopes and Sparse Solution of Underdetermined Linear Equations. *IEEE Transactions on Information Theory,* to appear.

[38] Donoho, D.L. and Tanner, J. (2005) Neighborliness of Randomly-Projected Simplices in High Dimensions. *Proc. National Academy of Sciences,* 102(27), 9446–9451.

[39] Donoho, D.L. (2006) Compressed Sensing. *IEEE Transactions on Information Theory,* 52, 4, 1289–1306.

[40] Donoho, D.L., Elad, M. and Temlyakov, V. (2006) Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. *IEEE Transactions on Information Theory,* 52, 1, 6–18.

[41] Dudley, R.M. (1978) Central Limit Theorems for Empirical Measures, *Annals of Probability,* 6,6, 899–929.

[42] Dudley, R.M. (1999) Uniform Central Limit Theorems. Cambridge University Press.

[43] Einmahl, U. and Mason, D. (2000) An empirical processes approach to the uniform consistency of kernel type function estimators. *J. Theoretical Probability,* 13, 1–37.

[44] Fromont, M. (2007) Model selection by bootstrap penalization for classification. *Machine Learning,* 66, 2-3, 165–207.

[45] van de Geer, S. (1999) Empirical Processes in M-estimation. Cambridge University Press. Cambridge.

[46] van de Geer, S. (2008) High-dimensional generalized linear models and the Lasso, *Annals of Statistics,* 36, 2, 614–645.

[47] Giné, E. and Zinn, J. (1984) Some limit theorems for empirical processes. *Annals of Probability* 12, 929-989.

[48] Giné, E. and Guillou, A. (2001) On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. I.H.Poincaré,* 4, 503-522.

[49] Giné, E., Koltchinskii, V. and Wellner, J. (2003) Ratio Limit Theorems for Empirical Processes. In: *Stochastic Inequalities,* Birkhäuser, pp. 249–278.

[50] Giné, E. and Koltchinskii, V. (2006) Concentration Inequalities and Asymptotic Results for Ratio Type Empirical Processes. *Annals of Probability,* 34, 3, 1143-1216.

[51] Giné, E. and Nickl, R. (2008) Adaptive Estimation of Distribution Function and its Density in Sup-Norm Loss by Wavelet and Spline Projection. Preprint.

[52] Johnstone, I.M. (1998) Oracle Inequalities and Nonparametric Function Estimation. In: Documenta Mathematica, *Journal der Deutschen Mathematiker Vereinigung,* Proc. of the International Congress of Mathematicians, Berlin, 1998, v.III, 267–278.

[53] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002) A Distribution-Free Theory of Nonparametric Regression. Springer.

[54] Klein, T. (2002) Une inégalité de concentration à gauche pour les processus empiriques. *C.R. Acad. Sci. Paris,* Ser I, 334, 500–505.

[55] Klein, T. and Rio, E. (2005) Concentration around the mean for maxima of empirical processes. *Annals of Probability,* 33,3, 1060–1077.

[56] Kohler, M. (2000) Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. of Statistical Planning and Inference*, 89, 1–23.

[57] Koltchinskii, V. (1981) On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics*, 24, 71–82.

[58] Koltchinskii, V. (2001) Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory,* 47(5), 1902–1914.

[59] Koltchinskii, V. (2006) Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Annals of Statistics,* 34, 6, 2593–2656.

[60] Koltchinskii, V. and Panchenko, D. (2000) Rademacher processes and bounding the risk of function learning. In: Giné, E., Mason, D. and Wellner, J. *High Dimensional Probability II,* 443–459.

[61] Koltchinskii, V. and Panchenko, D. (2002) Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics,* 30, 1, 1–50.

[62] Koltchinskii, V., Panchenko, D. and Lozano, F. (2003) Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins. *Ann. Appl. Probab.,* **13**, 1, 213–252.

[63] Koltchinskii, V. and Panchenko, D. (2005) Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics,* 33, 4, 1455–1496.

[64] Koltchinskii, V. (2005) Model selection and aggregation in sparse classification problems. *Oberwolfach Reports: Meeting on Statistical and Probabilistic Methods of Model Selection, October 2005.*

[65] Koltchinskii, V. (2007) Sparsity in penalized empirical risk minimization, *Annales Inst. H. Poincaré, Probabilités et Statistiques,* to appear.

[66] Koltchinskii, V. (2007) The Dantzig Selector and Sparsity Oracle Inequalities. Preprint.

[67] Koltchinskii, V. (2008) Sparse Recovery in Convex Hulls via Entropy Penalization, *Annals of Statistics,* to appear.

