NON-ASYMPTOTIC BOUNDS FOR PREDICTION PROBLEMS AND DENSITY ESTIMATION

A Thesis
Presented to
The Academic Faculty

by

Stanislav Minsker

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology
August 2012

NON-ASYMPTOTIC BOUNDS FOR PREDICTION PROBLEMS AND DENSITY ESTIMATION

Approved by:

Professor Vladimir Koltchinskii,
Advisor
School of Mathematics
Georgia Institute of Technology

Professor Christian Houdré
School of Mathematics
Georgia Institute of Technology

Professor Yuri Bakhtin
School of Mathematics
Georgia Institute of Technology

Professor Justin Romberg
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Maria-Florina Balcan
School of Computer Science
Georgia Institute of Technology

Date Approved: May 1, 2012

To the memory of my Grandparents.

ACKNOWLEDGEMENTS

There are many people who helped me to choose the right direction in a long, sometimes random walk through the graduate school, which eventually led to completion of this thesis. I was lucky to have Dr. Vladimir Koltchinskii as my advisor, and I want to thank him for all support, guidance and the right amount of academic freedom that I had during the last 5 years. His ability to make complicated things look simple had always helped to recharge my depleted supplies of optimism. I am also grateful to Professors Yuri Bakhtin, Nina Balcan, Christian Houdré and Justin Romberg for serving on my dissertation committee.

Academic success would be impossible without the people I am proud to call my friends, and especially without Marta, Anton and Janetta. Thank you for sharing all the joy and difficulties of life with me. Fortunately, the complete list of names is too long to be included here, and I will express my gratitude in person.

I want to thank the faculty and staff of the School of Mathematics, and especially Dr. Luca Dieci and Ms. Cathy Jacobson, for their guidance along the way and for creating friendly atmosphere in the department.

Next, I want to mention the key role that professors of the Novosibirsk State University, especially Dr. Igor S. Borisov, played in building the foundations of my mathematical background and development of my career.

Most of all, I am grateful to my parents, Nikolay and Elena, for their unconditional support and for giving me everything I could hope for.

TABLE OF CONTENTS

LIST OF FIGURES

List of Algorithms

SUMMARY

This dissertation investigates the learning scenarios where a high-dimensional parameter has to be estimated from a given sample of fixed size, often smaller than the dimension of the problem. The motivation for the questions we study comes from real-world applications where the data is often obtained from expensive experiments and has to be processed and analyzed efficiently.

The first part answers some open questions for the binary classification problem in the framework of *active learning*. Given a random couple $(X, Y) \in \mathbb{R}^d \times \{\pm 1\}$ with unknown distribution P, the goal of binary classification is to predict a label Y based on the observation X. The prediction rule is constructed from the observations $(X_i, Y_i)_{i=1}^n$ sampled from P. The concept of active learning can be informally characterized as follows: on every iteration, the algorithm is allowed to request a label Y for *any* instance X which it considers to be the most informative. The contribution of this work consists of two parts: first, we provide the minimax lower bounds for the performance of active learning methods. Second, we propose an active learning algorithm which attains nearly optimal rates over a broad class of underlying distributions and is adaptive with respect to the unknown parameters of the problem.

The second part of this thesis is related to sparse recovery in the framework of *dictionary learning*. Let (X, Y) be a random couple with unknown distribution P, with X taking its values in some metric space S and Y - in a bounded subset of $\mathbb{R}$. Given a collection of functions $\mathcal{H} = \{h_t\}_{t \in \mathbb{T}}$ mapping S to $\mathbb{R}$, the goal of dictionary learning is to construct a prediction rule for Y given by a linear(or convex) combination of the elements of $\mathcal{H}$. The problem is *sparse* if there exists a good

prediction rule that depends on a small number of functions from $\mathcal{H}$. We propose an estimator of the unknown optimal prediction rule based on penalized empirical risk minimization algorithm. We show that the proposed estimator is able to take advantage of the possible sparse structure of the problem by providing probabilistic bounds for its performance.

CHAPTER I

INTRODUCTION

In the past decade, numerous applications created a high demand for the tools to process and analyze high-dimensional data. As a result, new algorithms and statistical models were introduced. One of the challenges associated with the theoretical analysis of these methods is to develop tight *non-asymptotic* bounds which give performance guarantees holding with high probability for the input of any fixed cardinality. In particular, such bounds are important when the dimension of the problem is much larger than the amount of data available to a researcher. In general, consistent estimation of all parameters of the system is impossible in this case, at least without additional assumptions, such as sparsity. Another challenge originates from the problems where one tries to minimize the amount of data used to achieve a certain goal (for example, when the data is obtained via expensive experiments). In this case, we want to use the available budget efficiently and to *choose* the most informative observations from the given collection.

This dissertation describes two problems that are general enough to include many nontrivial examples and at the same time admit tight non-asymptotic performance bounds. The first problem is related to sparse recovery in the framework of dictionary learning while the second targets some open questions in the field of *active learning*.

We will continue by presenting the key objects of statistical learning theory related to the questions we study, and by introducing the necessary background. Since the material of the following two chapters in not closely related, we are not going to create the connection artificially and will provide separate detailed introductions for the topics of interest in each chapter.

1.1 Statistical framework for prediction problems

One of the main goals of statistical learning theory is to develop good models and methods for making predictions and gaining knowledge from the data, which is assumed to come from some underlying (but usually unknown) distribution. Statistical framework happened to be very well-suited for this type of problems, since much of the real-world data is not exact but contaminated by random noise. To the best of our knowledge, one of the first successful attempts to place the learning problem into the statistical "Probably Approximately Correct"(PAC) framework is the pioneering work by L. Valiant [87]. Another breakthrough that underlies many deep results in the field was the work of M. Talagrand on general concentration inequalities in product spaces [82],[81]. Later, O. Bousquet, T. Klein and E. Rio [15], [47] obtained explicit small values of constants in Talagrand's inequality, thus expanding the range of its applications. These concentration results, coupled with some powerful tools from the empirical processes theory and geometric functional analysis, allow to prove many nontrivial results.

Let S be a measurable space, $T \subset \mathbb{R}$, and let (X, Y) be a random couple in $S \times T$ with unknown distribution P. The marginal distribution of X will be denoted by Π. Usually, X is the observation (for example, a newspaper article) and Y is the unknown quantity that has to be predicted based on X (for example, the category of the text: politics, sports, arts, etc.). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the so-called *training data* consisting of n copies of (X, Y) which are observed – in our example this could be a collection of articles taken from the search engine output and classified by a human. Often, the pairs are also assumed to be independent and identically distributed; however, as we will show later in this thesis, the result can sometimes be improved if the observations are dependent: intuitively, this happens when one is allowed to pick the most informative data from the available collection.

Our goal is to construct a good prediction rule – a measurable function $f : S \mapsto T$ –

based on the available training data. The quality of a prediction rule is measured in terms of the *loss function* $\ell(y, f(x))$, and the associated average loss is $\mathbb{E}\ell(Y, f(X))$ where the expectation is taken with respect to P. The popular losses include

1. the binary loss $\ell(y, f(x)) := I\{y \neq f(x)\}$, where $I\{A\}$ is the indicator of event A, commonly used when Y is a discrete random variable;

2. squared loss $\ell(y, f(x)) := (y - f(x))^2$, used for the regression problem;

3. exponential loss $\ell(y, f(x)) := e^{-yf(x)}$, used in binary classification (meaning that $Y \in \{\pm 1\}$) as a convex majorant of the discrete loss,

among others. In what follows, we will denote by P_n the empirical distribution based on a given sample of n training examples, $P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$. Similarly, Π_n will denote the empirical measure based on the sample $(X_1, \ldots, X_n)$. The integrals with respect to P and P_n will be denoted by

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

In many cases, it is desirable to find a prediction rule f which minimizes the average loss $P\ell(Y, f(X))$ over some class $\mathcal{F}$. However, this problem cannot be solved in practice since distribution P remains unknown. Instead, in the case of iid observations, the true average loss can be approximated by the empirical loss:

$$P\ell(Y, f(X)) \approx P_n\ell(Y, f(X)),$$

and the hope is that the minimizer of the empirical loss is going to be "close" to the minimizer of the true average loss, thus being a good candidate for a prediction rule. Much of the theory is devoted to understanding how this procedure can be formalized and how the performance of the obtained prediction rule depends on the richness of class $\mathcal{F}$ and the number of observations n.

There are a lot well-written reviews and monographs on learning and pattern recognition. Among many others, these are the excellent sources of information of different

level: the classical text by V. Vapnik and A. Chervonenkis [95] (german translation of the original russian text [92]), a more recent monograph by V. Vapnik [93], a book by L. Devroye, L. Györfi and G. Lugosi [30] and a set of lecture notes by V. Koltchinskii containing many recent results in the field [53].

1.2 Technical background and related results

Concentration inequalities play a crucial role in theoretical analysis of machine learning methods. The classical result of S. N. Bernstein is well-known (see [91], Lemma 2.2.9):

Theorem 1.2.1. *Let $X_1, \dots, X_n$ be a sequence of independent random variables with zero mean. Assume that $|X_i| \le M,\ i = 1 \dots n$ almost surely and let $B_n^2 := \sum_{i=1}^{n} \mathbb{E}X_i^2$. Then*

$$\Pr\left(\left|\sum_{i=1}^{n} X_i\right| \ge t\right) \le 2\exp\left(\frac{-t^2/2}{B_n^2 + Mt/3}\right).$$

When uniform bound on X_i's is not available (or is too large), a version of Bernstein's inequality for random variables with sub exponential tails might be useful (see [53], section A.2 and [91], Lemma 2.2.11).

Let $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing function with $\psi(0) = 0$.

Definition 1.1. *The Orlicz norm of a random variable η is defined via*

$$\|\eta\|_\psi := \inf\left\{C > 0 :\ \mathbb{E}\psi\left(\frac{|\eta|}{C}\right) \le 1\right\}$$

By $\|\cdot\|_{\psi_1},\ \|\cdot\|_{\psi_2}$ we denote the Orlicz norms for $\psi_1(x) := e^x - 1$ and $\psi_2(x) := e^{x^2} - 1$, respectively; the following inequalities are elementary:

$$\|\eta\|_{\psi_1} \le \sqrt{\log 2}\|\eta\|_{\psi_2}, \tag{1.2.1}$$

$$\|\eta^2\|_{\psi_1} = \|\eta\|_{\psi_2}^2. \tag{1.2.2}$$

Theorem 1.2.2. *Let $X_1, \dots, X_n$ be a sequence of independent random variables with zero mean. Assume that $\|X_1\|_{\psi_1} \leq V$. Then*

$$\Pr\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t}{V}, \frac{t^2}{nV^2}\right)\right).$$

Next, we formulate the uniform versions of Bernstein's inequality, along with several related results. Let $(S, \mathcal{B})$ be a measurable space and let $\mathcal{F}$ be a class of functions

$$\mathcal{F} \ni f : S \mapsto [-1, 1].$$

If Z is a random process indexed by $\mathcal{F}$, define $\|Z\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |Z(f)|$. Given a collection of independent random variables $X_1, \dots, X_n$, $X_i \in S$ with distribution Π, let

$$\sqrt{n}Z_n(f) := \sqrt{n}(\Pi_n - \Pi)(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}f(X_i))$$

be the empirical process indexed by $\mathcal{F}$. Finally, let

$$\sigma_\Pi^2(\mathcal{F}) := \sup_{f\in\mathcal{F}}\left(\mathbb{E}f(X)^2 - (\mathbb{E}f(X))^2\right).$$

We will always assume that $\sigma_\Pi(\mathcal{F}) < \infty$. The following inequalities provide an estimate for the deviations of $\|Z_n\|_{\mathcal{F}}$ from its mean. The bounds in its present form are taken from [53].

Theorem 1.2.3 (Bousquet [15])**.**

$$\Pr\left(\|Z_n\|_{\mathcal{F}} \geq \mathbb{E}\|Z_n\|_{\mathcal{F}} + \sqrt{2t\left(\sigma_\Pi^2(\mathcal{F}) + \frac{2}{\sqrt{n}}\mathbb{E}\|Z_n\|_{\mathcal{F}}\right)} + \frac{t}{3\sqrt{n}}\right) \leq e^{-t}.$$

Theorem 1.2.4 (Klein-Rio [47])**.**

$$\Pr\left(\|Z_n\|_{\mathcal{F}} \leq \mathbb{E}\|Z_n\|_{\mathcal{F}} - \sqrt{2t\left(\sigma_\Pi^2(\mathcal{F}) + \frac{2}{\sqrt{n}}\mathbb{E}\|Z_n\|_{\mathcal{F}}\right)} - \frac{t}{\sqrt{n}}\right) \leq e^{-t}.$$

We will also use a version of concentration inequality for the classes that do not admit uniform upper bound. Given a class $\mathcal{F} : S \mapsto \mathbb{R}$, let F be measurable and such that $|f(x)| \leq F(x)$ for all $f \in \mathcal{F}$ all x.

Theorem 1.2.5 (Adamczak [1])**.**

$$\Pr\left(\|Z_n\|_{\mathcal{F}} \geq K\left[\mathbb{E}\|Z_n\|_{\mathcal{F}} + \sigma_P(\mathcal{F})\sqrt{t} + \|\max_{1\leq i\leq n} F(X_i)\|_{\psi_1}\frac{t}{\sqrt{n}}\right]\right) \leq e^{-t},$$
$$\Pr\left(\mathbb{E}\|Z_n\|_{\mathcal{F}} \geq K\left[\|Z_n\|_{\mathcal{F}} + \sigma_P(\mathcal{F})\sqrt{t} + \|\max_{1\leq i\leq n} F(X_i)\|_{\psi_1}\frac{t}{\sqrt{n}}\right]\right) \leq e^{-t}.$$

The following results provide powerful tools to bound the expected supremum $\mathbb{E}\|Z_n\|_{\mathcal{F}}$ and together with aforementioned concentration inequalities yield the estimates for $\|Z_n\|_{\mathcal{F}}$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables (that is, $\Pr(\varepsilon_i = \pm 1) = \frac{1}{2}$) which are also independent from $X_1, \ldots, X_n$. Define the Rademacher process indexed by $\mathcal{F}$ as

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i).$$

Informally, $R_n(f)$ measures correlation between the 'random noise' $\{\varepsilon_i\}_{i=1}^n$ and the vector $\{f(X_i)\}_{i=1}^n$. It turns out that the expected supremum of empirical process indexed by class $\mathcal{F}$ can be controlled in terms of expected supremum of $R_n(f)$. The latter can be estimated by Dudley's entropy integral (or, more generally, by generic chaining complexity) associated to the index set $\mathcal{F}$ equipped with a (random) pseudo-metric $d_n^2(f,g) := \frac{1}{n}\sum\limits_{i=1}^{n}(f(X_i) - g(X_i))^2$. This is possible due to the fact that, conditionally on $\{X_i\}_{i=1}^n$, Rademacher process $R_n(f)$ is subgaussian with respect to $d_n(\cdot,\cdot)$:

Definition 1.2. *Let (T,d) be a pseudo-metric space. A random process $\{Y(t),\ t \in T\}$ is called subgaussian with respect to d if for any $t, s \in T$*

$$\mathbb{E}e^{\lambda(Y(t)-Y(s))} \leq e^{\lambda^2 d^2(t,s)/2},\ \lambda \in \mathbb{R}.$$

Theorem 1.2.6 (Symmetrization inequality)**.** *For any convex function $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$*

$$E\Phi\left(\frac{1}{2}\|R_n\|_{\mathcal{F}_c}\right) \leq \mathbb{E}\Phi\left(\|Z_n\|_{\mathcal{F}}\right) \leq \mathbb{E}\Phi\left(2\|R_n\|_{\mathcal{F}}\right),$$

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

Proof. See Theorem 2.1 in [53]. □

For $\Phi(x) = x$, the lower bound is usually combined with the inequality

$$\mathbb{E}\|R_n\|_{\mathcal{F}_c} \geq \|R_n\|_{\mathcal{F}} - \frac{\sup_{f\in\mathcal{F}} Pf}{\sqrt{n}}.$$

Assume $\phi : \mathbb{R} \mapsto \mathbb{R}$ is such that $\phi(0) = 0$ and $|\phi(u) - \phi(v)| \leq |u - v|,\ u, v \in \mathbb{R}$, and let $\phi \circ \mathcal{F} := \{\phi(f(\cdot)),\ f \in \mathcal{F}\}$. Then the expected supremum of a Rademacher process indexed by $\phi \circ \mathcal{F}$ can be controlled by $\mathbb{E}\|R_n\|_{\mathcal{F}}$:

Theorem 1.2.7 (Contraction inequality)**.**

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

Proof. See Theorem 4.12 in [62]. □

Finally, we present Dudley's entropy bound (for a proof and modern viewpoint, see [80]).

Definition 1.3 (Covering number)**.** *Let (T, d) be totally bounded. The covering number $N(T, d, \varepsilon)$ is the minimal number of balls of radius ε (with respect to d) needed to cover T.*

Note that the centers of the balls in the definition above do not have to be in T.

Theorem 1.2.8 (Dudley's entropy bound)**.** *Let $D(T)$ be the diameter of (T, d). If $Y(t)$ is subgaussian with respect to d, then for some absolute constant $C > 0$*

$$\mathbb{E}\sup_{t\in T} Y(t) \leq C \int_0^{D(T)} \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon.$$

1.2.1 Lower bounds for the minimax risk

Below, we will mention a general result that allows to obtain lower bounds for the risk in many statistical applications.

Theorem 1.2.9. *Let Σ be a class of models, $\mathrm{d} : \Sigma \times \Sigma \mapsto \mathbb{R}$ - the pseudo-metric and $\{P_f,\ f \in \Sigma\}$ – a collection of probability measures associated with Σ. Assume there exists a subset $\{f_0, \ldots, f_M\}$ of Σ such that*

1. $\mathrm{d}(f_i, f_j) \geq 2s > 0$ *for all* $0 \leq i < j \leq M$;
2. $P_{f_j} \ll P_{f_0}$ *for every* $1 \leq j \leq M$;
3. $\frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}(P_{f_j}, P_{f_0}) \leq \alpha \log M, \quad 0 < \alpha < \frac{1}{8}$.

Then

$$\inf_{\hat{f}} \sup_{f \in \Sigma} P_f \left(\mathrm{d}(\hat{f}, f) \geq s \right) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \right),$$

where the infimum is taken over all possible estimators of f based on a sample from P_f and $\mathrm{KL}(\cdot, \cdot)$ is the Kullback-Leibler divergence.

Proof. For a proof and examples of applications, see Theorem 2.5 in [85]. □

The result above is often combined with a combinatorial lemma known as Gilbert-Varshamov bound:

Proposition 1.2.10 (Gilbert-Varshamov)**.** *For $m \geq 8$, there exists*

$$\{\sigma_0, \ldots, \sigma_M\} \subset \{-1, 1\}^m$$

such that $\sigma_0 = \{1, 1, \ldots, 1\}$, $\rho(\sigma_i, \sigma_j) \geq \frac{m}{8}\ \forall\ 0 \leq i < k \leq M$ and $M \geq 2^{m/8}$ where ρ stands for the Hamming distance $\rho(\sigma, \nu) := \sum_{i=1}^{m} I\{\sigma_i \neq \nu_i\}$.

Proof. See [85], Lemma 2.9. □

CHAPTER II

ACTIVE LEARNING

2.1 Introduction

In most prediction problems, the main and only source of information is the training data obtained through a sequence of experiments. It is convenient to represent each datapoint as a pair *(Instance, Label)*, where *'Label'* often takes only finitely many values.

It was observed that in some cases, the cost related to the process of collecting the data is associated with the labels, while the pool of instances itself is almost unlimited. Examples include speech recognition which requires trained linguists for careful annotation; document and media classification which requires a human to provide a correct label to each object, etc. (see [78] for an excellent review and discussion of the applications). However, in most cases there is freedom to allocate the resources – namely, one can *choose* the data given to the expert for further processing and labeling. This motivated researchers to start the investigation of the learning strategies that are able to mimic this heuristic framework. Intuitively, an effective learning strategy should decide which instances are the most informative for a present task, thus omitting the 'less informative' observations.

The most popular, and probably the simplest model for active learning methods is the binary classification problem where the label is allowed to take only two values ± 1, and we will focus our attention on this particular case. Note that multi-label classification can often be reduced to binary classification, so it is important to understand the potential improvements in the easiest case. In particular, such improvements are possible due to the fact that for the binary classification problem, it is enough to

know only the *sign* of the regression function. If one wants to estimate the whole regression curve (for example, with respect to the L_2 loss), then active learning might not be helpful. In particular, a negative result of R. Nowak, R. Castro and R. Willett (Theorem 3 in [24]) implies the following: if we only assume that the regression function is smooth (namely, belongs to some Hölder ball, see Definition 2.1), active learning does not give any advantage (in the minimax sense) over passive learning.

There are several popular active learning frameworks which we mention below. The so-called Selective Sampling scenario is closest to our setting. In this case, the observations are assumed to come independently, one at a time, from some underlying distribution, and the algorithm decides whether the corresponding label should be requested or not. This approach goes back to the works of D. Cohn, L. Atlas, R. Ladner et al.[26],[25]. Another approach is the so-called Learning with Membership Queries, where an algorithm can request the label for *any* observation, see [2],[3]. This framework is often more restrictive, but does not necessarily give advantage over Selective Sampling (for example, under some assumptions on the underlying design distribution). Interesting theoretical results for these scenarios were obtained in [35], where authors propose and analyze the so-called Query-By-Committee algorithm. Another popular approach is Pool-Based active learning, where an algorithm sequentially selects the data from some large but fixed pool of observations [64]. This framework fits many practical applications, see [84],[79].

Most of the existing theoretical analysis was focusing on the noiseless case, thus making an assumption that there exists a perfect classifier that always predicts the data correctly, and that the training labels come from this classifier. The development and analysis of noise-robust algorithms turned out to be a harder problem. One of the first methods that performs well in the agnostic (noisy) setting is the A^2(agnostic active) algorithm by M.-F. Balcan, A. Beygelzimer, and J. Langford [6], and the margin-based active learning algorithm [8] for the case of linear separators. Another

approach was suggested by S. Dasgupta, D. Hsu and C. Monteleoni [28]. S. Hanneke developed new tools for theoretical analysis of active learning methods, see [41], and the further improvements by S.Hanneke, M.-F. Balcan and J. Wortman [7], [42], and V. Koltchinskii [52]. It was discovered that in some cases the generalization error of a resulting classifier can converge to zero exponentially fast with respect to its label complexity (while the best rate for passive learning is usually polynomial with respect to the cardinality of the training data set). Most of the aforementioned methods are based on empirical risk minimization subroutine, and the high-level idea is as follows:

1. Obtain the empirical risk minimizer $\hat{f}$;

2. Take a 'ball' $B(\hat{f}, \delta)$ around $\hat{f}$ that contains the best possible classifier f_* with high probability;

3. Find the 'disagreement set' associated to $B(\hat{f}, \delta)$ defined as

$$\Big\{x : \exists f_1, f_2 \in B(\hat{f}, \delta) \text{ s.t. sign } f_1(x) \neq \text{sign } f_2(x)\Big\}$$

   – in some cases, this set is going to be much smaller than the initial domain of observations;

4. Obtain the next group of labeled observations supported on the disagreement set and go back to step 1.

Under some natural assumptions these methods give provable significant improvements over passive learning, and, moreover, are adaptive with respect to the unknown parameters of the problem. On the other hand, empirical risk minimization that is at the core of these methods is done with respect to the binary loss. For such non-convex problems, it is very hard to find the minimizers over nontrivial hypotheses classes, so the practical use of the methods is limited. Other practically successful methods, such as Support Vector Machine Active Learning [84] have not yet been supported

by theoretical justification, to the best of our knowledge. A different approach in the noisy setting, based on learning the one-dimensional threshold classifiers, was proposed by R. Castro and R. Nowak [23]. The viewpoint and the methods of this work are closer to our investigation compared to previously mentioned results, so we will give more details below. We just mention that the algorithm developed in [23] can be effectively implemented, but it is not adaptive and requires some unknown parameters of the distribution as an input. One of the goals of the present work was to construct active learning methods that are computationally tractable and at the same time are adaptive over a large class of underlying distributions.

2.2 Probabilistic framework

Let $(X, Y) \in [0,1]^d \times \{-1, 1\}$ be a random couple with unknown distribution P. The marginal distribution of design variable X will be denoted by Π. Let $\eta(x) := \mathbb{E}(Y|X = x)$ be the regression function. The level set $\left\{x \in [0,1]^d : \ \eta(x) = 0\right\}$ is called the *decision boundary*. The goal of *binary classification* is to predict a label Y based on the observation X. Prediction is based on a *classifier* - a measurable function $f : [0,1]^d \mapsto \{-1, 1\}$. The quality of a classifier is measured in terms of its generalization error, $R(f) = \Pr(Y \neq f(X))$. It is well-known, and intuitively clear, that the best possible classifier is the so-called *Bayes classifier* $g_*(x) := \text{sign}\, \eta(x)$. The situation when active learning methods outperform passive algorithms might occur when the so-called *Tsybakov's low noise assumption* [69],[86] is satisfied: there exist constants $K, \gamma > 0$ such that

$$\forall\ t > 0,\ \Pi(x : |\eta(x)| \leq t) \leq Kt^{\gamma}. \tag{2.2.1}$$

This assumption provides a convenient way to characterize the noise level of the problem and will play a crucial role in our investigation. The majority of the previous work in the field was done under standard complexity assumptions on the set of possible classifiers(such as polynomial growth of the covering numbers). Castro and Nowak

[23] derived their results under the regularity conditions on the decision boundary and the noise assumption which is slightly more restrictive than (2.2.1). Essentially, they proved that if the decision boundary is a graph of the Hölder smooth function $g \in \Sigma(\beta, K, [0,1]^{d-1})$ (see Section 2.3 for definitions) and the noise assumption is satisfied with $\gamma > 0$, then the minimax lower bound for the expected excess risk of the active classifier is of order $N^{-\frac{\beta(1+\gamma)}{2\beta+\gamma(d-1)}}$ and the upper bound is $\mathcal{O}\left((N/\log N)^{-\frac{\beta(1+\gamma)}{2\beta+\gamma(d-1)}}\right)$, where N is the label budget. However, construction of the classifier that achieves an upper bound assumes β and γ to be known.

In this thesis, we consider the problem of active learning under classical nonparametric assumptions on the regression function – namely, we assume that it belongs to a certain Hölder class $\Sigma(\beta, K, [0,1]^d)$ and satisfies the low noise condition (2.2.1) with some positive γ.

Remark. Note that our assumption is different from the framework in [23] where smoothness condition was imposed on the level set $\{x : \eta(x) = 0\}$ and not on η itself.

Under similar assumptions on regularity of the regression function, A. Tsybakov and J.-Y. Audibert [5] showed that *plug-in classifiers* attain optimal rates in the *passive* learning framework. In particular, their results imply that the expected excess risk of a classifier $\hat{g} = \operatorname{sign} \hat{\eta}$ is bounded above by $C \cdot N^{-\frac{\beta(1+\gamma)}{2\beta+d}}$ (which is the optimal rate), where $\hat{\eta}$ is the local polynomial estimator of the regression function and N is the size of the training data set (their construction of an estimator $\hat{\eta}$ assumes β to be known). We were able to partially extend this claim to the case of active learning: first, we obtain minimax lower bounds for the excess risk of an active classifier in terms of its label complexity. Second, we propose a new algorithm that is based on plug-in classifiers, attains almost optimal rates over a broad class of distributions and possesses adaptivity with respect to β, γ (within the certain range of these parameters).

The rest of the chapter is organized as follows: the next section introduces remaining notations and specifies the main assumptions. This is followed by a qualitative description of our learning algorithm. The second part contains the statements and proofs of our main results - upper and minimax lower bounds for the excess risk.

2.3 Notations and main assumptions

Our *active learning* framework is governed by the following rules:

1. Observations are sampled sequentially: X_k is sampled from the modified distribution $\hat{\Pi}_k$ that depends on $(X_1, Y_1), \ldots, (X_{k-1}, Y_{k-1})$.

2. Y_k is sampled from the conditional distribution $P_{Y|X}(\cdot|X = x)$. Labels are conditionally independent given the feature vectors $X_i,\ i \leq n$.

Usually, the distribution $\hat{\Pi}_k$ is supported on a set where classification is difficult.

Given the probability measure $\mathbb{Q}$ on $S \times \{-1, 1\}$, we denote the integral with respect to this measure by $\mathbb{Q}g := \int g d\,\mathbb{Q}$. Let $\mathcal{F}$ be a class of bounded, measurable functions. The risk and the excess risk of $f \in \mathcal{F}$ with respect to the measure $\mathbb{Q}$ are defined by

$$R_{\mathbb{Q}}(f) := \mathbb{Q}\mathcal{I}_{y \neq \text{sign } f(x)}$$
$$\mathcal{E}_{\mathbb{Q}}(f) := R_{\mathbb{Q}}(f) - \inf_{g \in \mathcal{F}} R_{\mathbb{Q}}(g),$$

where $\mathcal{I}_{\mathcal{A}}$ is the indicator of event $\mathcal{A}$. We will omit the subindex $\mathbb{Q}$ when the underlying measure is clear from the context. Recall that we denoted the distribution of (X, Y) by P. The minimal possible risk with respect to P is

$$R^* = \inf_{g: S \mapsto [-1,1]} \Pr\left(Y \neq \text{sign } g(X)\right),$$

where the infimum is taken over all measurable functions. It is well known that it is attained for any g such that sign $g(x) =$ sign $\eta(x)$ Π - a.s. Given $g \in \mathcal{F},\ A \in \mathcal{B},\ \delta > 0$,

define

$$\mathcal{F}_{\infty,A}(g;\delta) := \{f \in \mathcal{F} :\ \|f-g\|_{\infty,A} \le \delta\},$$

where $\|f-g\|_{\infty,A} = \sup_{x\in A} |f(x)-g(x)|$. For $A \in \mathcal{B}$, define the function class

$$\mathcal{F}|_A := \{f|_A,\ f \in \mathcal{F}\},$$

where $f|_A(x) := f(x)\mathcal{I}_A(x)$.

Let $B > 0$.

Definition 2.1. *We say that $g : \mathbb{R}^d \mapsto \mathbb{R}$ belongs to $\Sigma(\beta, B, [0,1]^d)$, the $(\beta, B, [0,1]^d)$ - Hölder class of functions, if g is $\lfloor\beta\rfloor$ times continuously differentiable and for all $x, x_1 \in [0,1]^d$ satisfies*

$$|g(x_1) - T_x(x_1)| \le B\|x - x_1\|_\infty^\beta,$$

where T_x is the Taylor polynomial of degree $\lfloor\beta\rfloor$ of g at the point x.

Definition 2.2. *$\mathcal{P}(\beta,\gamma)$ is the class of probability distributions on $[0,1]^d \times \{-1,+1\}$ with the following properties:*

1. *There exisits $K > 0$ such that $\forall\ t > 0,\ \Pi(x : |\eta(x)| \le t) \le Kt^\gamma$;*
2. *$\eta(x) \in \Sigma(\beta, B, [0,1]^d)$.*

Remark: note that $\eta(x)$ can be defined arbitrarily on the sets of measure zero (with respect to Π), and we only ask for the existence of one smooth representative. We do not mention the dependence of $\mathcal{P}(\beta,\gamma)$ on the fixed constants B, K explicitly, but this should not cause any uncertainty.

Finally, let us define $\mathcal{P}^*_U(\beta,\gamma)$ and $\mathcal{P}_U(\beta,\gamma)$, the subclasses of $\mathcal{P}(\beta,\gamma)$, by imposing two additional assumptions. Along with the formal descriptions of these assumptions, we shall try to provide some motivation behind them. The first condition is related to the marginal distribution Π. For an integer $M \ge 1$, let

$$\mathcal{G}_M := \left\{ \left(\frac{k_1}{M}, \dots, \frac{k_d}{M}\right),\ k_i = 1 \dots M,\ i = 1 \dots d \right\}$$

be the regular grid on the unit cube $[0,1]^d$ with mesh size M^{-1}. It naturally defines a partition into a set of M^d open cubes $R_i,\ i = 1 \dots M^d$ with edges of length M^{-1} and vertices in $\mathcal{G}_M$. Below, we consider the nested sequence of grids $\{\mathcal{G}_{2^m},\ m \geq 1\}$ and corresponding dyadic partitions $\{\mathcal{H}_{2^m},\ m \geq 1\}$ of the unit cube.

Definition 2.3. *We will say that* Π *is* (u_1, u_2) *– regular if there exists* $m \geq 1$ *such that* Π *is supported on the union of elements of* $\{\mathcal{H}_{2^m}\}$ *and is absolutely continuous with respect to Lebesgue measure, with a density* $p(x)$ *such that* $\forall x \in \mathrm{supp}(\Pi)$*,*

$$u_1 \leq p(x) \leq u_2,$$

where $0 < u_1 \leq u_2 < \infty$*.*

Assumption 2.1. Π *is* (u_1, u_2) *- regular.*

Let us mention that our definition of regularity is somewhat restrictive; slightly more general conditions are possible at the price of more technical statements and details. We believe that small benefits from imposing a weaker assumption do not justify the losses in the clarity of exposition. For most of the chapter, the reader might think of Π as being uniform on $[0,1]^d$(however, we need slightly more complicated marginal to construct the minimax lower bounds for the excess risk).

It is known that estimation of regression function in sup-norm is sensitive to the geometry of design distribution, mainly because the quality of estimation depends on the *local* amount of data at every point; conditions similar to our *Assumption* 2.1 were used in the previous works where this problem appeared, for example, *strong density assumption* in [4] and *Assumption D* in [37]. A useful characteristic of (u_1, u_2) - regular distribution Π is that this property remains valid for conditional distributions $\Pi_A(dx) = \Pi(dx|A)$ for certain subsets A of the unit cube. This fact fits our active learning framework particularly well.

Definition 2.4. *We say that* $\mathbb{Q}$ *belongs to* $\mathcal{P}_U(\beta, \gamma)$ *if* $\mathbb{Q} \in \mathcal{P}(\beta, \gamma)$ *and Assumption 2.1 is satisfied for some* u_1, u_2*.*

The second assumption is crucial in derivation of the upper bounds. First, let us introduce the family of piecewise–polynomial functions that is used to construct the estimators of $\eta(x)$. Given two nonnegative integers r and m, let

$$\mathcal{F}_m^r := \left\{ f = \sum_{i=1}^{2^{dm}} q_i(x_1, \ldots, x_d) I_{R_i} \right\}, \tag{2.3.1}$$

where $\mathcal{H}_m = \left\{R_i,\ 1 \le i \le 2^{dm}\right\}$ is the dyadic partition of the unit cube and $q_i(x_1, \ldots, x_d)$ are the polynomials of degree at most r in d variables. For example, when $r = 0$, $\mathcal{F}_m^0$ can be viewed as the linear span of first 2^{dm} Haar basis functions on $[0,1]^d$. Note that $\{\mathcal{F}_m^r,\ m \ge 1\}$ is a nested family, which is usually a desirable property for statistical model selection procedures. By $\bar{\eta}_m(x)$ we denote the $L_2(\Pi)$ - projection of regression function $\eta(x)$ onto $\mathcal{F}_m^r$. We explain the motivation behind this choice of estimators and discuss the approximation properties of $\mathcal{F}_m^r$ in Section 2.3.1 below.
Let $\eta \in \Sigma(\beta, B, [0,1]^d)$ for some $0 < \beta \le r + 1$.

Assumption 2.2. *Assume one of the following two conditions holds:*

1. *$\eta(x)$ belongs to $\mathcal{F}_{m_0}^r$ for some $m_0 \ge 1$;*

2. *There exists $B_2 := B_2(\eta, \Pi) > 0$ such that for all $m \ge 1$ the following holds true:*

$$\|\eta - \bar{\eta}_m\|_{\infty, \mathrm{supp}(\Pi)} \ge B_2 2^{-\beta m}.$$

Finally, we define $\mathcal{P}_U^*(\beta, \gamma)$:

Definition 2.5. *We say that $\mathbb{Q}$ belongs to $\mathcal{P}_U^*(\beta, \gamma)$ if $\mathbb{Q} \in \mathcal{P}_U(\beta, \gamma)$ and Assumption 2.2 is satisfied.*

Appearance of *Assumption 2.2* is motivated by the structure of our learning algorithm – namely, it is based on adaptive confidence bands for the regression function.

Nonparametric confidence bands form a big topic in statistical literature, and the review of this subject is not our goal. We just mention that it is impossible to construct adaptive confidence bands of optimal size over the whole $\bigcup_{0<\beta\le r+1} \Sigma\left(\beta, K, [0,1]^d\right)$. The subject is discussed in details in [68, 40], among others. Assumption 2.2 plays an important role in controlling the bias of the estimator as can be seen from the proof of Theorems 2.5.6, 2.6.2 below. We continue the discussion of Assumption 2.2 in Section 2.5.4 and provide explicit examples of functions satisfying this assumption (for example, it will be shown that all sufficiently smooth functions satisfy Assumption 2.2).

2.3.1 Comparison Inequalities

Before proceeding with the main results, let us recall the well-known connections between the binary risk and the $\|\cdot\|_\infty$, $\|\cdot\|_{L_2(\Pi)}$ - norm risks:

Proposition 2.3.1. *Under the low noise assumption,*

$$R_P(f) - R^* \le D_1 \|(f-\eta)\mathcal{I}\left\{\text{sign } f \ne \text{sign } \eta\right\}\|_\infty^{1+\gamma}; \tag{2.3.2}$$

$$R_P(f) - R^* \le D_2 \|(f-\eta)\mathcal{I}\left\{\text{sign } f \ne \text{sign } \eta\right\}\|_{L_2(\Pi)}^{\frac{2(1+\gamma)}{2+\gamma}}; \tag{2.3.3}$$

$$R_P(f) - R^* \ge D_3 \Pi(\text{sign } f \ne \text{sign } \eta)^{\frac{1+\gamma}{\gamma}}. \tag{2.3.4}$$

Proof. For (2.3.2) and (2.3.3), see [4], Lemmas 5.1 and 5.2 respectively, and for (2.3.4)—see [53], Lemma 5.2. □

2.4 Learning algorithm: the first look

We proceed by giving a brief description of our learning algorithm, since several definitions appear naturally in this context. First, let us emphasize that *the marginal distribution* Π *is assumed to be known to the learner.* This is not a restriction, since we are not limited in the use of unlabeled data and Π can be estimated to any desired accuracy. Our construction is based on so-called *plug-in* classifiers of the

form $\hat{f}(\cdot) = \text{sign}\ \hat{\eta}(\cdot)$, where $\hat{\eta}$ is a piecewise-polynomial estimator of the regression function. As we have already mentioned above, it was shown in [4] that in the passive learning framework plug-in classifiers provide optimal rate of convergence for the excess risk of order $N^{-\frac{\beta(1+\gamma)}{2\beta+d}}$, with $\hat{\eta}$ being the local polynomial estimator.

Our active learning algorithm iteratively improves the classifier by constructing shrinking confidence bands for the regression function. On every step k, the piecewise-polynomial estimator $\hat{\eta}_k$ is obtained via the model selection procedure which allows adaptation to the unknown smoothness(for Hölder exponent $\leq r+1$). The estimator is further used to construct a confidence band $\hat{\mathcal{F}}_k$ for $\eta(x)$. The *active set* associated with $\hat{\mathcal{F}}_k$ is defined as

$$\hat{A}_k = A(\hat{\mathcal{F}}_k) := \left\{ x \in \text{supp}(\Pi) :\ \exists f_1, f_2 \in \hat{\mathcal{F}}_k, \text{sign}\ f_1(x) \neq \text{sign}\ f_2(x) \right\}.$$

Clearly, this is the set where the confidence band crosses zero level and where classification is potentially difficult. $\hat{A}_k$ serves as a support of the modified distribution $\hat{\Pi}_{k+1}$: on step $k+1$, label Y is requested only for observations $X \in \hat{A}_k$, forcing the labeled data to concentrate in the domain where higher precision is needed. This allows one to obtain a tighter confidence band for the regression function restricted to the active set. Since $\hat{A}_k$ approaches the decision boundary, its size is controlled by the low noise assumption. The algorithm does not require a priori knowledge of the noise and regularity parameters, being adaptive for $\gamma > 0, 0 < \beta \leq r+1$. See Figure 1 for graphical illustration. Further details are given in Section 2.6.2.

2.5 Approximation and estimation by piecewise–polynomial functions

In this section, we continue the discussion of classes $\mathcal{F}_m^r$ introduced in Section 2.3 (see (2.3.1)) and give some examples related to Assumption 2.2. We also introduce a piecewise–polynomial estimator of the regression function and discuss its concentration properties relative to $\|\cdot\|_\infty$ norm. Polynomial estimators are very common

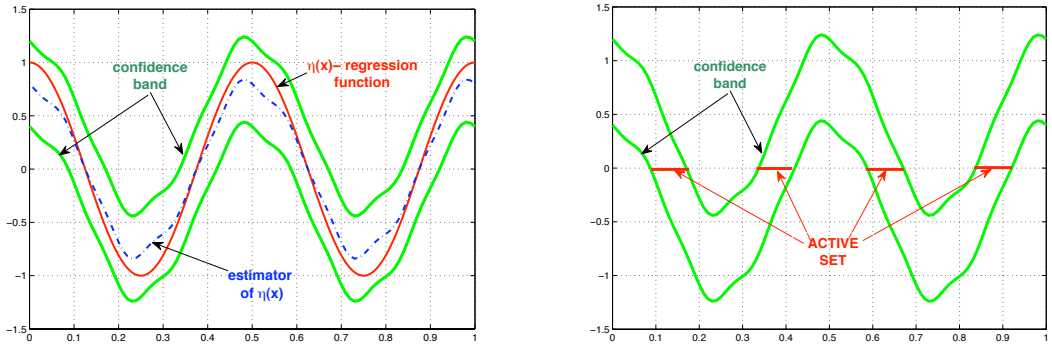

Figure 1: Active Learning Algorithm

in statistical literature on regression and density estimation due to their effectiveness and simplicity; most popular examples include B-splines and Bartle-Lemarié wavelets. In many cases it is desirable to construct a smooth estimator(for example, when it is known that the parameter of interest is smooth). However, our case is different in several aspects: due to the nature of the learning algorithm, the design distribution changes on every step and is often supported on a set with several connected components, so that using a globally smooth estimator does not necessarily give an advantage. It turns out that approximation by piecewise–polynomial functions better suits our goals, providing a simple and well-behaved estimator(which is only piecewise-smooth, however). In what follows, we often rely on the fact that the design distribution Π is known to the learner. In particular, we are able to construct $L_2(\Pi)$–orthonormal bases. It is possible to relax this condition by estimating all involved quantities to desired accuracy, but we omit these details to avoid deviations from our main exposition.

We proceed with a discussion of approximation properties of $\mathcal{F}_m^r$.

2.5.1 Approximation by piecewise–polynomial functions

It is a well known fact that the dimension of the space of polynomials in d variables of degree at most r on the unit cube is $\mathcal{D}_{d,r} := \binom{d+r}{r}$. This immediately implies

that $\dim \mathcal{F}_m^r = 2^{dm}\mathcal{D}_{d,r}$. Let $\phi_1, \dots, \phi_l,\ l \le 2^{dm}\mathcal{D}_{d,r}$ be any $L_2(\Pi)$ - orthonormal basis of $\mathcal{F}_m^r$ such that each ϕ_i is supported on a single dyadic cube $R_j,\ j = 1 \dots 2^{dm}$. A basis with required properties can be easily constructed as follows: take R_j such that $R_j \cap \operatorname{supp}(\Pi) \neq \emptyset$. Let $\phi_{j,1}, \dots, \phi_{j,\mathcal{D}_{d,r}}$ be the $L_2(\Pi)$–orthonormal basis of the restriction of $\mathcal{F}_m^r$ to R_j. The required basis of $\mathcal{F}_m^r$ is just the union of these bases over all j.

Let Proj_m be the $L_2(\Pi)$–projector on $\mathcal{F}_m^r$.

Proposition 2.5.1. *There exists a constant C that depends on Π, r, d but not on m such that for all Π-measurable, bounded $f : [0,1]^d \mapsto \mathbb{R}$ we have*

$$\|\operatorname{Proj}_m(f)\|_{\infty,\operatorname{supp}(\Pi)} \le C\|f\|_{\infty,\operatorname{supp}(\Pi)}.$$

Proof. Let $f_i := f \cdot \mathcal{I}_{R_i}$ be the restriction of f on the cube R_i, so that

$$f = \sum_{i:R_i \cap \operatorname{supp}(\Pi) \neq \emptyset} f_i.$$

Note that, due to the special structure of $\mathcal{F}_m^r$, $\operatorname{Proj}_m(f_i) = (\operatorname{Proj}_m(f))_i$, so it is enough to prove the claim for every f_i, with a constant C independent of i.

Let $\mathcal{P}_d^r$ be the space of polynomials of degree at most r on $[0,1]^d$, and assume that μ is the uniform distribution on a subset $S \subset [0,1]^d$ given by a union of dyadic cubes with edge length bounded from below by 2^{-m_0}. Since $\mathcal{P}_d^r$ is a finite dimensional subspace, for any $q \in \mathcal{P}_d^r$

$$\|q\|_{\infty,S} \le C(\mu, d, r)\|q\|_{L_2(\mu)}.$$

Note that the bound can be made uniform over all μ with described properties(since there are only finitely many candidates), with a constant C depending on m_0, d, r. Also, the same estimate clearly holds with $[0,1]^d$ replaced by *any* cube in $\mathbb{R}^d$ since it can be transformed into the unit cube by a composition of shifts, dilations and rotations.

Going back to our problem, consider a dyadic cube R_i with edge length 2^{-m} such

that $R_i \cap \operatorname{supp}(\Pi) \neq \emptyset$. Since, by Assumption 2.1, $R \cap \operatorname{supp}(\Pi)$ is a union of finitely many dyadic cubes, the previous argument implies

$$\|\mathrm{Proj}_m(f_i)\|_\infty \leq \frac{C(\Pi, d, r)}{\lambda\left(\operatorname{supp}(\Pi) \cap R_i\right)} \|\mathrm{Proj}_m(f_i)\|_{L_2(\lambda)}, \tag{2.5.1}$$

where λ is the Lebesgue measure on $\operatorname{supp}(\Pi) \cap R_i$ and the constant $C(\Pi, d, r)$ is independent from i. Recalling Assumption 2.1 once again, we have

$$\begin{aligned}\|\mathrm{Proj}_m(f_i)\|_{L_2(\lambda)} \leq \frac{1}{u_1}\|\mathrm{Proj}_m(f_i)\|_{L_2(\Pi)} \leq \frac{1}{u_1}\|f_i\|_{L_2(\Pi)} \leq \\ \leq \frac{\Pi(R_i)}{u_1}\|f_i\|_{\infty,\operatorname{supp}(\Pi)\cap R_i},\end{aligned}$$

where we used the fact that Proj_m is the projection, hence does not increase the norm. Together with (2.5.1), this gives

$$\begin{aligned}\|\mathrm{Proj}_m(f_i)\|_\infty \leq \frac{C(\Pi, d, r)}{\lambda\left(\operatorname{supp}(\Pi) \cap R_i\right)} \frac{\Pi(R_i)}{u_1}\|f_i\|_{\infty,\operatorname{supp}(\Pi)\cap R_i} \leq \\ \leq \frac{u_2}{u_1} C(\Pi, d, r)\|f_i\|_{\infty,\operatorname{supp}(\Pi)\cap R_i}\end{aligned}$$

where we used that $\frac{\Pi(R_i)}{\lambda(\operatorname{supp}(\Pi)\cap R_i)} \leq u_2$ by Assumption 2.1. Together with the initial observations, this completes the proof. □

The following corollary is straightforward(it is known as a 'Lebesgue lemma' in the approximation theory).