[68] Ledoux, M. and Talagrand, M. (1991) Probability in Banach Spaces. Springer-Verlag, New York.

[69] Lugosi, G. and Vayatis, N. (2004) On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics,* 32, 1, 30–55.

[70] Lugosi, G. and Wegkamp, M. (2004) Complexity regularization via localized random penalties. *Annals of Statistics* , 32, 4, 1679–1697.

[71] Massart, P. and Nedelec, E. (2006) Risk bounds for statistical learning. *Annals of Statistics,* 34, 5, 2326–2366.

[72] Mammen, E. and Tsybakov, A. (1999) Smooth discrimination analysis. *Annals of Statistics* 27, 1808–1829.

[73] Massart, P. (2000) Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse (IX),* 245–303.

[74] Massart, P. (2007) Concentration Inequalities and Model Selection. *Ecole d'ete de Probabilités de Saint-Flour 2003,* Lecture Notes in Mathematics, Springer.

[75] McAllester, D.A. (1998) Some PAC-Bayesian theorems. In *Proc. 11th Annual Conference on Learning Theory,* pp. 230–234, ACM Press.

[76] Mendelson, S. (2002) Improving the sample complexity using global data. *IEEE Transactions on Information Theory,* 48, 1977–1991.

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

[78] Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2007) Reconstruction and subgaussian operators in Asymptotic Geometric Analysis. *Geometric and Functional Analysis,* 17(4), 1248–1282.

[79] Nemirovski, A. (2000) Topics in non-parametric statistics, In: P. Bernard, editor, *Ecole d'Et'e de Probabilités de Saint-Flour XXVIII,* 1998, Lecture Notes in Mathematics, Springer, New York.

[80] Pollard, D. (1982) A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society*, A, 33, 235–248.

[81] Pollard, D.(1984) Convergence of Stochastic Processes, Springer, New York.

[82] Rudelson, M. and Vershynin, R. (2005) Geometric Approach to Error Correcting Codes and Reconstruction of Signals. *Int. Math. Res. Not.,* no 64, 4019-4041.

[83] Shen, X. and Wong, W.H. (1994) Convergence rate of sieve estimates. *Annals of Statistics,* 22, 580-615.

[84] Steinwart, I. (2005) Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory,* 51, 128–142.

[85] Talagrand, M. (1994) Sharper bounds for Gaussian and empirical processes. *Annals of Probability,* 22, 28-76.

[86] Talagrand, M. (1996) A new look at independence. *Annals of Probability,* 24, 1-34.

[87] Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.,* 126, 505-563.

[88] Talagrand, M. (2005) The Generic Chaining. Springer.

[89] Tibshirani, R. (1996) Regression shrinkage and selection via Lasso, *J. Royal Statist. Soc., Ser B,* **58**, 267–288.

[90] Tsybakov, A. (2003) Optimal rates of aggregation. In: *Proc. 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines,* Lecture Notes in Artificial Intelligence, **2777**, Springer, New York, 303–313.

[91] Tsybakov, A. (2004) Optimal aggregation of classifiers in statistical learning. *Annals of Statistics* , 32, 135–166.

[92] Tsybakov, A. and van de Geer, S. (2005) Square root penalty: adaptation to the margin in classification and in the edge estimation. *Annals of Statistics,* 33, 3, 1203–1224.

[93] Vapnik, V. (1998) Statistical Learning Theory. John Wiley & Sons, New York.

[94] Vapnik, V. and Chervonenkis, A. (1974) Theory of Pattern Recognition. Nauka, Moscow (in Russian).

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

[96] Vershik, A.M. and Sporyshev, P.V. (1992) Asymptotic behavior of the number of faces of random polyhedra and the neighborliness problem. *Selecta Math. Soviet.*, 11(2), 181–201.

[97] Yang, Y. (2000) Mixing strategies for density estimation. *Annals of Statistics,* **28**, 75–87.

[98] Yang, Y. (2004) Aggregating regression procedures for a better performance. *Bernoulli,* **10**, 25–47.

[99] Zhang, T. (2001) Regularized Winnow Method. In: *Advances in Neural Information Processing Systems 13 (NIPS2000),* T.K. Leen, T.G. Dietrich and V. Tresp (Eds), MIT Press, pp. 703–709.

[100] Zhang, T. (2004) Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics,* 32, 1, 56–134.

[101] Zhang, T. (2006) From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. *Annals of Statistics,* 34, 2180–2210.

[102] Zhang, T. (2006) Information Theoretical Upper and Lower Bounds for Statistical Estimation. *IEEE Transactions on Information Theory,* 52, 1307-1321.