Corollary 2.5.2. *Let C be as in Proposition 2.5.1. Then for any bounded Π-measurable f we have*

$$\|f - \mathrm{Proj}_m f\|_{\infty,\operatorname{supp}(\Pi)} \leq (C+1) \inf_{g \in \mathcal{F}_m^r} \|f - g\|_{\infty,\operatorname{supp}(\Pi)}.$$

Proof. For any $g \in \mathcal{F}_m^r$ we have

$$\begin{aligned}\|f - \mathrm{Proj}_m f\|_{\infty,\operatorname{supp}(\Pi)} \leq \|f - g\|_{\infty,\operatorname{supp}(\Pi)} + \|g - \mathrm{Proj}_m f\|_{\infty,\operatorname{supp}(\Pi)} = \\ = \|f - g\|_{\infty,\operatorname{supp}(\Pi)} + \|\mathrm{Proj}_m(g - f)\|_{\infty,\operatorname{supp}(\Pi)} \leq \\ \leq \|f - g\|_{\infty,\operatorname{supp}(\Pi)} + C\|f - g\|_{\infty,\operatorname{supp}(\Pi)}.\end{aligned}$$

□

A consequence of the previous results is the following important fact:

Corollary 2.5.3. *Assume $f \in \Sigma(\beta, B, [0,1]^d)$ for $\beta \leq r+1$. Then*

$$\|f - \mathrm{Proj}_m f\|_{\infty, \mathrm{supp}(\Pi)} \leq B(C+1)2^{-\beta m},$$

where C is a constant from Proposition 2.5.1.

Remark 1. In what follows, we will denote $B_1 := B(C+1)$, where B and C are as defined above.

Proof. The claim follows from Definition 2.1 and Corollary 2.5.2. □

Remark 2. It is also possible to prove similar approximation results with a different approach: namely, one could observe that the kernel $k(x,y) := \sum_{j=1}^{\mathcal{D}_{d,r}} \phi_j(x)\phi_j(y)$ reproduces polynomials up to degree r Π–almost everywhere, and use this fact together with Taylor expansion to prove the approximation bound.

2.5.2 Estimation of piecewise–polynomial functions

This section discusses the random error that occurs when estimating a function from the noisy data. Our goal is to show that the piecewise-polynomial projection estimator concentrates around its expectation in sup-norm. Let $\mathcal{B}_m$ be the sigma-algebra generated by the dyadic cubes $R_j,\ 1 \leq j \leq 2^{dm}$ forming the partition of $[0,1]^d$. Given $A \in \mathcal{B}_m$ with $A_\Pi := A \cap \mathrm{supp}(\Pi) \neq \emptyset$, define

$$\hat{\Pi}_A(dx) := \Pi(dx | x \in A_\Pi).$$

Moreover, set $d_{m,A} := \dim\left(\mathcal{F}_m^r\big|_{A_\Pi}\right)$. Let $(X_i, Y_i),\ i \leq N$ be iid observations with $X_i \sim \hat{\Pi}_A(dx)$ and $m \geq 1$ be the resolution level. Let $\Phi_A = \left\{\phi_1, \ldots, \phi_{d_{m,A}}\right\} \subset \mathcal{F}_m^r$ be the $L_2(\Pi)$–orthonormal basis of $\mathcal{F}_m^r\big|_{A_\Pi}$, such that each ϕ_i is supported on a single dyadic cube $R = R(\phi_i)$ with edge length 2^{-m}. It is easy to see that in this case $\left\{\phi_1\sqrt{\Pi(A)}, \ldots, \phi_{d_{m,A}}\sqrt{\Pi(A)}\right\}$ is the $L_2(\hat{\Pi}_A)$–orthonormal basis of $\mathcal{F}_m^r\big|_{A_\Pi}$.

Next, we introduce an estimator of the regression function on the set A_Π. Define the empirical Fourier coefficients by

$$\hat{\alpha}_i := \frac{1}{N}\sum_{j=1}^{N} Y_j \sqrt{\Pi(A)}\phi_i(X_j)$$

and the estimator $\hat{\eta}_{m,A}$ by

$$\hat{\eta}_{m,A}(x) := \begin{cases} \sum_{i=1}^{d_{m,A}} \hat{\alpha}_i \sqrt{\Pi(A)}\phi_i(x), & x \in R \text{ for some } R \cap A_\Pi \neq 0, \\ 0, & \text{else.} \end{cases} \tag{2.5.2}$$

Here, R is a dyadic cube with edge length 2^{-m}. It is easy to see that for $x \in A_\Pi$, the mean of $\eta_{m,A}(x)$ is equal to $\bar{\eta}_m(x)$, where $\bar{\eta}_m$ is the $L_2(\Pi)$– projection of η onto $\mathcal{F}_m^r$. Let α_i be such that $\bar{\eta}_m\big|_{A_\Pi} = \sum_{i=1}^{d_{m,A}} \alpha_i \phi_i$.

The goal of this section is to prove the following concentration result:

Theorem 2.5.4. *For any $t > 0$,*

$$\Pr\left(\sup_{x\in A_\Pi} |\hat{\eta}_{m,A}(x) - \bar{\eta}_m(x)| > C_1 t\sqrt{\frac{2^{dm}\Pi(A)}{N}}\right) \le 2d_{m,A}\exp\left(\frac{-t^2/2}{1 + C_2 t\sqrt{\frac{2^{dm}\Pi(A)}{N}}}\right).$$

Remark: we will often apply the bound of the theorem for a random resolution level m which is based on a sample independent from (X_i, Y_i), $i \le N$. In this case, the bound has to be applied conditionally on m. If it also known that m is bounded, namely, $\frac{2^{dm}\Pi(A)}{N} \le C$ (as it will be in our case), then $d_{m,A} \le C_1 \cdot 2^{dm}\Pi(A) \le C_2 \cdot N$ and the bound can be rewritten as

$$\begin{aligned} &\Pr\left(\sup_{x\in A_\Pi} |\hat{\eta}_{m,A}(x) - \bar{\eta}_m(x)| > C_3 \log\frac{N}{\alpha}\sqrt{\frac{2^{dm}\Pi(A)}{N}}\right) = \\ &= \mathbb{E}\Pr\left(\sup_{x\in A_\Pi} |\hat{\eta}_{m,A}(x) - \bar{\eta}_m(x)| > C_3 \log\frac{N}{\alpha}\sqrt{\frac{2^{dm}\Pi(A)}{N}}\,\middle|\, m\right) \le \\ &\le \mathbb{E}\left(C_4 N\exp\left(-C_5\log\frac{N}{\alpha}\right)\right) \le \alpha \end{aligned}$$

for C_3 large enough.

Proof. The estimate follows from Bernstein's inequality. Let $R \in \mathcal{B}_m$ be a dyadic cube with edge length 2^{-m} such that $R \cap A \neq \emptyset$ and assume $x \in R$. Without loss of generality, let $\phi_1, \dots, \phi_{\mathcal{D}_{d,r}}$ be the basis functions that are active on R(that is, not identically zero on R). Then

$$\left|\hat{\eta}_{m,A}(x) - \bar{\eta}_m(x)\right| = \left|\sum_{i=1}^{\mathcal{D}_{d,r}} (\hat{\alpha}_i \sqrt{\Pi(A)} - \alpha_i)\phi_i(x)\right| \leq$$

$$\leq \max_{i \leq \mathcal{D}_{d,r}} |\hat{\alpha}_i \sqrt{\Pi(A)} - \alpha_i| \sup_{x \in R} \sum_{i=1}^{\mathcal{D}_{d,r}} |\phi_i(x)| \leq C(\Pi, d, r) \frac{1}{\sqrt{\Pi(R)}} \max_{i \leq \mathcal{D}_{d,r}} |\hat{\alpha}_i \sqrt{\Pi(A)} - \alpha_i|,$$

where the last inequality follows from the equivalence of $\|\cdot\|_{L_2(\Pi)}$ and $\|\cdot\|_\infty$ as in Proposition 2.5.1.

Let $\xi_{i,j} := Y_j \phi_i(X_j) \Pi(A)$ so that $\hat{\alpha}_i \sqrt{\Pi(A)} = \frac{1}{N} \sum_{j=1}^{N} \xi_{i,j}$. Note that

$$\begin{aligned}
\mathbb{E}\xi_{i,j} &= \Pi(A) \int_R \eta(y)\phi_i(y) d\hat{\Pi}_A(y) = \alpha_i, \\
\mathbb{E}\xi_{i,j}^2 &= \Pi^2(A) \int_R \phi_i^2(y) d\hat{\Pi}_A(y) = \Pi(A), \\
|\xi_{i,j}| &\leq C(\Pi, d, r) \frac{\Pi(A)}{\sqrt{\Pi(R)}}, \text{ almost surely.}
\end{aligned}$$

Bernstein's inequality implies that

$$\Pr\left(\left|\hat{\alpha}_i \sqrt{\Pi(A)} - \alpha_i\right| > t\sqrt{\frac{\Pi(A)}{N}}\right) \leq 2\exp\left(\frac{-t^2/2}{1 + \frac{t}{3}\sqrt{\frac{\Pi(A)}{N\Pi(R)}}}\right).$$

The union bound over $i \leq \mathcal{D}_{d,r}$ implies

$$\Pr\left(\frac{1}{\sqrt{\Pi(R)}} \max_{i \leq \mathcal{D}_{d,r}} \left|\hat{\alpha}_i \sqrt{\Pi(A)} - \alpha_i\right| > t\sqrt{\frac{\Pi(A)}{N\Pi(R)}}\right) \leq 2\mathcal{D}_{d,r} \exp\left(\frac{-t^2/2}{1 + \frac{t}{3}\sqrt{\frac{\Pi(A)}{N\Pi(R)}}}\right).$$

It remains to recall that by our assumptions $\Pi(R) \geq u_1 2^{-dm}$. Together with the union bound over all R such that $R \cap A \neq \emptyset$, this completes the proof. □

2.5.3 Model selection

This section describes the tools that are needed to make our learning algorithm adaptive with respect to the unknown smoothness β. It turns out that if Assumption 2.2

is satisfied, then information about smoothness can be captured from the data. Our adaptation procedure is of "Lepski–type"[63]. The approach presented below was partially motivated by recent strong results of Giné and Nickl [40] on adaptive confidence bands in density estimation. We employ similar techniques as in the proof of Lemma 2 in the above–cited paper. An approach that is different from presented here(based on complexity-penalized model selection), which also allows to obtain adaptive confidence bands, is given in [73] for the case of piecewise-constant functions.

Given a sequence of finite dimensional subspaces $\mathcal{G}_m$(in our case, these are the piecewise–polynomial functions $\mathcal{F}_m^r$, possibly restricted to some subset of $[0,1]^d$), define the index set

$$\mathcal{J}(N) := \left\{ m \in \mathbb{N} : \ 1 \leq \dim \mathcal{G}_m \leq \frac{N}{\log^4 N} \right\} \tag{2.5.3}$$

which is the set of all possible resolution levels of an estimator from $\mathcal{G}_m$ based on a sample of size N. The upper bound on $\mathcal{J}(N)$ in imposed to make sure that the resulting estimator is consistent. For the model selection procedures described below, we will always assume that the index is chosen from the corresponding $\mathcal{J}(N)$.

Given a sample $(X_1, Y_1), \ldots, (X_N, Y_N)$ from P, let $\left\{ \hat{\eta}_m := \hat{\eta}_{m,[0,1]^d}, \ m \in \mathcal{J}(N) \right\}$ be a collection of estimators of η on the unit cube defined by formula (2.5.2). Our goal is to choose the resolution level m in an optimal way using the given sample. Optimality is understood as a balance between the bias term coming from the polynomial approximation and the random error coming from the use of noisy data. Given $t > 1$, define

$$\hat{m} := \hat{m}(t, N) = \min \left\{ m \in \mathcal{J}(N) : \ \forall l > m, \ l \in \mathcal{J}(N), \ \|\hat{\eta}_l - \hat{\eta}_m\|_\infty \leq K_1 t \sqrt{\frac{2^{dl} l}{N}} \right\}. \tag{2.5.4}$$

Remark 1: for brevity, everywhere in this section the sup–norm $\|\cdot\|_\infty$ stands for $\|f - g\|_{\infty, \mathrm{supp}(\Pi)} := \sup_{x \in \mathrm{supp}(\Pi)} |f(x) - g(x)|$.

Remark 2: Note that the actual value of $\hat{m}$ can be numerically computed once the estimators $\hat{\eta}_l$ have simple structure. For example, when $r = 0$ and $\hat{\eta}_l$ are piecewise-constant functions, the running time of the algorithm that finds $\hat{m}$ is $\mathcal{O}(N \log^2 N)$. This follows from the following considerations: first, we can write

$$\hat{\eta}_l(x) = \sum_{i: R_i \cap \mathrm{supp}(\Pi) \neq \emptyset} \frac{\sum_{j=1}^{N} Y_j \mathcal{I}_{R_i}(X_j)}{N \cdot \Pi(R_i)} \mathcal{I}_{R_i}(x),$$

where R_j have edge length 2^{-l}, hence $\hat{\eta}_l$ can be found in $\mathcal{O}(N)$ steps and takes at most $\mathcal{O}(N)$ different values. The sup-norm $\|\hat{\eta}_l - \hat{\eta}_m\|_\infty$ can be found in $\mathcal{O}(N)$ steps. There are at most $\mathcal{O}(\log N)$ models to consider, which totally gives at most $\mathcal{O}(N \log^2 N)$ running time.

We will compare $\hat{m}$ to the 'optimal' resolution level $\bar{m}$ defined by

$$\bar{m} := \min \left\{ m \in \mathcal{J}(N) : \|\eta - \bar{\eta}_m\|_\infty \leq K_2 \sqrt{\frac{2^{dm} m}{N}} \right\}. \tag{2.5.5}$$

For $\bar{m}$, we immediately get the following:

Lemma 2.5.5. *If $\eta \in \Sigma(B, [0,1]^d, \beta)$ for $0 < \beta \leq r + 1$, then*

$$2^{\bar{m}} \leq C_1 \cdot \left(\frac{N B_1^2}{\log(N B_1^2)} \right)^{1/(2\beta + d)},$$

where $B_1 = C \cdot B$ for some $C = C(\Pi, d, r)$.

Moreover, if condition 2 of Assumption 2.2 is satisfied with a constant B_2, then

$$2^{\bar{m}} \geq C_2 \cdot \left(\frac{N B_2^2}{\log(N B_2^2)} \right)^{1/(2\beta + d)}$$

Remark: note that we specify the dependence on B_1, B_2 since we will allow these constants to logarithmically depend on N later.

Proof. By the definition of $\bar{m}$ and Corollary 2.5.3, we have

$$\bar{m} \leq \min \left\{ m \in \mathcal{J}(N) : B_1 2^{-\beta m} \leq K_2 \sqrt{\frac{2^{dm} m}{N}} \right\}.$$

Clearly, the minimum can be bounded by m such that $2^m \simeq C_4 \left(\frac{NB_1^2}{\log(NB_1^2)}\right)^{1/(2\beta+d)}$, for appropriate C_4.

The second bound follows in a similar way, with Assumption 2.2 in place of Corollary 2.5.3 to estimate $\|\eta - \bar{\eta}_m\|_\infty$:

$$\bar{m} \geq \min\left\{m \in \mathcal{J}(N):\ B_2 2^{-\beta m} \leq K_2\sqrt{\frac{2^{dm}m}{N}}\right\} \geq$$
$$\geq \max\left\{m \in \mathcal{J}(N):\ B_2 2^{-\beta m} > K_2\sqrt{\frac{2^{dm}m}{N}}\right\},$$

and it is easy to see that $B_2 2^{-\beta m} > K_2\sqrt{\frac{2^{dm}m}{N}}$ is satisfied for $2^m \simeq C_5 \left(\frac{B_2^2 N}{\log(B_2^2 N)}\right)^{1/(2\beta+d)}$, if C_5 is sufficiently small. □

The main result of this subsection is formulated below.

Theorem 2.5.6. *Assume that $\eta \in \Sigma(B, [0,1]^d, \beta)$ and that Assumption 2.2 is satisfied with constant B_2. Then there exists $t_0 = t_0(\Pi, d, r) > 0$ and K_1 large enough such that for all $t \geq t_0$ we have*

$$\hat{m} \in \left(\bar{m} - \frac{1}{\beta}\left(\log_2 t + \log_2 \frac{B_1}{B_2} + h\right), \bar{m}\right]$$

with probability at least $1 - C2^{d\bar{m}} \log N \exp(-ct\bar{m})$, where h is some fixed positive number that depends on d, r, Π and $B_1 = C \cdot B$ for some $C = C(\Pi, d, r)$.

Remark: the bound above becomes useful if the endpoints of the interval are of the same order, which retranslates as a condition for the size of $\bar{m}$, or, using Lemma 2.5.5, can be further written in terms of N. Assume that we want $\hat{m} \in (\varepsilon\bar{m}, \bar{m})$ to hold with probability $\geq 1 - \alpha$. If, in addition, $\frac{B_1}{B_2} \leq C \log N$ (as we will assume in the sequel), then it is enough to require

$$t \geq c_1(2\beta + d)\left(1 + \frac{\log(1/\alpha)}{\log N}\right),$$
$$N \geq c_2 \left(\log N \vee t\right)^{(2+\varepsilon)(2\beta+d)/\beta}.$$

If A is such that $\alpha = N^{-A}$, then the last condition can be stated as

$$N \geq c_3(\beta, d) \cdot (\log N \vee A)^{(2+\varepsilon)(2\beta+d)/\beta}. \tag{2.5.6}$$

Proof. Let $m > \bar{m}$ be fixed; by definition of $\hat{m}$, we have

$$\Pr(\hat{m} = m) \leq \Pr\left(\exists l > m-1 : \ \|\hat{\eta}_l - \hat{\eta}_{m-1}\|_\infty > K_1 t \sqrt{\frac{2^{dl} l}{N}}\right) \leq$$
$$\leq \sum_{l \in \mathcal{J}(N):\ l > m-1} \Pr\left(\|\hat{\eta}_l - \hat{\eta}_{m-1}\|_\infty > K_1 t \sqrt{\frac{2^{dl} l}{N}}\right).$$

We will bound each term independently. Note that

$$\|\hat{\eta}_l - \hat{\eta}_{m-1}\|_\infty \leq \|\hat{\eta}_l - \bar{\eta}_l\|_\infty + \|\hat{\eta}_{m-1} - \bar{\eta}_{m-1}\|_\infty + \|\eta - \bar{\eta}_l\|_\infty + \|\eta - \bar{\eta}_{m-1}\|_\infty.$$

Since $l > m-1 \geq \bar{m}$, we have by the definition of $\bar{m}$

$$\|\eta - \bar{\eta}_l\|_\infty \leq K_2 \sqrt{\frac{2^{dl} l}{N}}, \quad \|\eta - \bar{\eta}_{m-1}\|_\infty \leq K_2 \sqrt{\frac{2^{d(m-1)}(m-1)}{N}} < K_2 \sqrt{\frac{2^{dl} l}{N}}.$$

If $\frac{1}{4} K_1 t > K_2$, this gives

$$\Pr\left(\|\hat{\eta}_l - \hat{\eta}_{m-1}\|_\infty > K_1 t \sqrt{\frac{2^{dl} l}{N}}\right) \leq$$
$$\leq \Pr\left(\|\hat{\eta}_l - \bar{\eta}_l\|_\infty > K_3 t \sqrt{\frac{2^{dl} l}{N}}\right) + \Pr\left(\|\hat{\eta}_{m-1} - \bar{\eta}_{m-1}\|_\infty > K_3 t \sqrt{\frac{2^{dl} l}{N}}\right),$$

where $K_3 = K_2/4$. It remains to apply Theorem 2.5.4(for $A := [0,1]^d$) to estimate the probabilities above, and to take the sum over all $l > m-1$. Note that, by definition of $\mathcal{J}(N)$, $\sqrt{\frac{2^{dl} l}{N}}$ is bounded above by a constant independent of l and that the dimension factor $d_{l,[0,1]^d}$ from Theorem 2.5.4 is bounded by $2^{dl}\mathcal{D}_{d,r} \sim C 2^{dl}$. Finally, for t large enough

$$\Pr(\hat{m} = m) \leq \sum_{l \in \mathcal{J}(N):\ l > m-1} C 2^{dl} \exp\left(\frac{-t^2 l}{c_1 t}\right) +$$
$$+ C 2^{d(m-1)} \sum_{l \in \mathcal{J}(N):\ l > m-1} \exp\left(-c_3 t 2^{d(l-m+1)} l\right) \leq C 2^{d(m-1)} \exp\left(-ct(m-1)\right).$$

Since $m > \bar{m}$ was picked arbitrarily, it remains to use the union bound over all such m to get that

$$\Pr\left(\hat{m} > \bar{m}\right) \leq C2^{d\bar{m}} \exp\left(-ct\bar{m}\right). \tag{2.5.7}$$

Now we turn our attention to the reverse inequality. Let m be such that

$$m < \bar{m} - \frac{1}{\beta}\left(\log_2 t + \log_2 \frac{B_1}{B_2} + h\right) \tag{2.5.8}$$

where h is some fixed positive integer that depends on d, r, Π. By the definition (2.5.4) of $\hat{m}$, we have

$$\Pr\left(\hat{m} = m\right) \leq \Pr\left(\|\hat{\eta}_m - \hat{\eta}_{\bar{m}}\|_\infty \leq K_1 t\sqrt{\frac{2^{d\bar{m}}\bar{m}}{N}}\right). \tag{2.5.9}$$

Next, by triangle inequality

$$\|\hat{\eta}_m - \hat{\eta}_{\bar{m}}\|_\infty \geq \|\bar{\eta}_m - \eta\|_\infty - \|\bar{\eta}_{\bar{m}} - \eta\|_\infty - \|\hat{\eta}_m - \bar{\eta}_m - \hat{\eta}_{\bar{m}} + \bar{\eta}_{\bar{m}}\|_\infty.$$

By Assumption 2.2, $\|\bar{\eta}_m - \eta\|_\infty \geq B_2 2^{-\beta m}$. By the definition of $\bar{m}$, $\|\bar{\eta}_{\bar{m}} - \eta\|_\infty \leq K_2\sqrt{\frac{2^{d\bar{m}}\bar{m}}{N}}$. Together with the inequality (2.5.9), this implies

$$\Pr\left(\hat{m} = m\right) \leq \Pr\left(\|\hat{\eta}_m - \bar{\eta}_m - \hat{\eta}_{\bar{m}} + \bar{\eta}_{\bar{m}}\|_\infty \geq B_2 2^{-\beta m} - (K_1 t + K_2)\sqrt{\frac{2^{d\bar{m}}\bar{m}}{N}}\right). \tag{2.5.10}$$

Since we also have(by definition of $\bar{m}$) that $B_1 2^{-\beta(\bar{m}-1)} \geq K\sqrt{\frac{2^{d(\bar{m}-1)(\bar{m}-1)}}{N}}$, our assumption (2.5.8) on m implies, for h large enough,

$$B_2 2^{-\beta m} - (K_1 t + K_2)\sqrt{\frac{2^{d\bar{m}}\bar{m}}{N}} \geq c_1 t\sqrt{\frac{2^{d\bar{m}}\bar{m}}{N}}.$$

It remains to apply Theorem 2.5.4 to further bound (2.5.10) from above. We get

$$\begin{aligned}\Pr\left(\hat{m} = m\right) &\leq \Pr\left(\|\hat{\eta}_m - \bar{\eta}_m\|_\infty \geq \frac{c_1}{2} t\sqrt{\frac{2^{d\bar{m}}\bar{m}}{N}}\right) + \Pr\left(\|\hat{\eta}_{\bar{m}} - \bar{\eta}_{\bar{m}}\|_\infty \geq \frac{c_1}{2} t\sqrt{\frac{2^{d\bar{m}}\bar{m}}{N}}\right) \\ &\leq C2^{d\bar{m}} \exp\left(-ct\bar{m}\right).\end{aligned}$$

The union bound over all m satisfying condition (2.5.8)(which gives $\mathcal{O}(\log N)$ choices) completes the proof. □

Remark. It easily follows from the definition of $\bar{m}$ and the previous theorem that the following inequality holds with probability $\geq 1-\alpha$ (if $\frac{B_1}{B_2} \lesssim \log N$):

$$2^{-\beta \hat{m}_0} \leq \frac{C}{\beta} \left(\log \frac{N}{\alpha} \right)^{1+\frac{d}{2\beta}} \sqrt{\frac{2^{d\hat{m}} \hat{m}}{N}}. \tag{2.5.11}$$

This inequality will be useful in the sequel to control the bias, since $\frac{2^{d\hat{m}}\hat{m}}{N}$ is a purely data-dependent quantity. If we assume that $\beta \geq \nu$ for some $\nu > 0$, then the upper bound only depends on known parameters.

2.5.4 Functions satisfying Assumption 2.2

The goal of this subsection is to get more transparent description of the limitations imposed by Assumption 2.2 on the class of underlying distributions. We will show that all "nice" functions satisfy the requirements. The meaning of the following two propositions can be summarized in the informal way as follows: functions that satisfy Assumption 2.2 are the functions whose smoothness β can be learned from the data.

Proposition 2.5.7. *Assume $\eta \in C^{r+1}\left([0,1]^d\right)$, the space of $(r+1)$–times continuously differentiable functions. Then Assumption 2.2 is satisfied.*

Proof. First, note that, if $(D^{r+1}\eta)\big|_{\mathrm{supp}(\Pi)} \equiv 0$, then the first condition of Assumption 2.2 holds. Otherwise, there exist $x_0 \in \operatorname{int}\operatorname{supp}(\Pi)$ such that $D^{r+1}\eta(x_0) \neq 0$, meaning that at least one of the partial derivatives of order $r+1$ is nonzero. Define

$$M_0 := \max_{|\alpha|=r+1} \left| \frac{D^\alpha f}{\alpha!}(x_0) \right|$$

where α is a multi index, and we employ the standard multi index notation below. Let $R \in \mathcal{H}_{2^m}$ be a cube with edge length 2^{-m}, such that $x_0 \in \operatorname{int} R$. Assume that z_0 is the vertex of R closest to the origin, so that the change of variables $y = 2^m(x - z_0)$ transforms R into a unit cube. Our main step is based on the following fact: there exists $c(r+1,d) > 0$ such that for all monic polynomials p of power $r+1$ in d variables

$$\sup_{x \in [0,1]^d} |p(x)| \geq c(r+1,d). \tag{2.5.12}$$

Remark: we will say that a polynomial $p = \sum\limits_{|\alpha|\le k} c_\alpha x^\alpha$ of degree k in d variables is *monic* iff

$$\max_{|\alpha|=k} |c_\alpha| = 1.$$

In what follows, $\mathcal{M}(k,d)$ denotes the set of all monic polynomials in d variables of degree k.

One way to see (2.5.12) is as follows: first, since the norms on the space of polynomials of bounded degree are equivalent, we have $\sup_{x\in[0,1]^d} |p(x)| \ge c_1(r,d)\|p(x)\|_{L_2(dx)}$, where $\|\cdot\|_{L_2(dx)}$ is the usual L_2–norm with respect to the Lebesgue measure. Next, for $p(x) = \sum_{|\alpha|\le r+1} c_\alpha x^\alpha$, we have

$$\|p\|^2_{L_2(dx)} = \int\limits_{[0,1]^d} \left(\sum_{|\alpha|\le r+1} c_\alpha x^\alpha \right)^2 dx = \sum_{|\alpha|,|\beta|\le r+1} c_\alpha c_\beta A_{\alpha,\beta} = c^T A c,$$

where $A_{\alpha,\beta} = \int\limits_{[0,1]^d} x^\alpha x^\beta dx$ and c is a vector of coefficients of p (ordered lexicographically, for example). Since the set of monomials is linearly independent, A is positive definite as a Gram matrix of a linearly independent set, hence

$$c^T A c \ge \lambda_{\min}(A)\|c\|_2^2 := c_2(r,d)\|c\|_2^2.$$

It remains to notice that $\|c\|_2 \ge 1$ for monic polynomials.

Going back to the main argument, let $T_{r+1}(x;x_0)$ be the Taylor polynomial of η at x_0, so that

$$\eta(x) = T_{r+1}(x;x_0) + E(x-x_0),$$

where $E(x-x_0) = o(\|x-x_0\|_\infty^{r+1})$. We have

$$\begin{aligned}\|T_{r+1}(x;x_0) - \bar\eta_m(x)\|_{\infty,R} &= \|T_{r+1}(2^{-m}y + z_0;x_0) - \bar\eta_m(2^{-m}y+z_0)\|_{\infty,[0,1]^d} \ge \\ &\ge M_0 2^{-m(r+1)} \inf_{p\in\mathcal{M}(r+1,d)} \|p\|_{\infty,[0,1]^d} \ge c(r+1,d)M_0 2^{-m(r+1)}\end{aligned}$$

If $m_0 = m_0(\eta,x_0)$ is such that $|E(x-x_0)| < \frac{1}{2}c(r+1,d)M_0\|x-x_0\|^{r+1}$ for any

$\|x - x_0\|_\infty \le 2^{-m}$, then for any $m \ge m_0$

$$\|\eta - \bar{\eta}_m\|_{\infty,R} \ge \|\eta - T_{r+1}(\cdot; x_0)\|_{\infty,R} - \|E(x)\|_{\infty,R} \ge$$
$$\ge \frac{1}{2} c(r+1,d) M_0 2^{-m(r+1)},$$

concluding the proof. □

When $\beta \neq r+1$, we have the following result:

Proposition 2.5.8. *Assume that* $\{\beta\} = \beta - \lfloor\beta\rfloor > 0$ *and define*

$$f_u(t) := \eta(x_0 + tu),\ t \in \mathbb{R},\ u \in \mathbb{R}^d,\ \|u\|_\infty = 1.$$

If there exist $x_0 \in (0,1)^d$ *and* u *such that*

$$\lim_{t\to 0} \left| \frac{f_u^{(\lfloor\beta\rfloor)}(t) - f_u^{(\lfloor\beta\rfloor)}(0)}{|t|^{\{\beta\}}} \right| = M > 0,$$

then Assumption 2.2 is satisfied.

Proof. For $-1 \le s \le 1$, define $g_j(s) := \frac{f_u^{(\lfloor\beta\rfloor)}(2^{-j}s) - f_u^{(\lfloor\beta\rfloor)}(0)}{2^{-\{\beta\}j}}$. Note that

$$g_j(s) \xrightarrow[j\to\infty]{} M \cdot g(s),$$

where $g(s)$ can be one of four functions $g_{1,2}(s) := \pm|s|^{\{\beta\}}$ or $g_{3,4}(s) := \pm|s|^{\{\beta\}}\mathrm{sign}(s)$, and, moreover, convergence is uniform in s over any compact interval, due to our assumptions. Indeed, the function $h(t) := f_u^{(\lfloor\beta\rfloor)}(t) - f_u^{(\lfloor\beta\rfloor)}(0)$ must be nonzero in some neighborhood $t \in (-\delta,\delta) \setminus \{0\}$ (since $M > 0$), so it can only change the sign at $t = 0$. The claim about uniformity is also straightforward:

$$\sup_s |g_j(s) - M \cdot g(s)| = \sup_s \left| \left| \frac{f_u^{(\lfloor\beta\rfloor)}(2^{-j}s) - f_u^{(\lfloor\beta\rfloor)}(0)}{(2^{-j}s)^{\{\beta\}}} \right| - M \right| \cdot |s|^{\{\beta\}} \xrightarrow[j\to\infty]{} 0.$$

We will also define (motivated by the integral form of the remainder term in Taylor expansion)

$$q_j(t) = \int_0^t g_j(u)(t-u)^{\lfloor\beta\rfloor - 1} du.$$

Clearly, $q_j(t) \xrightarrow[j\to\infty]{} M \cdot q(t) := M \int\limits_0^t g(u)(t-u)^{\lfloor\beta\rfloor-1}du$ uniformly over compact subsets. Our previous observations imply (for example, direct evaluation of $q(t)$ gives $|q(t)| = |t|^\beta \int\limits_0^1 u^{\{\beta\}}(1-u)^{\lfloor\beta\rfloor-1}du = C|t|^\beta$) that for any closed interval I containing 0

$$D = D(r, \eta, I) := \mathrm{dist}_{\infty,I}(q, \mathcal{P}_r) > 0$$

– in other words, the distance in $C(I)$ from $q(t)$ to the subspace $\mathcal{P}_r$ of polynomials of degree at most r is positive. Due to uniform convergence, there exists j_0 such that that for any $j \geq j_0$, $\mathrm{dist}_\infty(q_j, \mathcal{P}_r) \geq M\frac{D}{2}$. As a consequence, for a dyadic cube R containing x_0 and $I = I(x_0) \subset \mathbb{R}$ such that

$$x_0 + 2^{-m}tu \in R \text{ for } t \in I$$

we have

$$\|\eta - \bar{\eta}_m\|_{\infty,R} \geq \|f_u(2^{-m}t) - \bar{\eta}_m(x_0 + 2^{-m}tu)\|_{\infty,I} =$$
$$= \left\| T_{\lfloor\beta\rfloor}(2^{-m}t; 0) + \frac{2^{-\beta m}}{(\lfloor\beta\rfloor-1)!} \int\limits_0^t g_m(u)(t-u)^{\lfloor\beta\rfloor-1}du - \bar{\eta}_m(x_0 + 2^{-m}tu) \right\|_{\infty,I},$$

where $T_{\lfloor\beta\rfloor}(2^{-m}t; 0)$is the Taylor polynomial of $f_u(t)$ of degree $\lfloor\beta\rfloor$ expanded around 0 and g_m is defined above. Consequently, for $m \geq j_0$,

$$\|\eta - \bar{\eta}_m\|_{\infty,R} \geq \frac{2^{-\beta m}}{(\lfloor\beta\rfloor-1)!} \inf_{p\in\mathcal{P}_r} \|q_m(t) - p(t)\|_{\infty,I} \geq \frac{2^{-\beta m}}{(\lfloor\beta\rfloor-1)!} M\frac{D}{2},$$

completing the proof. □

To end this section, we will mention that results of the same flavor appeared in a recent work of E. Giné and R. Nickl on adaptive density estimation [40], where a condition similar to our Assumption 2.2 was studied in the case of wavelet projection estimators. In particular, due to a nice characterization of smooth classes in terms of wavelet coefficients, authors were able to show that the functions that do not satisfy two-sided inequalities for approximation by wavelet projection form a nowhere dense subset of the corresponding smoothness class.

2.6 Main results

The question we address below is: what are the best possible rates that can be achieved by active learning algorithms in our framework and how these rates can be attained. We start this section by investigating the theoretical limitations of active learning under our assumptions on the underlying distribution. The second subsection is devoted to the detailed description and analysis of the learning algorithm introduced earlier.

2.6.1 Minimax lower bounds for the excess risk

The goal of this section is to prove that for $P \in \mathcal{P}(\beta, \gamma)$, no active learner can output a classifier with expected excess risk converging to zero faster than $N^{-\frac{\beta(1+\gamma)}{2\beta+d-\beta\gamma}}$. Our result builds upon the minimax bounds appeared in the works of A. Tsybakov, J.-Y. Audibert [4] and R. Nowak, R. Castro [23].

Remark The theorem below is proved for a smaller class $\mathcal{P}_U^*(\beta, \gamma)$, which implies the result for $\mathcal{P}(\beta, \gamma)$.

Theorem 2.6.1. *Let β, γ, d be such that $\beta\gamma \leq d$. Then there exists $C > 0$ such that for all n large enough and for any active classifier $\hat{f}_n(x)$ we have*

$$\sup_{P \in \mathcal{P}_U^*(\beta,\gamma)} \mathbb{E}R_P(\hat{f}_n) - R^* \geq CN^{-\frac{\beta(1+\gamma)}{2\beta+d-\beta\gamma}}.$$

Proof. We proceed by constructing the appropriate family of classifiers $f_\sigma(x) = \text{sign } \eta_\sigma(x)$, in a way similar to Theorem 3.5 in [4], and then apply Theorem 1.2.9 mentioned in Chapter 1.

Let $q = 2^l,\ l \geq 1$ and

$$G_q := \left\{ \left(\frac{2k_1 - 1}{2q}, \ldots, \frac{2k_d - 1}{2q} \right),\ k_i = 1 \ldots q,\ i = 1 \ldots d \right\}$$

be the grid on $[0, 1]^d$. For $x \in [0, 1]^d$, let

$$n_q(x) = \text{argmin } \{\|x - x_k\|_2 :\ x_k \in G_q\}.$$

If $n_q(x)$ is not unique, we choose a representative with the smallest $\|\cdot\|_2$ norm. The unit cube is partitioned with respect to G_q as follows: x_1, x_2 belong to the same subset if $n_q(x_1) = n_q(x_2)$. Let $' \succ'$ be some order on the elements of G_q such that $x \succ y$ implies $\|x\|_2 \geq \|y\|_2$. Assume that the elements of the partition are enumerated with respect to the order of their centers induced by $' \succ'$: $[0,1]^d = \bigcup_{i=1}^{q^d} R_i$. Fix $1 \leq m \leq q^d$ and let

$$S := \bigcup_{i=1}^{m} R_i$$

Note that the partition is ordered in such a way that there always exists $1 \leq k \leq q\sqrt{d}$ with

$$B_+\left(0, \frac{k}{q}\right) \subseteq S \subseteq B_+\left(0, \frac{k+3\sqrt{d}}{q}\right), \tag{2.6.1}$$

where $B_+(0,R) := \left\{x \in \mathbb{R}^d_+ : \ \|x\|_2 \leq R\right\}$. In other words, (2.6.1) means that that the difference between the radii of inscribed and circumscribed spherical sectors of S is of order $C(d)q^{-1}$.

Let $v > r_1 > r_2$ be three integers satisfying

$$2^{-v} < 2^{-r_1} < 2^{-r_1}\sqrt{d} < 2^{-r_2}\sqrt{d} < 2^{-1}. \tag{2.6.2}$$

Define $u(x) : \mathbb{R} \mapsto \mathbb{R}_+$ by

$$u(x) := \frac{\int_x^{\infty} U(t)dt}{\int_{2^{-v}}^{1/2} U(t)dt}, \tag{2.6.3}$$

where

$$U(t) := \begin{cases} \exp\left(-\frac{1}{(1/2-x)(x-2^{-v})}\right), & x \in (2^{-v}, \frac{1}{2}) \\ 0 & \text{else.} \end{cases}$$

Note that $u(x)$ is an infinitely differentiable function such that $u(x) = 1, \ x \in [0, 2^{-v}]$ and $u(x) = 0, \ x \geq \frac{1}{2}$. Finally, for $x \in \mathbb{R}^d$ let

$$\Phi(x) := Cu(\|x\|_2),$$

where $C := C_{L,\beta}$ is chosen such that $\Phi \in \Sigma(\beta, L, \mathbb{R}^d)$.

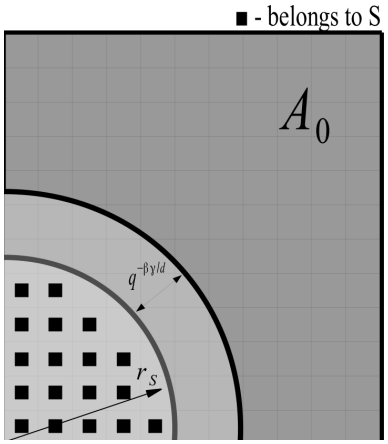

Figure 2: Geometry of the support

Let $r_S := \inf\{r > 0 : \ B_+(0,r) \supseteq S\}$ and

$$A_0 := \left\{ \bigcup_i R_i : \ R_i \cap B_+\left(0, r_S + q^{-\frac{\beta\gamma}{d}}\right) = \emptyset \right\}.$$

Note that

$$r_S \leq c\frac{m^{1/d}}{q}, \tag{2.6.4}$$

since $\mathrm{Vol}\,(S) = mq^{-d}$.

Define $\mathcal{H}_m = \{P_\sigma : \sigma \in \{-1,1\}^m\}$ to be the hypercube of probability distributions on $[0,1]^d \times \{-1,+1\}$. The marginal distribution Π of X is independent of σ: define its density p by

$$p(x) = \begin{cases} \frac{2^{d(r_1-1)}}{2^{d(r_1-r_2)}-1}, & x \in B_\infty\left(z, \frac{2^{-r_2}}{q}\right) \setminus B_\infty\left(z, \frac{2^{-r_1}}{q}\right), \ z \in G_q \cap S, \\ c_0, & x \in A_0, \\ 0 & \text{else.} \end{cases}$$

where $B_\infty(z,r) := \{x : \ \|x-z\|_\infty \leq r\}$, $c_0 := \frac{1-mq^{-d}}{\mathrm{Vol}(A_0)}$(note that $\Pi(R_i) = q^{-d} \quad \forall i \leq m$) and r_1, r_2 are defined in (2.6.2). In particular, Π satisfies *Assumption* 2.1 since it is supported on the union of dyadic cubes and has bounded above and below on $\mathrm{supp}(\Pi)$ density. Let

$$\Psi(x) := u\left(1/2 - q^{\frac{\beta\gamma}{d}}\mathrm{dist}_2(x, B_+(0,r_S))\right),$$

where $u(\cdot)$ is defined in (2.6.3) and $\mathrm{dist}_2(x, A) := \inf\{\|x - y\|_2,\ y \in A\}$.

Finally, the regression function $\eta_\sigma(x) = \mathbb{E}_{P_\sigma}(Y|X = x)$ is defined via

$$\eta_\sigma(x) := \begin{cases} \sigma_i q^{-\beta}\Phi(q[x - n_q(x)]), & x \in R_i,\ 1 \le i \le m \\ \frac{1}{C_{L,\beta}\sqrt{d}}\,\mathrm{dist}_2(x, B_+(0, r_S))^{\frac{d}{\gamma}} \cdot \Psi(x), & x \in [0,1]^d \setminus S. \end{cases}$$

The graph of η_σ is a surface consisting of small "bumps" spread around S and tending away from 0 monotonically with respect to $\mathrm{dist}_2(\cdot, B_+(0, r_S))$ on $[0,1]^d \setminus S$. Clearly, $\eta_\sigma(x)$ satisfies smoothness requirement, [1] since for $x \in [0,1]^d$

$$\mathrm{dist}_2(x, B_+(0, r_S)) = (\|x\|_2 - r_S) \vee 0.$$

Let's check that it also satisfies the low noise condition. Since $|\eta_\sigma| \ge Cq^{-\beta}$ on the support of Π, it is enough to consider $t = Czq^{-\beta}$ for $z > 1$:

$$\begin{aligned} \Pi(|\eta_\sigma(x)| \le Czq^{-\beta}) &\le mq^{-d} + \Pi\left(\mathrm{dist}_2(x, B_+(0, r_S)) \le Cz^{\gamma/d}q^{-\frac{\beta\gamma}{d}}\right) \le \\ &\le mq^{-d} + C_2\left(r_S + Cz^{\gamma/d}q^{-\frac{\beta\gamma}{d}}\right)^d \le \\ &\le mq^{-d} + C_3 mq^{-d} + C_4 z^\gamma q^{-\beta\gamma} \le \\ &\le \widehat{C}t^\gamma, \end{aligned}$$

if $mq^{-d} = O(q^{-\beta\gamma})$. Here, the first inequality follows from considering η_σ on S and A_0 separately, and second inequality follows from (2.6.4) and direct computation of the sphere volume.

Finally, η_σ satisfies *Assumption* 2.2 with some $B_2 := B_2(q)$ by Proposition 2.5.7 since η_σ is infinitely differentiable. The next step in the proof is to choose the subset of $\mathcal{H}$ which is "well-separated": this can be done due to Gilbert-Varshamov bound, see Proposition 1.2.10 in Chapter 1. Let $\mathcal{H}' := \{P_{\sigma_0}, \dots, P_{\sigma_M}\}$ be chosen such that $\{\sigma_0, \dots, \sigma_M\}$ satisfies the Proposition 1.2.10. Next, following the proof of Theorems 1 and 3 in [23], we note that $\forall \sigma \in \mathcal{H}',\ \sigma \ne \sigma_0$

$$\mathrm{KL}(P_{\sigma,N}\|P_{\sigma_0,N}) \le 8N \max_{x\in[0,1]}(\eta_\sigma(x) - \eta_{\sigma_0}(x))^2 \le 32C_{L,\beta}^2 N q^{-2\beta}, \tag{2.6.5}$$

[1] $\Psi(x)$ is introduced to provide extra smoothness at the boundary of $B_+(0, r_S)$.

where $P_{\sigma,N}$ is the joint distribution of $(X_i, Y_i)_{i=1}^N$ under hypothesis that the distribution of couple (X, Y) is P_σ. Let us briefly sketch the derivation of (2.6.5); see also the proof of Theorem 1 in [23]. Denote

$$\bar{X}_k := (X_1, \dots, X_k),$$
$$\bar{Y}_k := (Y_1, \dots, Y_k).$$

Then $dP_{\sigma,N}$ admits the following factorization:

$$dP_{\sigma,N}(\bar{X}_N, \bar{Y}_N) = \prod_{i=1}^N P_\sigma(Y_i|X_i) dP(X_i|\bar{X}_{i-1}, \bar{Y}_{i-1}),$$

where $dP(X_i|\bar{X}_{i-1}, \bar{Y}_{i-1})$ does not depend on σ but only on the active learning algorithm. As a consequence,

$$\begin{aligned}
\mathrm{KL}(P_{\sigma,N}\|P_{\sigma_0,N}) = \mathbb{E}_{P_{\sigma,N}} \log \frac{dP_{\sigma,N}(\bar{X}_N, \bar{Y}_N)}{dP_{\sigma_0,N}(\bar{X}_n, \bar{Y}_N)} = \mathbb{E}_{P_{\sigma,N}} \log \frac{\prod_{i=1}^N P_\sigma(Y_i|X_i)}{\prod_{i=1}^N P_{\sigma_0}(Y_i|X_i)} = \\
= \sum_{i=1}^N \mathbb{E}_{P_{\sigma,N}} \left[\mathbb{E}_{P_\sigma} \left(\log \frac{P_\sigma(Y_i|X_i)}{P_{\sigma_0}(Y_i|X_i)} \big| X_i \right) \right] \leq \\
\leq N \max_{x \in [0,1]^d} \mathbb{E}_{P_\sigma} \left(\log \frac{P_\sigma(Y_1|X_1)}{P_{\sigma_0}(Y_1|X_1)} \big| X_1 = x \right) \leq \\
\leq 8N \max_{x \in [0,1]^d} (\eta_\sigma(x) - \eta_{\sigma_0}(x))^2,
\end{aligned}$$

where the last inequality follows from Lemma 1 in [23]. Also, note that we have $\max_{x\in[0,1]^d}$ in our bounds rather than the average over x that would appear in the passive learning framework.

It remains to choose q, m in appropriate way: set $q = \lfloor C_1 N^{\frac{1}{2\beta+d-\beta\gamma}} \rfloor$ and $m = \lfloor C_2 q^{d-\beta\gamma} \rfloor$ where $C_1,\ C_2$ are such that $q^d \geq m \geq 1$ and $32 C_{L,\beta}^2 N q^{-2\beta} < \frac{m}{64}$ which is possible for N big enough. In particular, $mq^{-d} = O(q^{-\beta\gamma})$. Together with the bound (2.6.5), this gives

$$\frac{1}{M} \sum_{\sigma \in \mathcal{H}'} \mathrm{KL}(P_\sigma \| P_{\sigma^0}) \leq 32 C_{L,\beta}^2 N q^{-2\beta} < \frac{m}{8^2} = \frac{1}{8} \log |\mathcal{H}'|,$$

so that conditions of Theorem 1.2.9 are satisfied. Setting

$$f_\sigma(x) := \operatorname{sign} \eta_\sigma(x),$$

we finally have $\forall \sigma_1 \neq \sigma_2 \in \mathcal{H}'$

$$d(f_{\sigma_1}, f_{\sigma_2}) := \Pi(\text{sign } \eta_{\sigma_1}(x) \neq \text{sign } \eta_{\sigma_2}(x)) \geq \frac{m}{8q^d} \geq C_4 N^{-\frac{\beta\gamma}{2\beta+d-\beta\gamma}},$$

where the lower bound just follows by construction of our hypotheses. Since under the low noise assumption $R_P(\hat{f}_n) - R^* \geq c\Pi(\hat{f}_n \neq \text{sign } \eta)^{\frac{1+\gamma}{\gamma}}$ (see (2.3.4)), we conclude by Theorem 1.2.9 that

$$\inf_{\hat{f}_N} \sup_{P \in \mathcal{P}_U^*(\beta,\gamma)} \Pr\left(R_P(\hat{f}_n) - R^* \geq C_4 N^{-\frac{\beta(1+\gamma)}{2\beta+d-\beta\gamma}}\right) \geq$$
$$\geq \inf_{\hat{f}_N} \sup_{P \in \mathcal{P}_U^*(\beta,\gamma)} \Pr\left(\Pi(\hat{f}_n(x) \neq \text{sign } \eta_P(x)) \geq \frac{C_4}{2} N^{-\frac{\beta\gamma}{2\beta+d-\beta\gamma}}\right) \geq \tau > 0.$$

□

2.6.2 Upper bounds for the excess risk

In this subsection we continue the discussion of an active learning algorithm introduced earlier. We will present detailed analysis and provide tight probabilistic bounds for its performance. In particular, we show that the classifier constructed by the algorithm attains the rates of Theorem 2.6.1, up to polylogarithmic factor, if $0 < \beta \leq r + 1$. Moreover, the algorithm is adaptive with respect to β, γ and is computationally tractable. The analysis builds upon our previous work [73] where only the case of piecewise-constant estimators was treated. However, this limited the adaptation range to $0 < \beta \leq 1$. It turns out that the general case involving piecewise–polynomial estimators of higher degrees is more difficult in several aspects. In particular, the algorithm itself requires some changes, and the associated analysis is performed under slightly more restrictive assumptions. The main difference is related to the geometry of the active set (see Section 2.4 for definitions). For piecewise–constant estimators, the active set is always given by a union of dyadic cubes of fixed size – same as the domain of an estimator. The level sets of polynomials of higher degree have more complicated structure, so we need to approximate them by certain

'regular' sets which allows the algorithm to maintain suitable structure for the subsequent iterations.

This section is organized as follows: first, the basic case of piecewise-constant estimators is presented($r = 0$). The rigorous description of the learning procedure for this case is summarized in Algorithm 1. The main facts about the performance of the method are formulated in Theorem 2.6.2. Then we proceed with a general case $r \geq 1$ and, building upon the foundations of Theorem 2.6.2, prove our main result, Theorem 2.6.3. In the end, we provide the running time analysis and computer simulation results.

2.6.3 Learning with piecewise-constant functions

It would be convenient for us to define the stopping rule for Algorithm 1 in terms of the *label threshold* $\bar{N}$. The algorithm stops after completing the first iteration on which the total number of used labels exceeds $\bar{N}$ (so that $\bar{N}$ gives only approximate bound for the total number of requested labels). Another (and, sometimes, more convenient) way to define the stopping rule is to do it in term of the prescribed confidence level and excess risk that the resulting classifier should attain. This requires an additional validation subroutine that is able to keep track of the excess risk by estimating it from the data, but we are currently not aware of the method allowing to implement such procedure into our algorithm. This is a minor disadvantage of our approach relative to the popular empirical risk minimization techniques where the excess risk can often be adaptively estimated from the data.

In what follows, $\bar{N}$ will stand for the label threshold and N – for the total number of used labels. Recall that, given a set A and $g \in \mathcal{F}_m^0$ (the space of piecewise-constant functions on the dyadic partition $\mathcal{H}_{2^m}$ of $[0,1]^d$),

$$\mathcal{F}_{\infty,A}(g;\delta) := \left\{ f \in \mathcal{F}_m^0 : \ \|f - g\|_{\infty,A} \leq \delta \right\}.$$

We briefly go over the main stages of Algorithm 1: first, a small part of the label

Algorithm 1: Active Learning Algorithm, $r = 0$.

input : label threshold $\bar{N}$; confidence α; minimal regularity $0 < \nu < 1$

output: $\hat{g} := \operatorname{sign} \hat{\eta}$

$\omega := 2 + \frac{d}{2\nu}$;

$k = 0,\ \hat{A}_0 := [0,1]^d$;

$N_0 := \lfloor \sqrt{\bar{N}} \rfloor$;

$LB := \bar{N} - 2N_0$;

for $i = 1$ **to** $2N_0$ **do**

  sample i.i.d. $\left(X_i^{(0)}, Y_i^{(0)}\right)$ with $X_i^{(0)} \sim \Pi$;

$S_{0,1} := \left\{ \left(X_i^{(0)}, Y_i^{(0)}\right),\ i \le N_0 \right\}, \quad S_{0,2} = \left\{ \left(X_i^{(0)}, Y_i^{(0)}\right),\ N_0 + 1 \le i \le 2N_0 \right\}$;

$\hat{m}_0 := \hat{m}(s, N_0; S_{0,1})$ `/* see equation (2.5.4) in Section 2.5.3 */`;

$\hat{\eta}_0 := \hat{\eta}_{\hat{m}_0, [0,1]^d; S_{0,2}}$ `/* see equation (2.5.2) in Section 2.5.2 */`;

while $LB > 0$ **do**

  $\hat{\mathcal{F}}_k := \left\{ f \in \mathcal{F}^0_{\hat{m}_k} :\ f|_{\hat{A}_k} \in \mathcal{F}_{\infty, \hat{A}_k}(\hat{\eta}_k; \delta_k),\quad f|_{[0,1]^d \setminus \hat{A}_k} \equiv \hat{\eta}_{k-1}|_{[0,1]^d \setminus \hat{A}_k} \right\}$

  `/* confidence band around` $\hat{\eta}_k$ `*/`;

  $k := k + 1$;

  $\hat{A}_k := \left\{ x \in [0,1]^d :\ \exists f_1, f_2 \in \hat{\mathcal{F}}_{k-1}, \operatorname{sign}\left(f_1(x)\right) \ne \operatorname{sign}\left(f_2(x)\right) \right\}$;

  if $\hat{A}_k \cap \operatorname{supp}(\Pi) = \emptyset$ **then**

    break

  else

    $\hat{m}_k := \hat{m}_{k-1} + 1$;

    $\tau_k := \frac{\hat{m}_k}{\hat{m}_{k-1}}$;

    $N_k := \lfloor N_{k-1}^{\tau_k} \rfloor$;

    for $i = 1$ **to** $\lfloor N_k \cdot \Pi(\hat{A}_k) \rfloor$ **do**

      sample i.i.d. $\left(X_i^{(k)}, Y_i^{(k)}\right)$ with $X_i^{(k)} \sim \hat{\Pi}_k := \Pi(dx | x \in \hat{A}_k)$;

      $S_k := \left\{ \left(X_i^{(k)}, Y_i^{(k)}\right),\ i \le \lfloor N_k \cdot \Pi(\hat{A}_k) \rfloor \right\}$;

    $\hat{\eta}_k := \hat{\eta}_{\hat{m}_k, \hat{A}_k}$ `/* estimator based on` S_k `*/`;

    $\delta_k := \tilde{D} (\log \frac{\bar{N}}{\alpha})^{\omega \frac{\hat{m}_k}{\hat{m}_0}} \cdot \sqrt{\frac{2^{d\hat{m}_k}}{N_k}}$ `/* size of the confidence band */`;

    $LB := LB - \lfloor N_k \cdot \Pi(\hat{A}_k) \rfloor$;

    $\hat{\eta} := \hat{\eta}_k$ `/* keeping track of the most recent estimator */`;

budget of cardinality $2N_0$ is used to select the optimal resolution level and to construct a preliminary estimator $\hat{\eta}_0$: the first N_0 pairs (X_i, Y_i) denoted by $S_{0,1}$ are used to select $\hat{m}_0$, and the rest (denoted $S_{0,2}$) are used to construct $\hat{\eta}_0$. The iterative part of the algorithm works as follows: based on the current estimator $\hat{\eta}_k$, the confidence band $\hat{\mathcal{F}}_k$ is constructed. In turn, $\hat{\mathcal{F}}_k$ is used to obtain the active set $\hat{A}_k$ that serves as a support for the updated design distribution $\hat{\Pi}_k$. The cardinality of a new sample from $\hat{\Pi}_k$ guarantees that the local amount of data increases on every step, allowing tighter concentration in sup-norm. N_k is chosen based on the requirement that on every step, $2^{\hat{m}_k} \approx N_k^{1/(2\beta+d)}$(this motivates our definition of τ_k). After the label threshold is exceeded, algorithm outputs the sign of the most recent estimator as a result.
For convenience, we summarize our main assumptions before stating the theorems.

(i) $P \in \mathcal{P}_U^*(\beta, \gamma)$, meaning that the low noise assumption (see (2.2.1)) is satisfied with exponent γ and for all $m \geq 1$

$$B_2 2^{-\beta m} \leq \|\eta - \bar{\eta}_m\|_{\infty, \mathrm{supp}(\Pi)} \leq B_1 2^{-\beta m}. \tag{2.6.6}$$

(ii) $B_1 \leq \sqrt{\log \bar{N}}, \quad B_2 \geq \frac{1}{\sqrt{\log \bar{N}}}$, where $\bar{N}$ is the label threshold. This assumption allows one to construct explicit non-asymptotic confidence bands, and it can be replaced by any other known bounds on B_1, B_2 (note that instead of $\sqrt{\cdot}$ we could use any other power function which would affect only the logarithmic factor in the resulting bounds).

(iii) If A is such that $\alpha = \bar{N}^{-A}$ and $0 < \nu \leq 1$, then

$$\bar{N} \geq C(\nu, d) \cdot \left(\log \bar{N} \vee A\right)^{5(2\nu+d)/\nu}.$$

This condition comes from Theorem 2.5.6, and the number ν essentially determines the range of uniformity of our results.

(iv) This last assumption is only used in Theorem 2.6.3: there exist $K_1, K_2\gamma > 0$ such that

$$\forall\ t > 0,\ K_2 t^\gamma \le \Pi(x : |\eta(x)| \le t) \le K_1 t^\gamma,$$

where $\frac{K_1}{K_2} \le \log \bar{N}$ ($\log \bar{N}$ can be replaced by any polynomial in $\log \bar{N}$ with corresponding changes in the resulting log-factor). This condition is a more restrictive version of the low noise assumption and is similar to condition used in a work of R. Nowak and R. Castro [23]. Note that in the simplest case of piecewise constant estimators $r = 0$ the lower bound is not necessary.

Below, we will mainly concentrate on the hardest case when the graph of the regression function hits or crosses the decision boundary $\{x : \eta(x) = 0\}$ in the interior or boundary of supp(Π). In terms of the parameters of the distribution, this case can be characterized by a condition $(\beta \wedge 1)\gamma \le d$ (see Proposition 3.4 in [5] for details). The easier case when $|\eta(x)|$ is bounded away from 0 on supp(Π) (often called "the bounded noise condition") can be handled similarly, and under our assumptions the Bayes classifier $f_* = \operatorname{sign} \eta$ can be learned with high probability in finitely many steps.

Theorem 2.6.2. *If the aforementioned assumptions (i-iii) are satisfied, then the following holds uniformly over all $0 < \nu \le \beta \le 1$ and $\gamma > 0$: with probability at least $1-\alpha$, the classifier $\hat{g}$ returned by* **Algorithm 1** *with label threshold $\bar{N}$ and confidence α, satisfies*

$$R_P(\hat{g}) - R^* \le C \cdot N^{-\frac{\beta(1+\gamma)}{2\beta+d-\beta\gamma}} \log^p \frac{\bar{N}}{\alpha},$$

where $p \le \left(\frac{4+2d}{\nu}\right)^2 (1+\gamma)\left(1 + \frac{\beta\gamma}{2\beta+d-\beta\gamma}\right)$ and N is the label complexity - the total number of requested labels.

Remarks:

1. Note that when $\beta\gamma > \frac{d}{3}$, $N^{-\frac{\beta(1+\gamma)}{2\beta+d-\beta\gamma}}$ is a *fast rate*, that is, faster than $N^{-\frac{1}{2}}$; at

the same time, the passive learning rate $N^{-\frac{\beta(1+\gamma)}{2\beta+d}}$ is guaranteed to be fast only when $\beta\gamma > \frac{d}{2}$, see [4] (we assume here that $0 < \beta \leq 1$).

2. For $\hat{\alpha} \simeq N^{-\frac{\beta(1+\gamma)}{2\beta+d-\beta\gamma}}$ **Algorithm 1** returns a classifier $\hat{g}_{\hat{\alpha}}$ that satisfies the following *average excess risk* bound:

$$\mathbb{E}R_P(\hat{g}_{\hat{\alpha}}) - R^* \leq \text{Const} \cdot N^{-\frac{\beta(1+\gamma)}{2\beta+d-\beta\gamma}} \log^p N.$$

This is a direct corollary of Theorem 2.6.2 and the inequality

$$\mathbb{E}|Z| \leq t + \|Z\|_\infty \Pr(|Z| \geq t).$$

Proof. Our main goal will be to construct high probability bounds for the size of the sets $\hat{A}_k$ defined by Algorithm 1. In turn, these bounds depend on the size of the confidence bands for $\eta(x)$ (denoted by δ_k). Suppose L is the number of steps performed by the algorithm before termination.

Let $N_k^{\text{act}} := \lfloor N_k \cdot \Pi(\hat{A}_k) \rfloor$ be the number of labels requested on k-th iteration of the algorithm. Claim: the following bounds hold uniformly for all $1 \leq k \leq L$ with probability at least $1 - \alpha$:

$$\begin{aligned} \|\eta - \hat{\eta}_k\|_{\infty, \hat{A}_k} &\leq C \left(\log \frac{\bar{N}}{\alpha} \right)^{\omega \frac{\hat{m}_k}{\hat{m}_0}} \cdot N_k^{-\beta/(2\beta+d)} \\ \Pi(\hat{A}_k) &\leq C \left(\log \frac{\bar{N}}{\alpha} \right)^{\gamma\omega\bar{\tau}} \cdot N_{k-1}^{-\beta\gamma/(2\beta+d)} \end{aligned} \tag{2.6.7}$$

where $\omega = 2 + \frac{d}{2\nu}$ and $\bar{\tau} = 4 + \frac{2d}{\nu}$.

Let us first assume that (2.6.7) has already been established and derive the result from it. Let $\bar{m}_0$ be the "optimal" resolution level for the corresponding sample of size N_0, see formula (2.5.5) and line 3 of Algorithm 1. First, we make a useful observation that, with high probability, numbers N_k grow geometrically: indeed, we have by the definition of $\hat{m}_k$

$$N_{k+1} = \lfloor N_k^{\hat{m}_{k+1}/\hat{m}_k} \rfloor \leq N_k \cdot N_k^{1/\hat{m}_k} \leq N_k \cdot \left(N_0^{\frac{\hat{m}_k}{\bar{m}_0}} \right)^{\frac{1}{\hat{m}_k}} = N_k \cdot N_0^{\frac{1}{\bar{m}_0}},$$

and by Theorem 2.5.6 (see the remark following the main statement), if N_0 is sufficiently large, as guaranteed by our assumptions,

$$\frac{\log_2 N_0}{\bar{m}_0} \le \log_2 N_0^{1/\hat{m}_0} \le \frac{\log_2 N_0}{\frac{1}{2}\bar{m}_0}$$

with probability $\ge 1-\alpha$. Finally, Lemma 2.5.5 gives

$$\frac{1}{2\beta+d}\log N_0 + c \ge \bar{m}_0 \ge \frac{1}{2\beta+d}(\log N_0 - 2\log\log N_0) - c$$

which shows that $0 < C_1 \le \log_2 N_0^{1/\hat{m}_0} \le C_2$.

Next, inequality (2.6.7) implies, together with the previous observation, that the number of labels requested on step $k \ge 1$ satisfies

$$N_k^{\text{act}} = \lfloor N_k \Pi(\hat{A}_k) \rfloor \le C \cdot N_{k-1}^{\frac{2\beta+d-\beta\gamma}{2\beta+d}} \left(\log \frac{\bar{N}}{\alpha}\right)^{\gamma\omega\bar{\tau}}$$

with probability $\ge 1-2\alpha$. If N is the total number of labels requested by the Algorithm, then

$$N = \sum_{k=0}^{L} N_k^{\text{act}} \le C_3 \left(\log \frac{\bar{N}}{\alpha}\right)^{\gamma\omega\bar{\tau}} \sum_{k=0}^{L} N_k^{\frac{2\beta+d-\beta\gamma}{2\beta+d}} \le C_4 \left(\log \frac{\bar{N}}{\alpha}\right)^{\gamma\omega\bar{\tau}} N_L^{\frac{2\beta+d-\beta\gamma}{2\beta+d}},$$

one easily deduces that on the last iteration L we have

$$N_L \ge c(\nu,\gamma,\Pi,d)\left(\frac{N}{\log^{\gamma\omega\bar{\tau}}(\bar{N}/\alpha)}\right)^{\frac{2\beta+d}{2\beta+d-\beta\gamma}} \tag{2.6.8}$$

To obtain the risk bound of the theorem from (2.6.8), we apply inequality (2.3.2) from Proposition 2.3.1:

$$R_P(\hat{g}) - R^* \le D_1 \|(\hat{\eta}_L - \eta)\cdot \mathcal{I}\{\text{sign } \hat{\eta}_L \ne \text{sign } \eta\}\|_\infty^{1+\gamma}. \tag{2.6.9}$$

Since $\{\text{sign } \hat{\eta}_L \ne \text{sign } \eta\} \cap \text{supp}(\Pi) \subseteq \hat{A}_L$ whenever the bounds (2.6.7) hold, it remains to estimate $\|\hat{\eta}_L - \eta\|_{\infty,\hat{A}_L}$. Recalling the first inequality of (2.6.7) once again (for $k = L$), we get

$$\|(\hat{\eta}_L - \eta)\cdot \mathcal{I}\{\text{sign } \hat{\eta}_L \ne \text{sign } \eta\}\|_\infty \le C\left(\log\frac{\bar{N}}{\alpha}\right)^{\omega\bar{\tau}} \left(\frac{N}{\log^{\gamma\omega\bar{\tau}}(\bar{N}/\alpha)}\right)^{-\frac{\beta}{(2\beta+d-\beta\gamma)}}$$

which together with (2.6.9) yields the final result, after some simple algebra to estimate the power of the logarithm.

It remains to show (2.6.7). The main tools are given by Theorem 2.5.4 and Theorem 2.5.6. Let $\hat{\eta}_k$ be the estimator obtained on step k. For $k = 0$, we have

$$\|\eta - \hat{\eta}_0\|_{\infty,\mathrm{supp}(\Pi)} \le \|\eta - \bar{\eta}_{\hat{m}_0}\|_{\infty,\mathrm{supp}(\Pi)} + \|\bar{\eta}_{\hat{m}_0} - \hat{\eta}_0\|_{\infty,\mathrm{supp}(\Pi)}.$$

By Theorem 2.5.4 (applied conditionally on $S_{0,1}$, see a remark after the theorem for details), with probability $\ge 1 - \alpha$

$$\|\bar{\eta}_{\hat{m}_0} - \hat{\eta}_0\|_{\infty,\mathrm{supp}(\Pi)} \le C \log(\bar{N}/\alpha) \sqrt{\frac{2^{d\hat{m}_0}}{N_0}}.$$

For the bias term $\|\eta - \bar{\eta}_{\hat{m}_0}\|_{\infty,\mathrm{supp}(\Pi)}$, there are two possibilities: if the first condition of Assumption 2.2 is satisfied, then the bias is 0 for $m \ge m_0$, and we are only left to control the random error as above. Otherwise, by our assumptions on η (see Corollary 2.5.2),

$$\|\eta - \bar{\eta}_{\hat{m}_0}\|_{\infty,\mathrm{supp}(\Pi)} \le B_1 2^{-\beta\hat{m}_0}.$$

By a remark (2.5.11) after Theorem 2.5.6, with probability $\ge 1 - \alpha$

$$2^{-\beta\hat{m}_0} \le \frac{C}{\beta} \left(\log \frac{\bar{N}}{\alpha}\right)^{1+\frac{d}{2\beta}} \sqrt{\frac{2^{d\hat{m}_0}\hat{m}_0}{N_0}}, \tag{2.6.10}$$

and by Theorem 2.5.6 and Lemma 2.5.5, with probability $\ge 1 - \alpha$

$$\frac{2^{d\hat{m}_0}}{N_0} \le \frac{2^{d\bar{m}_0}}{N_0} \le C_1 N_0^{-2\beta/(2\beta+d)}, \tag{2.6.11}$$

so that, with probability $\ge 1 - 2\alpha$,

$$\begin{aligned} \|\eta - \hat{\eta}_0\|_{\infty,\mathrm{supp}(\Pi)} &\le C(\beta,\Pi) \left(\log \frac{\bar{N}}{\alpha}\right)^{3/2+\frac{d}{2\beta}} \sqrt{\frac{2^{d\hat{m}_0}\hat{m}_0}{N_0}} := \frac{\delta_0}{2} \le \\ &\le C(\beta,\Pi) \left(\log \frac{\bar{N}}{\alpha}\right)^{2+\frac{d}{2\beta}} N_0^{-\frac{\beta}{2\beta+d}}. \end{aligned} \tag{2.6.12}$$

For $k \ge 1$, we have in a similar way

$$\|\eta - \hat{\eta}_k\|_{\infty,\hat{A}_k} \le \|\eta - \bar{\eta}_{\hat{m}_k}\|_{\infty,\hat{A}_k} + \|\bar{\eta}_{\hat{m}_k} - \hat{\eta}_k\|_{\infty,\hat{A}_k}.$$

By (2.6.11) and Theorem 2.5.4 applied for $A := \hat{A}_k$ and $N := N_k^{\text{act}}$ (conditionally on $\bigcup_{i=0}^{k-1} S_k$), with probability $\geq 1-\alpha$

$$\|\bar{\eta}_{\hat{m}_k} - \hat{\eta}_k\|_{\infty,\hat{A}_k} \leq C\log(\bar{N}/\alpha)\sqrt{\frac{2^{d\hat{m}_k}}{N_k}} \leq C\log(\bar{N}/\alpha)\left(\sqrt{\frac{2^{d\hat{m}_0}}{N_0}}\right)^{\prod_{i=1}^k \tau_k} \leq$$

$$\leq C\log(\bar{N}/\alpha)N_k^{-\beta/(2\beta+d)}. \tag{2.6.13}$$

Once again, for the bias term, we only need to consider the case when the second condition of Assumption 2.2 is satisfied. By (2.6.10), we have

$$\begin{aligned}
\|\eta - \bar{\eta}_{\hat{m}_k}\|_{\infty,\hat{A}_k} &\leq CB_1 2^{-\beta\hat{m}_k} = CB_1\left(2^{-\beta\hat{m}_0}\right)^{\prod_{i=1}^k \tau_i} \leq \\
&\leq C(\beta,\Pi)B_1\left[\left(\log\frac{\bar{N}}{\alpha}\right)^{1+\frac{d}{2\beta}}\sqrt{\frac{2^{d\hat{m}_0}\hat{m}_0}{N_0}}\right]^{\prod_{i=1}^k \tau_i} \leq \\
&\leq C(\nu,\Pi)\left[\left(\log\frac{\bar{N}}{\alpha}\right)^{2+\frac{d}{2\nu}}\right]^{\prod_{i=1}^k \tau_i}\sqrt{\frac{2^{d\hat{m}_k}}{N_k}} := \frac{\delta_k}{2} \leq \\
&\leq C(\nu,\Pi)\left[\left(\log\frac{\bar{N}}{\alpha}\right)^{2+\frac{d}{2\nu}}\right]^{\prod_{i=1}^k \tau_i} N_k^{-\beta/(2\beta+d)},
\end{aligned} \tag{2.6.14}$$

which holds with probability $\geq 1-\alpha$ and gives together with (2.6.13) that

$$\|\eta - \hat{\eta}_k\|_{\infty,\hat{A}_k} \leq \frac{\delta_k}{2} \leq C(\nu,\Pi)\left[\left(\log\frac{\bar{N}}{\alpha}\right)^{2+\frac{d}{2\nu}}\right]^{\prod_{i=1}^k \tau_i} N_k^{-\beta/(2\beta+d)} \tag{2.6.15}$$

with probability $\geq 1-2\alpha$. Finally, it remains to note that for all $1 \leq k \leq L$, with probability $\geq 1-2\alpha$,

$$\prod_{i=1}^k \tau_i \leq \prod_{i=1}^L \tau_i \leq 2\frac{2\nu+d}{\nu} := \bar{\tau}. \tag{2.6.16}$$

Indeed, on each iteration, the set $\hat{A}_k$ contains at least one dyadic cube with edge length $2^{-\hat{m}_{k-1}}$, and by our assumption on Π, Theorem 2.5.6 and Lemma 2.5.5,

$$\begin{aligned}
\Pi(\hat{A}_k) \geq u_1 2^{-d\hat{m}_{k-1}} &\geq u_1\left(2^{-d\hat{m}_0}\right)^{\hat{m}_{k-1}/\hat{m}_0} \geq u_1\left(2^{-d\bar{m}_0}\right)^{\hat{m}_{k-1}/\hat{m}_0} \geq \\
&\geq C\left(N_0^{-d/(2\beta+d)}\right)^{\hat{m}_{k-1}/\hat{m}_0} \geq CN_k^{-d/(2\beta+d)}.
\end{aligned}$$

Also, because of the geometric growth of N_k, the number of labels requested on a last step cannot exceed $C\bar{N}$, in other words,

$$C\bar{N} \geq \lfloor N_L \Pi(A_L) \rfloor \geq C_1 N_L^{2\beta/(2\beta+d)} \geq C_2 \left(N_0^{2\beta/(2\beta+d)} \right)^{\prod_{i=1}^{L} \tau_i} \geq C_2 \left(N_0^{2\nu/(2\nu+d)} \right)^{\prod_{i=1}^{L} \tau_i},$$

and, since $N_0 = \lfloor \sqrt{\bar{N}} \rfloor$, this implies (2.6.16).

The union bound over all $0 \leq k \leq L$ gives that, with probability $\geq 1 - 4(L+1)\alpha$, on every iteration we have

$$\|\eta - \hat{\eta}_k\|_{\infty, \hat{A}_k} \leq \frac{\delta_k}{2} \leq \bar{C}(\nu, \Pi) \left(\log \frac{\bar{N}}{\alpha} \right)^{\bar{\tau}(2+\frac{d}{2\nu})} N_k^{-\beta/(2\beta+d)}, \tag{2.6.17}$$

where we used that $\prod_{i=1}^{k} \tau_i \leq \bar{\tau}$ for any k (on this event). With $\tilde{D}$ from the definition of δ_k in line 24, Algorithm 1, satisfying $\tilde{D} = 2\bar{C}(\nu, \Pi)$ (where $\bar{C}(\nu, \Pi)$ is the constant from (2.6.17)), it can be easily seen that the necessary condition for $x \in \operatorname{supp}(\Pi) \cap \hat{A}_k$ is

$$|\eta(x)| \leq 3\bar{C} \cdot \left(\log \frac{\bar{N}}{\alpha} \right)^{\bar{\tau}(2+\frac{d}{2\nu})} N_{k-1}^{-\beta/(2\beta+d)}.$$

Indeed, by triangle inequality,

$$x \in \operatorname{supp}(\Pi) \cap \hat{A}_k \implies |\hat{\eta}_k(x)| \leq \delta_k \implies$$
$$|\eta(x)| \leq |\hat{\eta}_k(x)| + |\eta(x) - \hat{\eta}_k(x)| \leq \frac{3}{2}\delta_k.$$

This gives, by the low noise assumption,

$$\Pi(\hat{A}_k) = \Pi(\hat{A}_k \cap \operatorname{supp}(\Pi)) \leq \Pi \left(|\eta(x)| \leq 3\bar{C} \cdot \left(\log \frac{\bar{N}}{\alpha} \right)^{\bar{\tau}(2+\frac{d}{2\nu})} N_{k-1}^{-\beta/(2\beta+d)} \right) \leq$$
$$\leq K \log(\bar{N}/\alpha)^{\gamma\bar{\tau}(2+\frac{d}{2\nu})} \cdot N_{k-1}^{-\beta\gamma/(2\beta+d)}$$

for every $1 \leq k \leq L$, with probability $\geq 1 - 4L\alpha$, hence proving the claim (since the number of steps L is surely bounded by the label threshold $\bar{N}$, the confidence can be raised to $1 - \alpha$ if we use $\frac{\alpha}{\bar{N}}$ in place of α, which only affects the constants). $\square$

2.6.4 Learning with piecewise-polynomial functions

Finally, we present and analyze the learning algorithm in the case when more complex estimators ($r \geq 1$) are used. The main difference with the case $r = 0$ is that for 'highly regular' functions (i.e., with $\beta > 1$), the size of the confidence bands decays faster resulting in a smaller active set. While this is a positive observation on the one hand, we encounter some technical difficulties arising from the fact that the natural 'resolution level' for the active set might be much smaller than the 'resolution level' of the corresponding estimator. One might think of a set having a small measure but many connected components, and the problem of constructing an estimator for the regression function supported on such a set and attaining small sup-norm error is difficult.

Recall that $\mathcal{B}_m$ is the sigma-algebra generated by the dyadic cubes $R_j,\ 1 \leq j \leq 2^{dm}$ forming the partition of $[0,1]^d$. We briefly mention the main differences between Algorithm 2 and Algorithm 1. As we have already observed, the "true" active set (denoted Act_k below) associated to the confidence band can be quit hard to work with, so instead the algorithm constructs its approximation by a union of dyadic cubes of suitable size, denoted $\hat{A}_k$, which is at most $C \log \bar{N}$ times larger (with respect to Π). This allows to maintain the structure suitable for the iterative nature of our method.

Theorem 2.6.3. *If the aforementioned assumptions (i-iv) are satisfied and* $(\beta \wedge 1)\gamma \leq d$*, then the following holds uniformly over all* $0 < \nu \leq \beta \leq r+1$ *and* $\gamma > 0$*: with probability at least* $1-\alpha$*, the classifier* $\hat{g}$ *returned by* **Algorithm 2** *with label threshold* $\bar{N}$ *and confidence* α*, satisfies*

$$R_P(\hat{g}) - R^* \leq C \cdot N^{-\frac{\beta(1+\gamma)}{2\beta+d-(\beta\wedge 1)\gamma}} \log^p \frac{\bar{N}}{\alpha},$$

where $p \leq \left(\frac{4+2d}{\nu} \vee (r+1)(2(r+1)+d)\right)^2 (1+\gamma)\left(1+\frac{\beta\gamma}{2\beta+d-\beta\gamma}\right)$ *and* N *is the total number of label requests.*

Algorithm 2: Active Learning Algorithm, $r \geq 1$.

input : label threshold $\bar{N}$; confidence α; minimal regularity $0 < \nu < 1$

output: $\hat{g} := \operatorname{sign} \hat{\eta}$

$\omega := 2 + \frac{d}{2\nu}$;

$k = 0,\ \tau_0 = 1,\ \hat{A}_0 := [0,1]^d$;

$N_0 := \lfloor \bar{N}^{1/(2(r+1))} \rfloor$;

$LB := \bar{N} - 2N_0$;

for $i = 1$ **to** $2N_0$ **do**

  sample i.i.d. $\left(X_i^{(0)}, Y_i^{(0)}\right)$ with $X_i^{(0)} \sim \Pi$;

$S_{0,1} := \left\{ \left(X_i^{(0)}, Y_i^{(0)}\right),\ i \leq N_0 \right\}, \quad S_{0,2} = \left\{ \left(X_i^{(0)}, Y_i^{(0)}\right),\ N_0 + 1 \leq i \leq 2N_0 \right\}$;

$\hat{m}_0 := \hat{m}(s, N_0; S_{0,1})$ `/* see equation (2.5.4) in Section 2.5.3 */`;

$\hat{\eta}_0 := \hat{\eta}_{\hat{m}_0,[0,1]^d;S_{0,2}}$ `/* see equation (2.5.2) in Section 2.5.2 */`;

while $LB > 0$ **do**

  $\delta_k := \tilde{D} \left(\log \frac{\bar{N}}{\alpha} \right)^{\omega} \prod_{i=0}^{k} \tau_i \sqrt{\frac{2^{d\hat{m}_k}}{N_k}}$ `/* size of the confidence band */`;

  $\hat{\mathcal{F}}_k := \left\{ f \in \mathcal{F}^r_{\hat{m}_k} :\ f|_{\hat{A}_k} \in \mathcal{F}_{\infty,\hat{A}_k}(\hat{\eta}_k; \delta_k),\quad f|_{[0,1]^d \setminus \hat{A}_k} \equiv \hat{\eta}_{k-1}|_{[0,1]^d \setminus \hat{A}_k} \right\}$;

  $k := k + 1$;

  $\mathrm{Act}_k := \left\{ x \in [0,1]^d :\ \exists f_1, f_2 \in \hat{\mathcal{F}}_{k-1}, \operatorname{sign}\left(f_1(x)\right) \neq \operatorname{sign}\left(f_2(x)\right) \right\}$ `/* the ''true" active set */`;

  if $\mathrm{Act}_k \cap \operatorname{supp}(\Pi) = \emptyset$ **then**

    break

  else

    $\hat{m}_k :=$

    $\min \left\{ m \geq [\tau_{k-1}\hat{m}_{k-1} \vee (\hat{m}_{k-1} + 1)] :\ \min_{A \in \mathcal{B}_m,\ A \supset \mathrm{Act}_k} \Pi(A) \leq C \log \bar{N}\, \Pi(\mathrm{Act}_k) \right\}$;

    $\hat{A}_k := \bigcap \{A :\ A \in \mathcal{B}_{\hat{m}_k},\ A \supset \mathrm{Act}_k\}$ `/* regular approximation of` Act_k `*/`;

    $\tau_k := \frac{\hat{m}_k}{\hat{m}_{k-1}}$;

    $N_k := \lfloor N_{k-1}^{\tau_k} \rfloor$;

    for $i = 1$ **to** $\lfloor N_k \cdot \Pi(\hat{A}_k) \rfloor$ **do**

      sample i.i.d. $\left(X_i^{(k)}, Y_i^{(k)}\right)$ with $X_i^{(k)} \sim \hat{\Pi}_k := \Pi(dx | x \in \hat{A}_k)$;

      $S_k := \left\{ \left(X_i^{(k)}, Y_i^{(k)}\right),\ i \leq \lfloor N_k \cdot \Pi(\hat{A}_k) \rfloor \right\}$;

    $\hat{\eta}_k := \hat{\eta}_{\hat{m}_k, \hat{A}_k}$ `/* estimator based on` S_k `*/`;

    $LB := LB - \lfloor N_k \cdot \Pi(\hat{A}_k) \rfloor$;

    $\hat{\eta} := \hat{\eta}_k$ `/* keeping track of the most recent estimator */`;

Remark. Note that the main difference with the bound of Theorem 2.6.2 is that we have $(\beta \wedge 1)\gamma$ instead of $\beta\gamma$ (which coincide iff $\beta \leq 1$) in the exponent. We will further discuss the sources of this difference below.

Proof. The argument follows exactly the same pattern as Theorem 2.6.2, so we will only outline the necessary changes. Note that the resolution levels $\hat{m}_k$ do not have to grow arithmetically anymore, implying, in particular, that the sequence N_k controlling the 'number of labeled observations per unit volume' might grow exponentially. Our first step is to show that

(a) in the case $r + 1 \geq \beta > 1$, with high probability we have $\frac{\hat{m}_{k+1}}{\hat{m}_k} \leq \beta$ for every $k \geq 0$, implying that $N_{k+1} \leq N_k^{\beta}$;

(b) at the same time, for $\beta \leq 1$ we have that for every $k \geq 1$,

$$\tau_k = \tau_1 := \frac{\hat{m}_1}{\hat{m}_0} = 1 + \frac{1}{\hat{m}_0}$$

with high probability, so that N_k grow geometrically and the proof goes through as in Theorem 2.6.2 without further changes.

To obtain the desired inequalities, we will compare two estimators of η: the first is the piecewise–polynomial estimator $\hat{\eta}_k$ constructed by the algorithm on step k and the second is the piecewise–constant estimator $\bar{\eta}_k$ that has similar to $\hat{\eta}_k$ approximation properties (for example, this can be a projection of η onto the space of piecewise–constant functions with a suitable resolution level). As a result, we will be able to relate the "active sets" associated to these estimators, taking advantage of the fact that the active set associated to $\bar{\eta}_k$ is always a union of dyadic cubes, and conclude that $\hat{m}_{k+1}$ cannot exceed the resolution level of $\bar{\eta}_k$.

We proceed by inductive argument. As we have already seen before, for $k = 0$ we have $\|\hat{\eta}_0 - \eta\|_{\infty,\mathrm{supp}(\Pi)} \leq \delta_0/2$, where δ_0 is defined in line 11 of Algorithm 2. Note

that the following inclusions hold:

$$\{x : |\eta(x)| < \delta_0/2\} \subseteq \mathrm{Act}_1 \subseteq \{x : |\eta(x)| < 3\delta_0/2\}\,.$$

Indeed,

$$|\eta(x)| < \delta_0/2 \implies |\hat{\eta}_0(x)| < \delta_0/2 + |\eta(x) - \hat{\eta}_0(x)| < \delta_0 \implies x \in \mathrm{Act}_1$$

and

$$x \in \mathrm{Act}_1 \implies |\hat{\eta}_0(x)| < \delta_0 \implies |\eta(x)| < 3\delta_0/2.$$

Let $\bar{\eta}_0$ be a projection of η onto $\mathcal{F}^0_{l_0}$ (the space of piecewise-constant functions on the dyadic partition of the unit cube) where l_0 is such that $|\eta - \bar{\eta}_0| \leq \delta_0/2$, let $\bar{\mathcal{F}}_0$ be the band of size $2\delta_0$ around $\bar{\eta}_0$ and $\bar{A}_1 = \big\{x : \exists f_1, f_2 \in \bar{\mathcal{F}}_0 : \text{sign }(f_1(x)) \neq \text{sign }(f_2(x))\big\}$. By a similar argument we have

$$\{x : |\eta(x)| < 3\delta_0/2\} \subseteq \bar{A}_1 \subseteq \{x : |\eta(x)| < 5\delta_0/2\}\,,$$

which implies

$$\{x : |\eta(x)| < \delta_0/2\} \subseteq \mathrm{Act}_1 \subseteq \bar{A}_1 \subseteq \{x : |\eta(x)| < 5\delta_0/2\}\,.$$

Note that

1. $\bar{A}_1$ is the union of dyadic cubes with edge length 2^{-l_0};
2. $\frac{\Pi(\bar{A}_1)}{\Pi(\mathrm{Act}_1)} \leq \frac{\Pi(\{x:|\eta(x)|\leq 5\delta_0/2\})}{\Pi(\{x:|\eta(x)|\leq \delta_0/2\})} \leq 5^\gamma \log \bar{N}$ by assumption (iv),

meaning that $\bar{A}_1$ gives the required "regular approximation" of Act_1, hence $\hat{m}_1 \leq l_0$. When $\beta > 1$ (case (a)), we have that

$$\|\eta - \bar{\eta}_0\|_\infty \leq B_1 2^{-(\beta\wedge 1)l_0} \leq \sqrt{\log \bar{N}} \cdot 2^{-l_0}. \tag{2.6.18}$$

Recalling the definition of $\delta_0 := C\left(\log \frac{\bar{N}}{\alpha}\right)^{2+\frac{d}{2\nu}} \sqrt{\frac{2^{d\hat{m}_0}}{N_0}}$, we see that it clearly suffices to take

$$2^{l_0} \simeq \frac{\sqrt{\frac{N_0}{2^{d\hat{m}_0}}}}{\left(\log \frac{\bar{N}}{\alpha}\right)^{3/2+\frac{d}{2\nu}}}$$

to have $\|\eta - \bar{\eta}_0\|_\infty \le \delta_0/2$. At the same time, by (2.6.10) we have

$$2^{\beta\hat{m}_0} \gtrsim \frac{\sqrt{\frac{N_0}{2^{d\hat{m}_0}}}}{\left(\log\frac{\bar{N}}{\alpha}\right)^{1+\frac{d}{2\nu}}}$$

with probability $\ge 1-\alpha$, giving that on the same event $2^{\beta\hat{m}_0} \ge 2^{l_0}$, hence $\frac{l_0}{\beta\hat{m}_0} \le 1$, implying $\hat{m}_1 \le l_0 \le \beta\hat{m}_0$, as desired.

Proceeding in a similar way, we get that the following holds with probability $\ge 1-2L\alpha$ uniformly for every $1 \le k \le L$ (for the definition of δ_k, see Algorithm 2):

$$\begin{aligned}
&1)\ \|\hat{\eta}_k - \eta\|_{\infty,\hat{A}_k} \le \delta_k/2,\\
&2)\ \{x : |\eta(x)| < \delta_k/2\} \subseteq \mathrm{Act}_{k+1} \subseteq \bar{A}_{k+1} \subseteq \{x : |\eta(x)| < 5\delta_k/2\},\\
&3)\ 2^{l_k} \simeq \frac{\sqrt{\frac{N_k}{2^{d\hat{m}_k}}}}{\left(\log\frac{\bar{N}}{\alpha}\right)^{(3/2+\frac{d}{2\nu})\frac{\hat{m}_k}{\hat{m}_0}}},\\
&4)\ 2^{\beta\hat{m}_k} \gtrsim C\frac{\sqrt{\frac{N_k}{2^{d\hat{m}_k}}}}{\left(\log\frac{\bar{N}}{\alpha}\right)^{(1+\frac{d}{2\nu})\frac{\hat{m}_k}{\hat{m}_0}}},
\end{aligned} \tag{2.6.19}$$

hence on this event $\frac{l_k}{\hat{m}_k} \le \beta$ and $\hat{m}_{k+1} \le \beta\hat{m}_k$. Similar reasoning gives the second part of the claim (the case $\beta \le 1$, with the only change occurring due to $\beta \wedge 1 = \beta$, see (2.6.18)).

It remains to bound $\frac{\hat{m}_L}{\hat{m}_0}$ to control the power of the logarithmic term. One way to do this is as follows: let L be the number of the last iteration before termination. Since the number of labels requested on $(L-1)$st iteration does not exceed $\bar{N}$, we have $2\bar{N} \ge N_{L-1}\Pi(A_{L-1})$. On the event where inequalities (2.6.19) hold A_{L-1} contains at least one dyadic cube with edge length $2^{-\hat{m}_{L-1}}$, hence by our assumptions on Π and Theorem 2.5.6

$$\begin{aligned}
\Pi(A_{L-1}) \ge u_1 2^{-d\hat{m}_{L-1}} = u_1\left(2^{-d\hat{m}_0}\right)^{\hat{m}_{L-1}/\hat{m}_0} \ge u_1\left(2^{-d\bar{m}_0}\right)^{\hat{m}_{L-1}/\hat{m}_0} \ge\\
\ge u_1\left(CN_0^{-d/(2\beta+d)}\right)^{\hat{m}_{L-1}/\hat{m}_0} \ge CN_{L-1}^{-d/(2\beta+d)}.
\end{aligned}$$

This gives

$$2\bar{N} \geq N_{L-1}\Pi(A_{L-1}) \geq cN_{L-1}^{2\beta/(2\beta+d)} \geq c\left(N_0^{2\beta/(2\beta+d)}\right)^{\hat{m}_{L-1}/\hat{m}_0},$$

and since $N_0 := \lfloor \bar{N}^{1/2(r+1)} \rfloor$ by definition, we have that with probability $\geq 1 - 2L\alpha$, $\frac{\hat{m}_{L-1}}{\hat{m}_0} \leq \frac{(r+1)(2\beta+d)}{\beta}$ and

1. $\frac{\hat{m}_L}{\hat{m}_0} = \frac{\hat{m}_L}{\hat{m}_{L-1}}\frac{\hat{m}_{L-1}}{\hat{m}_0} \leq \beta\frac{(r+1)(2\beta+d)}{\beta} \leq (r+1)(2(r+1)+d)$ in the case $\beta > 1$ and

2. $\frac{\hat{m}_L}{\hat{m}_0} \leq 4 + \frac{2d}{\nu}$ if $\beta \leq 1$, see (2.6.16).

Set $\bar{\tau} := (r+1)(2(r+1)+d) \vee \left(4 + \frac{2d}{\nu}\right)$. As before, the final result is implied once we have the lower bound on N_L, L being the index of the last iteration.

In the case $\beta > 1$, the required bound is obtained as follows. First, note that for $k \geq 2$

$$q_k := N_k N_{k-1}^{-\frac{\beta\gamma}{2\beta+d}} \geq N_{k-1}^{\tau_k}\left(N_{k-2}^{\tau_{k-1}}\right)^{-\frac{\beta\gamma}{2\beta+d}} \geq \left(N_{k-1}N_{k-2}^{-\frac{\beta\gamma}{2\beta+d}}\right)^{\tau_{k-1}} = q_{k-1}^{\tau_{k-1}},$$

since τ_k is nondecreasing. Moreover, the sequence $\{q_k\}$ grows exponentially fast. In particular, $\sum_{i=1}^{j} q_i \leq Cq_j$ for any $j \geq 2$.

This implies, together with the inequality $\Pi(\hat{A}_k) \leq \Pi(|\eta(x)| \leq \frac{3}{2}\delta_{k-1}) \leq K\delta_{k-1}^{\gamma}$,

$$N = \sum_{k=0}^{L} \lfloor N_k \Pi(\hat{A}_k) \rfloor \leq C\left(\log\frac{\bar{N}}{\alpha}\right)^{\bar{\tau}\gamma(2+d/2\nu)} \sum_{k=0}^{L} q_k \leq$$
$$\leq C\left(\log\frac{\bar{N}}{\alpha}\right)^{\bar{\tau}\gamma(2+d/2\nu)} q_L, \tag{2.6.20}$$

Since on event (2.6.19) $\tau_L \leq \beta$, we have

$$q_L \leq N_{L-1}^{\tau_L} N_{L-1}^{-\frac{\beta\gamma}{2\beta+d}} \leq N_{L-1}^{\tau_L\left(1-\frac{\gamma}{2\beta+d}\right)}. \tag{2.6.21}$$

Combining inequalities (2.6.20),(2.6.21), it is easy to see that with high probability (at least $1 - 2L\alpha$)

$$N_L = \lfloor N_{L-1}^{\tau_L} \rfloor \geq C\left(\frac{N}{\left(\log\frac{\bar{N}}{\alpha}\right)^{\bar{\tau}\gamma(2+d/2\nu)}}\right)^{\frac{2\beta+d}{2\beta+d-\gamma}},$$

and the final result now follows from inequality (2.3.2) of Proposition 2.3.1.

In the case $\beta \leq 1$, the proof is similar to Theorem 2.6.2. The final form of the bound is a concatenation of the estimates for $\beta > 1$ and $\beta \leq 1$. □

2.7 Running time analysis

We continue this section by discussing the running time of Algorithm 1. Algorithm 2 is of mostly theoretical interest since it involves exact computation of the level sets of multivariate polynomials of high degree. Assume that the algorithm has access to the sampling subroutine that, given $A \subset [0,1]^d$ with $\Pi(A) > 0$, generates i.i.d. (X_i, Y_i) with $X_i \sim \Pi(dx|x \in A)$.

Proposition 2.7.1. *The running time of* **Algorithm 1** *with label budget* $\bar{N}$ *is*

$$\mathcal{O}(\bar{N} \log^2 \bar{N}).$$

Remark In view of Theorem 2.6.2, the running time required to output a classifier $\hat{g}$ such that $R_P(\hat{g}) - R^* \leq \varepsilon$ with probability $\geq 1 - \alpha$ is

$$\mathcal{O}\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2\beta+d-\beta\gamma}{\beta(1+\gamma)}} \operatorname{poly}\left(\log \frac{1}{\varepsilon\alpha}\right)\right),$$

given that the label threshold is large enough.

Proof. We will use the notations of Theorem 2.6.2. Let N_k^{act} be the number of labels requested by the algorithm on step k. The resolution level $\hat{m}_k$ is always chosen such that $\hat{A}_k$ is partitioned into at most N_k^{act} dyadic cubes. This means that the estimator $\hat{\eta}_k$ takes at most N_k^{act} distinct values and can be found in $\mathcal{O}(N_k^{\text{act}})$ steps. The key observation is that for any k, the active set $\hat{A}_k$ is always represented as the union of a finite number(at most N_{k-1}^{act}) of dyadic cubes: to determine if a cube $R_j \subset \hat{A}_{k+1}$, it is enough to take a point $x \in R_j$ and compare $\operatorname{sign}(\hat{\eta}_k(x) - \delta_k)$ with $\operatorname{sign}(\hat{\eta}_k(x) + \delta_k)$: $R_j \in \hat{A}_{k+1}$ only if the signs are different(so that the confidence band crosses zero level). This can be done in $\mathcal{O}(N_k^{\text{act}})$ steps, so the whole k-th iteration running time is

$\mathcal{O}(N_k^{\text{act}})$. Next, resolution level $\hat{m}_0$ can be found in $\mathcal{O}(N_0 \log^2 \bar{N})$ steps, see Remark 2 after (2.5.4). Since $\sum_k N_k^{\text{act}} = \mathcal{O}(\bar{N})$ and $N_0 = \mathcal{O}(\bar{N}^{1/2})$, the result follows. □

2.8 Simulation results

A (version of) Algorithm 1 was implemented in Matlab. The following model was used for simulations:

$$Y_i = \text{sign } [f(X_i) + \varepsilon_i], \ \varepsilon \sim \mathcal{N}(0, \sigma^2), \ i = 1 \dots 34$$
$$f(x) = x \left(1 + \sin \frac{5}{x}\right) \sin(4\pi x), \quad \sigma^2 = 0.2.$$

Note that in this case sign η = sign f, where η is the regression function $\eta(x) = \mathbb{E}(Y|X = x)$, see figure 3. Figure 4 shows the classifier output by the active algorithm

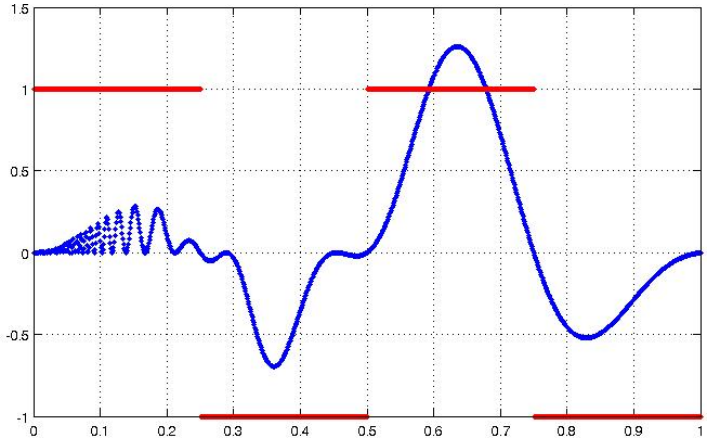

Figure 3: Graph of $f(x)$ and sign $f(x)$

with label budget $N = 34$ that performed 3 iterations. Figure 5 shows the plug-in classifier based on the wavelet threshold estimator produced by Matlab wavelet toolbox. Clearly, the output of the active learning algorithm is closer to the true underlying model.

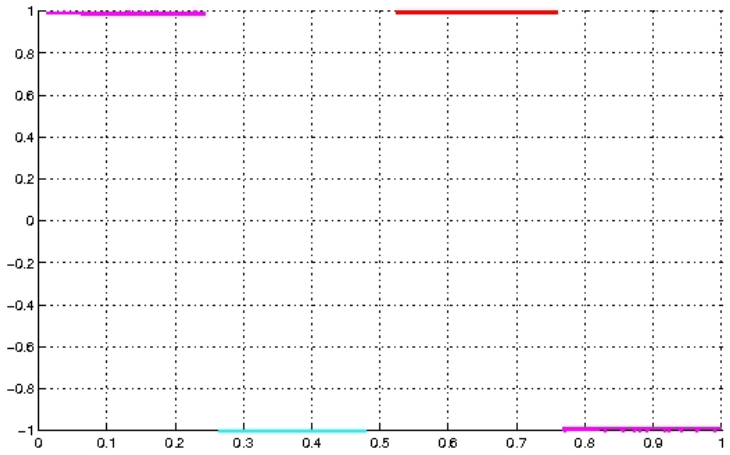

Figure 4: Classifier produced by Algorithm 1; each iteration marked with different color

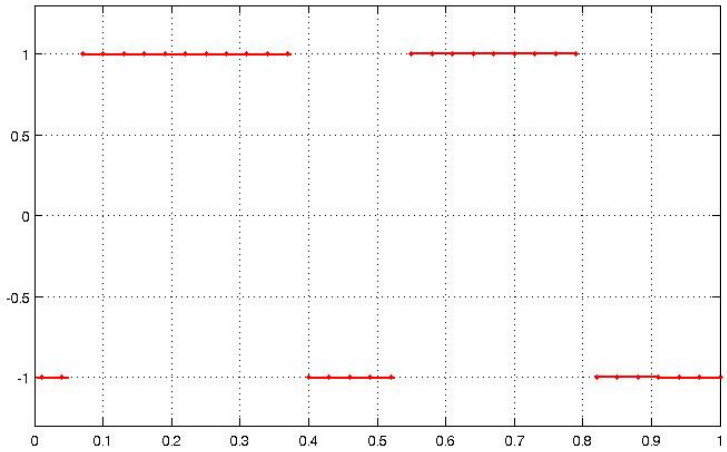

Figure 5: Plug-in classifier based on wavelet threshold estimator

2.9 Concluding remarks

The results above give some insight into the limitations and possible improvements of active methods over passive learning for a broad class of underlying distributions. However, as seen from Theorems 2.6.1, 2.6.2, 2.6.3, there is a gap between the lower and upper bounds following from our analysis. For the case $0 < \beta \leq 1$, the gap is logarithmic but we consider it satisfactory for our purposes. For the general case, there is a difference in the polynomial rate of convergence.

We believe that the reason for this difference is the fact that Theorem 2.6.3 gives

the worst-case analysis, in a sense that the regular approximation of the active set is assumed to have very high resolution level (in other words, it uses the dyadic cubes with small edge length) compared to the 'optimal' resolution level associated to the estimator of regression function. However, in many cases the situation might be far from the worst case. For example, if the sequence of resolution levels $\{\hat{m}_k\}$ grows arithmetically and $\beta\gamma \leq d$, then the rate of convergence resulting from the same analysis will match the lower bound of Theorem 2.6.1 up to logarithmic terms. However, at this point we are unaware of the natural and convenient way to describe this gradation.

Another question one might ask is if the piecewise-polynomial estimators give a provable advantage over piecewise-constant estimators during the intermediate stages of the algorithm. An alternative method could proceed as Algorithm 1, using the piecewise-constant estimators to reduce the size of the active set, and construct the more powerful piecewise-polynomial estimator on the last iteration. While such an approach seems appealing for practical purposes due to computational simplicity, it results in suboptimal convergence rate $N^{-\mu\frac{\beta(1+\gamma)}{2\beta+d-(\beta\wedge 1)\gamma}}$ which differs from the rate of Theorem 2.6.3 by a multiplicative factor $\mu := \frac{2(\beta\wedge 1)+d}{2(\beta\vee 1)+d} \leq 1$.

Finally, we note that our algorithm can be viewed as a method for level set estimation. This gives a possibility of applications to multi-label classification tasks viewed as simultaneous estimation of several level sets. A work in this direction is one of our future priorities.

CHAPTER III

SPARSE RECOVERY IN INFINITE DICTIONARIES

3.1 Introduction

Many prediction problems encountered in today's world involve high-dimensional data, often resulting in a situation when the number of available observations is smaller than the number of parameters. At the same time, it was noticed that the number of significant features might be a lot smaller (this is often referred to as "sparsity assumption"). If these features were known in advance, classical parametric methods would provide satisfactory solutions. Unfortunately, significance of parameters has to be learned from the data as well, and this problem led to development of modern methods such as LASSO [83] and Dantzig Selector [20], along with several modifications. Other methods used in large margin classification, such as Boosting [34], combine simple classifiers from a high-dimensional class to produce a linear combination with very strong generalization properties, and avoid overfitting at the same time. Much effort has been made to understand the reasons for this type of behavior (e.g., [77], [60]).

Most of the aforementioned problems can be stated in the framework of *dictionary learning*, and we proceed with its more detailed description.

3.2 Dictionary learning: probabilistic framework

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

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

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

Similar notations will be used for Π, Π_n and other measures. Let $\ell(y, \cdot)$ be the loss function such that for all $y \in T$, $\ell(y, \cdot)$ is convex. As suggested by its name, $\ell(y, f(x))$ measures the loss suffered from predicting Y by $f(X)$. Choice of the loss is usually motivated by the nature of the problem, for example the squared loss $\ell(y, f(x)) = (y - f(x))^2$ for regression or $\ell(y, f(x)) = \phi(yf(x))$ in binary classification, where ϕ is a convex nonincreasing function with $\phi(0) = 1$. The latter is sometimes called *the surrogate loss* [9] and serves as a convex majorant of the (non-convex) binary loss $\ell(y, f(x)) = I\{yf(x) \leq 0\}$. Common choices for the function ϕ are $\phi(z) = e^{-z}$ (boosting), $\phi(z) = \log_2(1 + e^{-z})$ (logistic regression), among others. For a function $f : S \mapsto \mathbb{R}$, let $(\ell \bullet f)(x, y) := \ell(y, f(x))$.

A *dictionary* (or a *base class*) is a given family $\mathcal{H}$ of measurable functions equipped with a σ-algebra and with a finite measure μ. For most results, we will assume that the elements of the dictionary are uniformly bounded by a constant $M > 0$ (and will take $M = 1$ for simplicity). This assumption is very natural when the response variable is known to be bounded (e.g., in binary classification problem). In other cases, such as density estimation, it is sometimes possible to relax boundedness assumption.

The goal of dictionary learning is to estimate the optimal (unknown) prediction rule $g_* := \operatorname{argmin} P\ell(y, g(x))$ (where the minimum is taken over all measurable functions) by a linear (convex) combination of the elements of $\mathcal{H}$. Common examples of the dictionaries include:

1. A subset of a basis, such as wavelets, splines, trigonometric polynomials, etc.

2. A collection of pre-defined estimators of g_* obtained by different methods. In this case, the goal is to "aggregate" these estimators to obtain a new one which

will perform at least as good as each of the initial candidates.

3. A location family $\mathcal{H} = \{\phi(\cdot - t),\ t \in \mathbb{R}\}$, where ϕ is a bounded density function. This dictionary can be useful in deconvolution problem.

Cardinality of the dictionary can be very large (or even infinite), so the sparsity assumptions translates into the belief that there exists a linear combination with only few non-zero coordinates. A common way to recover sparse solution is to solve a penalized empirical risk minimization problem. For example, in a popular case of regression with a dictionary of cardinality N it takes the form

$$\hat{\lambda} := \operatorname{argmin}_{\lambda \in \mathbb{R}^N} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{N} \lambda_j h_j(X_i) \right)^2 + \operatorname{pen}(\lambda),$$

where $\operatorname{pen}(\lambda)$ is a penalty term. The natural choice of the penalty would be $\operatorname{pen}(\lambda) := \varepsilon \|\lambda\|_0 = \varepsilon \sum_{i=1}^{N} I\{\lambda_i \neq 0\}$, but there is little hope to solve the resulting problem. Instead, $\operatorname{pen}(\lambda) := \varepsilon \|\lambda\|_1 = \varepsilon \sum_{i=1}^{N} |\lambda_i|$ is used (here, ε is a properly chosen regularization parameter). In the latter case, the problem becomes convex and can be solved by existing methods. This is one of the possible formulations of LASSO (Least Absolute Shrinkage and Selection Operator). This problem attracted a lot of attention of the research community due to excellent performance on the real-world data. Significant contributions to theoretical explanations of this method were made in the works of D. Donoho [31], [32], who established connections to the properties of high-dimensional polytopes.

It turns out that in some cases LASSO exactly identifies relevant features in the support of λ. In particular, this happens when the dictionary possesses some "almost-orthogonality" properties with respect to $L_2(\Pi)$, such as the "restricted isometry" condition in the works of E. Candès, J. Romberg and T. Tao [22], [20]. More specifically, the restricted isometry constant $\delta_d(\Pi)$ is the minimal δ such that for all $\lambda \in \mathbb{R}^N$

with at most d non-zero coefficients

$$(1-\delta)\|\lambda\|_2 \leq \left\|\sum_{i=1}^{N}\lambda_i h_i\right\|_{L_2(\Pi)} \leq (1+\delta)\|\lambda\|_2.$$

A result in [21] (also see [22]) states that in the noiseless case, (a version of) LASSO is able to *exactly* recover an unknown λ_* with d non-zero coefficients whenever $\delta_{2d} < \sqrt{2}-1$. In the presence of noise, recovery is still possible under the same assumption, with an error being proportional to the noise rate.

Other important works on the variants of LASSO include [17], [89], [49], [96], [11], [67], to name a few. In a notable paper by P. Bartlett, S. Mendelson and J. Neeman [10], powerful analysis methods were used to handle situations when elements of the dictionary are not uniformly bounded. In [90], authors analyze and compare different conditions on the dictionary, including the "restricted eigenvalue" condition introduced by P. Bickel, Ritov and Tsybakov [11] and "compatibility condition" of S. van de Geer [88].

In a series of papers [51], [49], [50] V. Koltchinskii introduced the notion of *"alignment coefficient"* and successfully applied it to study the variants of LASSO and Dantzig selector. Our work continues the line of research started in [51]: this paper investigates the problem of sparse mixture recovery, meaning that the unknown "true" solution is believed to have a good approximation in the convex hull of the dictionary (rather than in the linear span). A canonic example of this type of problems is density estimation with L_2 loss (in particular, the dictionary is a family of probability density functions). In this case, it is possible to replace $\|\lambda\|_1$ penalty by the (negative) entropy $H(\lambda) := \sum_{i=1}^{N}\lambda_i \log \lambda_i$. The advantage of this penalty is strict convexity, which allows to study random error and approximation error separately and leads to interesting theoretical results. Similar ideas were previously investigated in [27] in the context of aggregation with exponential weights.

Our main goal is to understand what happens in the case of (possibly uncountable)

infinite dictionary and continuos mixtures. In particular, such dictionaries might contain highly correlated (or even linearly dependent) elements, and the restricted isometry-type conditions do not have direct analogues. Instead, we define a version of alignment coefficient, and it turns out that geometric assumptions can often be expressed in terms of Sobolev-type norms; the details are discussed below.

We shall identify relevant parameters which control the risk, and prove sparsity oracle inequalities for prediction problems (Corollary 3.5.6) and density estimation (Corollary 3.6.3). In the latter case, the oracle inequality is exact (see the remark after Theorem 3.6.1).

Some parts of this work appeared previously in the joint paper with V. Koltchinskii [54], and the present chapter is mainly based on it (namely, the part devoted to prediction problems). Results of section 3.6 on L_2-density estimation were not previously published.

3.3 Problem statement, notations and main assumptions

Suppose we are given a probability measure Λ on $\mathcal{H}$ such that $\lambda = \frac{d\Lambda}{d\mu}$. The (negative) entropy $H(\lambda)$ is defined via

$$H(\lambda) := \int\limits_{\mathcal{H}} \lambda(h) \log \lambda(h) d\mu(h).$$

In what follows, we shall only consider densities with finite entropies. Assume that the mapping $S \times \mathcal{H} \ni (x, h) \mapsto h(x)$ is measurable, and let f_λ denote the mixture of the functions from dictionary $\mathcal{H}$ with respect to λ:

$$f_\lambda(\cdot) := \int\limits_{\mathcal{H}} h(\cdot)\lambda(h) d\mu(h).$$

The excess risk $\mathcal{E}(f_\lambda)$ of the estimator f_λ is defined as the difference of true risk and the minimal risk, that is

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

Throughout the chapter, we make the following assumption:

Assumption 3.1. $\inf_{f:S\mapsto R} P(\ell \bullet f)$, *where the infimum is taken over all measurable functions, is attained at some uniformly bounded function* f_*.

Let $\mathbb{D}$ be a convex set of probability densities on $\mathcal{H}$ having finite entropies. Consider the following penalized risk minimization problem:

$$\lambda_\varepsilon := \operatorname{argmin}_{\lambda\in\mathbb{D}}\Big[P(\ell \bullet f_\lambda) + \varepsilon H(\lambda)\Big] \tag{3.3.1}$$

together with its empirical version:

$$\hat{\lambda}_\varepsilon := \operatorname{argmin}_{\lambda\in\mathbb{D}}\Big[P_n(\ell \bullet f_\lambda) + \varepsilon H(\lambda)\Big]. \tag{3.3.2}$$

Since $\hat{\lambda}_\varepsilon$ depends only on the data, we can use it as an estimator of unknown λ_ε. Note that, due to the convexity of the loss, both (3.3.1) and (3.3.2) are convex optimization problems. We will use the notations $\Lambda_\varepsilon, \hat{\Lambda}_\varepsilon$ for the probability measures with densities $\lambda_\varepsilon, \hat{\lambda}_\varepsilon$, respectively.

One main goal will be to show that the "approximate sparsity" of the true penalized solution λ_ε implies that the corresponding empirical solution $\hat{\lambda}_\varepsilon$ possesses the same property with a high probability. More precisely, it will be said that λ_ε is "approximately sparse" if there exists a measurable set $\mathcal{H}' \subset \mathcal{H}$ such that $\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$ is small and, at the same time, there exists a subspace $L \subset L_2(\Pi)$ of 'small' dimension d that provides a good $L_2(\Pi)$-approximation of the functions from the set $\mathcal{H}'$. We will show that in this case the empirical solution $\hat{\lambda}_\varepsilon$ is also approximately supported on the same set $\mathcal{H}'$ in the sense that $\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$ is small. Thus, both the empirical solution $\hat{\lambda}_\varepsilon$ and the true solution λ_ε follow the same "sparsity pattern": they are concentrated on the same set of functions $\mathcal{H}'$ which can be well approximated by a linear subspace of small dimension. Our next goal is to obtain probabilistic bounds on the *random error* $\big|\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) - \mathcal{E}(f_{\lambda_\varepsilon})\big|$ in terms of characteristics of the sparsity of the problem, such as the measure $\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$ and the dimension d of the approximating space L. We

focus only on the losses of quadratic type (see Definition 3.1 below) and this allows us to reduce the problem to bounding the $L_2(\Pi)$-error $\|f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}\|^2_{L_2(\Pi)}$. At the same time, we derive upper bounds on the Kullback-Leibler type distance between $\hat{\lambda}_\varepsilon$ and λ_ε.

Another problem is to bound the approximation error $\mathcal{E}(f_{\lambda_\varepsilon})$ which, for the losses of quadratic type, is equivalent to bounding the $L_2(\Pi)$-approximation error $\|f_{\lambda_\varepsilon} - f_*\|^2_{L_2(\Pi)}$. We show that the size of this error is small if there exists an oracle $\lambda \in \mathbb{D}$ that is "sparse" in the sense that it is concentrated on a "small" set of functions $\mathcal{H}' \subset \mathcal{H}$ and, at the same time, possesses certain regularity properties, often expressed in terms of Sobolev-type norms of $\log \lambda$.

As it has been observed previously in the case of sparse recovery problems for finite dictionaries (see [49], [51]), the fact that the penalty is strictly convex allows us to study the random error independently of the approximation error, but geometric parameters of the dictionary needed to control these quantities are not the same.

In the end of this Chapter, we present some applications and examples showing how the quantities involved in the bounds can be computed. This allows us to state the final versions of oracle inequalities for these specific cases. The main tools used in the proofs include Talagrand's concentration inequality, Dudley's entropy bound for subgaussian processes, symmetrization and contraction inequalities. Details and references for these results are given in Chapter 1.

3.4 Preliminaries

Below, we formulate some basic results that become a starting point of subsequent detailed analysis.

3.4.1 Assumptions on the loss

Assume that:

(1) for all $y \in T$, $\ell(y,\cdot)$ is a convex twice differentiable function, ℓ''_u is uniformly bounded in $T \times [-1,1]$ and

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

(2)

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

Definition 3.1. *The loss function ℓ satisfying assumptions (1), (2) will be called the loss of quadratic type.*

In particular, our assumptions imply that

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

where $C = \sup_{y,u} \ell''_u(y,u)$. Moreover, the following proposition also holds for the losses of quadratic type:

Proposition 3.4.1. *There exists a constant $C > 0$ depending only on ℓ such that for all $\lambda, \bar\lambda \in \mathbb{D}$,*

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

Proof. Since ℓ is a loss function of quadratic type, the first order Taylor expansion implies that

$$(\ell \bullet f_{\bar\lambda})(x,y) - (\ell \bullet f_\lambda)(x,y) = (\ell' \bullet f_\lambda)(x,y)(f_{\bar\lambda} - f_\lambda)(x) + \rho(x,y),$$

where $|\rho(x,y)| \le C\left(f_{\bar\lambda} - f_\lambda\right)^2(x)$. Integrating with respect to P yields

$$\mathcal{E}(f_{\bar\lambda}) - \mathcal{E}(f_\lambda) = \langle \ell' \bullet f_\lambda, f_{\bar\lambda} - f_\lambda\rangle_{L_2(P)} + P\rho, \tag{3.4.1}$$

where

$$|P\rho| \le C\|f_{\bar\lambda} - f_\lambda\|^2_{L_2(\Pi)} \tag{3.4.2}$$

Since, for all uniformly bounded functions $h : S \mapsto \mathbb{R}$, $P(\ell' \bullet f_*)h = 0$, we have

$$\langle \ell' \bullet f_\lambda, f_{\bar{\lambda}} - f_\lambda \rangle_{L_2(P)} = \langle \ell' \bullet f_\lambda - \ell' \bullet f_*, f_{\bar{\lambda}} - f_\lambda \rangle_{L_2(P)},$$

and, using the Lipschitz condition for the loss ℓ, we get

$$\left| \langle \ell' \bullet f_\lambda, f_{\bar{\lambda}} - f_\lambda \rangle_{L_2(P)} \right| \le C \| f_\lambda - f_* \|_{L_2(\Pi)} \| f_{\bar{\lambda}} - f_\lambda \|_{L_2(\Pi)} \le C \sqrt{\frac{\mathcal{E}(f_\lambda)}{\tau}} \| f_{\bar{\lambda}} - f_\lambda \|_{L_2(\Pi)}. \tag{3.4.3}$$

It remains to combine (3.4.3) with (3.4.1) and (3.4.2). □

Previously mentioned loss functions, such as quadratic loss $\ell(y, f_\lambda(\cdot)) = (y - f_\lambda(\cdot))^2$, exponential loss $\ell(y, f_\lambda(\cdot)) = e^{-y f_\lambda(\cdot)}$ and the logit loss $\ell(y, f_\lambda(\cdot)) = \log_2 \left(1 + e^{-y f_\lambda(\cdot)}\right)$ satisfy our assumptions.

3.4.2 Assumptions on the dictionary

Complexity of the dictionary $\mathcal{H}$ will be characterized in terms of the continuity modulus of a certain (conditionally) Gaussian process. In particular, our assumptions allow for the unified treatment of conditions on covering and bracketing numbers. For $h \in \mathcal{H}$, let

$$G_n(h) := \frac{1}{n} \sum_{i=1}^{n} g_i h(X_i), \ h \in \mathcal{H},$$

where $\{g_i\}_{i=1}^n$ is a sequence of iid $N(0,1)$ random variables, independent of X_i's. Conditionally on X's, this is a Gaussian process. Let

$$\mathcal{H}(u) := \left\{ h_1 - h_2 : \ h_i \in \mathcal{H}, \ \| h_1 - h_2 \|_{L_2(\Pi)} \le u \right\}.$$

We will make the following assumption on $G_n(\cdot)$: there exists a sequence of functions $w_n(u)$ such that

$$\sqrt{n}\, \mathbb{E} \sup_{f \in \mathcal{H}(u)} |G_n(f)| \le w_n(u),$$

$$w(u) := \limsup_{n \to \infty} w_n(u), \ \lim_{u \to 0} w(u) = 0. \tag{3.4.4}$$

In particular, this implies that $\mathcal{H}$ is Π-Donsker [91]. Moreover, let

$$\Omega^{1/2}(\delta) := K \int_{\delta}^{\sqrt{2}\mathcal{D}} w(u)\frac{du}{u^2},$$

where K is a numerical constant and $\mathcal{D} = \mathrm{diam}_{L_2(\Pi)}(\mathcal{H})$. Expression for $w_n(u)$ can often be obtained from the following *complexity assumption* on the base class $\mathcal{H}$: there exists a nonnegative non-increasing function T such that $T(u) \to \infty$ as $u \to 0$, $T(1/u)$ is regularly varying at ∞ with exponent $\alpha \in [0, 2)$ [1] and, with probability 1,

$$\log N(\mathcal{H}; L_2(\Pi_n); u/2) \le T\left(\frac{u}{\|F\|_{L_2(P_n)}}\right),\ u > 0, \tag{3.4.5}$$

where $N(\mathcal{H}; L_2(\Pi_n); u/2)$ is the covering number of $\mathcal{H}$ with respect to $L_2(\Pi)$ (see Definition 1.3) and F is the (measurable) envelope of the class $\mathcal{H}$, meaning that $|h(x)| \le F(x)$ for every $h \in \mathcal{H}$ and $x \in S$. Typical examples include $T(u) = \log \frac{A}{u},\ A > 0$; $T(u) = u^{-\alpha},\ \alpha \in (0, 2)$.

When working with regularly varying functions, we will frequently use some well-known properties. One important fact states that

$$\int_0^x \sqrt{T(u)}du \le C(T)x\sqrt{T(x)}$$

for all $x > 0$. For this and other facts, see [75], in particular, Theorem 2.1.

3.4.3 Uniformly bounded base classes

If the elements of $\mathcal{H}$ are uniformly bounded by 1 (so that $F \equiv 1$ is the envelope function), we have the following result:

Proposition 3.4.2. *Assume (3.4.5) is satisfied. Then*

$$w_n(u) \le C\left[u\sqrt{T(u)} \bigvee \frac{T(u)}{\sqrt{n}}\right],$$

$$w(u) \le Cu\sqrt{T(u)},$$

$$\Omega^{1/2}(\delta) \le \tilde{K}\sqrt{T(\delta)}\log\frac{1}{\delta}.$$

[1] this means $\lim_{u\to\infty} \frac{T(1/(su))}{T(1/u)} \to s^\alpha$ for some $\alpha \in [0, 2)$ and any $s > 0$.

Proof. The proof repeats an argument of Theorem 3, [39], but one has to apply contraction inequality for the Gaussian sums (see Corollary 3.17 in [62]) instead of the similar result for Rademacher processes.
Slightly different argument (which does not use contraction inequality) is given in Proposition 3.4.4 below. □

In particular, one can take $T(u) = \log N$ for a base class of finite cardinality N or $T(u) = (2V + 1) \log \left(\frac{A}{u}\right)$ if a base class is a subset of a VC-subgraph family of VC-dimension V (for the definition of VC-dimension, see [91], Chapter 2.6).

Assumption on the covering numbers with respect to $\| \cdot \|_{L_2(P_n)}$ can be replaced by the similar assumption on the bracketing entropy with respect to a single norm $\| \cdot \|_{L_2(P)}$.

Definition 3.2 (Bracketing)**.** *Given two real-valued functions l and u such that $l \leq u$ and $\|l-u\|_{L_2(\Pi)} \leq \varepsilon$, the ε-bracket $[l, u]$ is the set of all $f : \ l \leq f \leq u$. The bracketing number $N_{[]}(\mathcal{H}, L_2(P), \varepsilon)$ is the minimal number of ε-brackets needed to cover $\mathcal{H}$. Here, the upper and lower bounds u and l do not need to belong to $\mathcal{H}$.*

Proposition 3.4.3. *Suppose that there exists a nonnegative non-increasing function T_1 such that $T_1(u) \to \infty$ as $u \to 0$, T_1 is regularly varying of exponent $0 < \alpha < 2$ and*

$$\log N_{[]}(\mathcal{H}, L_2(P), u) \leq T_1(u). \tag{3.4.6}$$

Then

$$w_n(u) \leq C \int_0^{2u} \sqrt{1 + T_1(s)} ds.$$

Proof. First, we reduce the bound on the continuity modulus of $G_n(h)$ to the corresponding bounds on the empirical process $Z_n(h) := \sqrt{n}(P_n - P)h$, since the bracketing entropy controls the latter quantity. Reduction is done with the help of desymmetrization and general multiplier inequalities. First, note that by Theorem

1.2.6 for any $k \geq 1$

$$\sqrt{k}\,\mathbb{E} \sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} |R_k(h_1-h_2)| \leq 2\mathbb{E} \sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} |Z_k(h_1-h_2)| + u, \qquad (3.4.7)$$

where we used that by Hölder's inequality $\sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} P|h_1-h_2| \leq u$. Here, $R_k(h)$ is the Rademacher process, see Chapter 1 for definition. Next, by the general multiplier inequality (Lemma 2.9.1 in [91])

$$\sqrt{n}\,\mathbb{E} \sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} G_n(h_1-h_2) \leq C \max_{1\leq k\leq n} \sqrt{k}\,\mathbb{E} \sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} |R_k(h_1-h_2)|.$$

Combined with (3.4.7), this gives

$$\sqrt{n}\,\mathbb{E} \sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} G_n(h_1-h_2) \leq C\left(u + \max_{1\leq k\leq n} \mathbb{E} \sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} |Z_k(h_1-h_2)|\right). \qquad (3.4.8)$$

Note that the bracketing number for the class $\mathcal{H}-\mathcal{H}$ satisfies

$$N_{[]}(\mathcal{H}-\mathcal{H}, L_2(P), \varepsilon) \leq N^2_{[]}(\mathcal{H}, L_2(P), \varepsilon/2).$$

Finally, a sharp version of M. Ossiander's bracketing theorem due to M. Talagrand (Theorem 2.7.10 in [80]) combined with Dudley's entropy integral estimate for the generic chaining complexity (Proposition 2.7.10 in [80]) yields

$$\mathbb{E} \sup_{\|h_1-h_2\|_{L_2(\Pi)}\leq u} |Z_k(h_1-h_2)| \leq C \int_0^{2u} \sqrt{1+T_1(u)}du. \qquad (3.4.9)$$

The result now follows from (3.4.8). □

Bracketing entropy is useful when one has to deal, for example, with base classes that are Lipschitz in parameter: assume that $\mathcal{H} = \{h_t,\ t \in I\}$ where I is the index space equipped with distance $\mathcal{D}(\cdot,\cdot)$. If for any t, s in I,

$$|h_s(x) - h_t(x)| \leq \mathcal{D}(s,t)F_1(x).$$

and $\mathbb{E}F_1^2(X) < \infty$, then the bracketing entropy $\log N_{[]}(\varepsilon, \mathcal{H}, L_2(P))$ is bounded by $\log N\left(\frac{\varepsilon}{\|F\|}, I, \mathcal{D}\right)$-the metric entropy of I with respect to $\mathcal{D}$.

3.4.4 Beyond the uniformly bounded base classes

If the envelope F is not uniformly bounded or $\|F\|_\infty$ is prohibitively large, one can use a slightly different approach which only requires that $F(X)$ possesses certain exponential moments.

Let $\Psi := \|F\|_{\psi_2}$, where $\|F\|_{\psi_2} := \inf\left\{C > 0 : \ \mathbb{E}\exp\left(\frac{F^2(X)}{C^2}\right) \le 2\right\}$ (see also Definition 1.1). Moreover, suppose (3.4.5) holds for a suitable function $T(u)$.

Proposition 3.4.4. *If the aforementioned conditions are satisfied, then*

$$w_n(u) \le C \cdot U(u,n) T^{1/2}\left(\frac{U(u,n)}{2\Psi}\right),$$
$$w(u) \le C u T^{1/2}\left(\frac{u}{2\Psi}\right),$$
$$\Omega^{1/2}(u) \le K T^{1/2}\left(\frac{u}{2\Psi}\right)\log\frac{\Psi}{u},$$

where $U(u,n) := \left[u \bigvee C\left(\frac{\Psi\sqrt{T(1/\sqrt{n})\log n}}{n}\right)^{1/2}\right]$.

Proof. The argument follows main steps of the proof of Theorem 3 in [39]. Let $u_n^2 := \sup\limits_{f\in\mathcal{H}(u)} P_n f^2$. By Theorem 1.2.8 and assumptions on T,

$$w_n(u) \le C\mathbb{E}\int\limits_0^{2u_n}\sqrt{T\left(\frac{\varepsilon}{\|F\|_{L_2(P_n)}}\right)}d\varepsilon = C\mathbb{E}\int\limits_0^{2u_n}\sqrt{T\left(\frac{\varepsilon}{\|F\|_{L_2(P_n)}}\right)}d\varepsilon I\left\{\|F\|_{L_2(P_n)} \le 2\Psi\right\}$$
$$+ C\mathbb{E}\int\limits_0^{2u_n}\sqrt{T\left(\frac{\varepsilon}{\|F\|_{L_2(P_n)}}\right)}d\varepsilon I\left\{\|F\|_{L_2(P_n)} > 2\Psi\right\}.$$

Note that $\mathbb{E}F^2(X) \le \Psi^2$ (by elementary properties of the Orlicz norm). By Theorem 1.2.2,

$$P\left(\|F\|^2_{L_2(P_n)} > 4\Psi^2\right) \le 2e^{-cn}.$$

Moreover, making change of variables $u = \frac{\varepsilon}{\|F\|_{L_2(P_n)}}$ we get

$$\int\limits_0^{2u_n}\sqrt{T\left(\frac{\varepsilon}{\|F\|_{L_2(P_n)}}\right)}d\varepsilon \le \|F\|_{L_2(P_n)}\int\limits_0^2\sqrt{T(u)}du \le C(T)\|F\|_{L_2(P_n)}.$$

Now Hölder's inequality implies that the second term in the sum is bounded by

$$C(T)\|F\|_{L_2(P)}\exp(-cn). \tag{3.4.10}$$

To bound the first term, we will first show that

$$\mathbb{E}u_n^2 \le u^2 + C\left[u\Psi^{1/2}\sqrt{\frac{(T(1/\sqrt{n})\log n)^{1/2}}{n}}\bigvee \Psi\frac{(T(1/\sqrt{n})\log n)^{1/2}}{n}\right] := B^2 \tag{3.4.11}$$

Indeed, by symmetrization inequality (Theorem 1.2.6)

$$\mathbb{E}u_n^2 \le u^2 + 2\mathbb{E}\sup_{f\in\mathcal{H}(u)}|R_n(h^2)|.$$

To bound $\mathbb{E}\sup\limits_{f\in\mathcal{H}(u)}|R_n(h^2)|$, we will apply Theorem 3.16 from [53]. It implies that

$$\mathbb{E}\sup_{f\in\mathcal{H}(u)}|R_n(h^2)| \le K\left[u\sqrt{\frac{\Gamma_{n,\infty}(\mathcal{H}(u))}{n}}\bigvee\frac{\Gamma_{n,\infty}(\mathcal{H}(u))}{n}\right],$$

where $\Gamma_{n,\infty}(\mathcal{H}(u))$ is the so-called *generic chaining complexity* that can be bounded by Dudley's entropy integral as follows [80]:

$$\Gamma_{n,\infty}(\mathcal{H}(u)) \le \mathbb{E}\int\limits_0^{2\|F\|_{L_\infty(P_n)}}\sqrt{\log N(\mathcal{H}(u),\varepsilon,L_\infty(P_n))}d\varepsilon.$$

To estimate the latter quantity, note that $N(\mathcal{H}(u),\varepsilon,L_\infty(P_n)) \le N\left(\mathcal{H}(u),\frac{\varepsilon}{\sqrt{n}},L_2(P_n)\right)$, hence

$$\log N(\mathcal{H}(u),\varepsilon,L_\infty(P_n)) \le \sqrt{T\left(\frac{\varepsilon}{\sqrt{n}\|F\|_{L_\infty(P_n)}}\right)}$$

and

$$\Gamma_{n,\infty}(\mathcal{H}(u)) \le \mathbb{E}\int\limits_0^{2\|F\|_{L_\infty(P_n)}}\sqrt{T\left(\frac{\varepsilon}{\sqrt{n}\|F\|_{L_\infty(P_n)}}\right)}d\varepsilon \le$$
$$\le CT^{1/2}\left(\frac{1}{\sqrt{n}}\right)\mathbb{E}\|F\|_{L_\infty(P_n)}.$$

It remains to use the well-known fact that for random random variables $\xi_1,\dots,\xi_n$ with finite ψ_2-norms we have $\mathbb{E}\max_i|\xi_i| \le C\sqrt{\log n}\max_i\|\xi_i\|_{\psi_2}$ (see Lemma 2.2.2 in

[91]). Taking $\xi_i := F(X_i)$ yields (3.4.11).

The rest follows from a simple computation: by Jensen's inequality, we have

$$\mathbb{E}\int_0^{2u_n}\sqrt{T\left(\frac{\varepsilon}{\|F\|_{L_2(P_n)}}\right)}d\varepsilon I\left\{\|F\|_{L_2(P_n)}\le 2\Psi\right\}\le \int_0^{2\mathbb{E}^{1/2}u_n^2}\sqrt{T\left(\frac{\varepsilon}{2\Psi}\right)}d\varepsilon$$
$$\le C\cdot B\,T^{1/2}\left(\frac{B}{2\Psi}\right),$$

which together with (3.4.11) implies the result. □

When class $\mathcal{H}$ is finite or has finite VC-dimension, we have the following result:

Corollary 3.4.5. *Assume that conditions of the previous Proposition hold with (a)* $T_1(u)\equiv \log N$ *or with (b)* $T_2(u)=V\log\frac{K}{u}$. *Then*

(a) $w_n(u)\le C\sqrt{\log N}\left[u\bigvee \Psi^{1/2}\frac{(\log n\log N)^{1/4}}{\sqrt{n}}\right]$;

(b) $w_n(u)\le C\sqrt{V}\cdot U(u,n)\log^{1/2}\frac{2K\Psi}{U(u,n)}$, *where* $U(u,n):=\left[u\bigvee C\sqrt{\frac{\sqrt{V}\Psi\log n}{n}}\right]$.

3.4.5 Existence of solutions

We continue our investigation with a general study of problem (3.3.1) and provide sufficient conditions of existence of a solution. Recall that all densities are assumed to have finite entropies.

Lemma 3.4.6. *The entropy functional is lower semi-continuous in* $L_1(\mu)$ *and Hölder continuous in every bounded subset of* $L_p(\mu)$ *for* $p>1$.

Proof. The functional is lower semi-continuos iff the level sets $\mathcal{L}_t=\{\lambda:\ H(\lambda)\le t\}$ are closed. Suppose $\lambda_n\in\mathcal{L}_t,\ \lambda_n\to\lambda_0$ in L_1. We can extract the subsequence λ_{n_k} converging to $\lambda_0\ \ \mu$ - a.s. Noting that $s\log(s)+e^{-1}\ge 0$ and applying the Fatou lemma to the sequence $\{\lambda_{n_k}\log(\lambda_{n_k})\}$, we derive the result. Note that here we did not use the assumption that μ is a finite measure.

To prove the second part of the claim, we use a different approach that allows to get a stronger result in case of finite measure μ.

Note that for positive t,

$$t \log t = t + \int_0^t \log(x) dx,$$

which implies

$$t_1 \log t_1 - t_2 \log t_2 = (t_1 - t_2) + \int_{t_2}^{t_1} \log(x) dx \tag{3.4.12}$$

The following two elementary inequalities are true for $\alpha, \beta > 0$ and $\gamma > 1$:

$$|\log(t)| \leq C_{\alpha,\beta}(t^\alpha + t^{-\beta}),$$

$$(t+s)^\gamma \leq 2^{\gamma-1}\left(t^\gamma + s^\gamma\right).$$

Let $\alpha = p-1,\ \gamma = \frac{p}{p-1}$. Combined with (3.4.12), this yields to

$$|H(\lambda_2) - H(\lambda_1)| \leq \int_{\mathcal{H}} |\lambda_1 - \lambda_2| d\mu + C_1 \int_{\mathcal{H}} |{\lambda_1}^{1-\beta} - {\lambda_2}^{1-\beta}| d\mu + C_2 \int_{\mathcal{H}} |\lambda_1^p - \lambda_2^p| d\mu.$$

Given $\|\lambda_1 - \lambda_2\|_p = \varepsilon$, we can estimate every term using Hölder's inequality:

$$\int_{\mathcal{H}} |\lambda_1 - \lambda_2| d\mu \leq \mu(\mathcal{H})^{\frac{p-1}{p}} \cdot \varepsilon,$$

$$\int_{\mathcal{H}} |{\lambda_1}^{1-\beta} - {\lambda_2}^{1-\beta}| d\mu \leq \int_{\mathcal{H}} |\lambda_1 - \lambda_2|^{1-\beta} d\mu \leq \mu(\mathcal{H})^{\frac{p-1+\beta}{p}} \cdot \varepsilon^{1-\beta},$$

$$\int_{\mathcal{H}} |\lambda_1^p - \lambda_2^p| d\mu \leq \int_{\mathcal{H}} p(\lambda_1^{p-1} + \lambda_2^{p-1}) |\lambda_1 - \lambda_2| d\mu \leq p 2^{\frac{1}{p}} \left(\|\lambda_1\|_p + \|\lambda_2\|_p\right)^{\frac{p-1}{p}} \cdot \varepsilon.$$

Finally, this implies that for $p > 1$ the entropy functional is Hölder continuous in L_p ball of any finite radius with Hölder exponent less then 1. □

Now we are ready to prove the existence of a solution of (3.3.1), (3.3.2).

Theorem 3.4.7. *Problems (3.3.1), (3.3.2) have unique solutions in every convex weakly compact subset $\mathbb{D}$ of L_p, $p \geq 1$. Moreover, if there exists $\lambda \in \mathbb{D}$ which is positive μ - a.e., the solutions are positive μ-a.e.*

Proof. Let

$$F(\lambda) := P(\ell \bullet f_\lambda) + \varepsilon H(\lambda),\ \lambda \in \mathbb{D}$$

Under the assumptions on the loss, F is convex, bounded from below and lower semi-continuous. This follows from Lemma 3.4.6 for the entropy term; it is easy to see that $P(\ell \bullet f_\lambda)$ is continuous as well: assume $\|\lambda_n - \lambda_0\|_1 \to 0$. Then, denoting $v := \lambda_n - \lambda_0$,

$$|P(\ell \bullet f_{\lambda_n}) - P(\ell \bullet f_{\lambda_0})| = |P(\ell' \bullet f_{\lambda_0 + \tau v})(f_{\lambda_n} - f_{\lambda_0})| \leq$$
$$\leq CP\,|f_{\lambda_n} - f_{\lambda_0}| \leq C\|\lambda_n - \lambda_0\|_1 \to 0,$$

where we used uniform boundedness of the dictionary and convexity of ℓ with respect to the second variable.

Now we can conclude that the level sets $\mathcal{L}_t = \{\lambda :\ F(\lambda) \leq t\}$ are closed and convex. Mazur's theorem (see [61], Theorem 2.1) implies that they are also closed in weak topology, so F is weakly lower semi-continuos.

Given a minimizing sequence λ_n, we can extract a weakly convergent subsequence

$$\lambda_{n_k} \xrightarrow{\sigma} \lambda_\infty,$$

and conclude that $\lambda_\infty \in \mathbb{D},\ -\infty < F(\lambda_\infty) \leq \liminf_{k\to\infty} F\left(\lambda_{n_k}\right).$

Convexity of the set $\mathbb{D}$ and strict convexity of the functional F implies the uniqueness of the solution of (3.3.1). Replacing P by P_n, we get similar statements for (3.3.2).

It remains to prove the last claim. Suppose that $\lambda_\varepsilon = \operatorname{argmin}_{\lambda\in\mathbb{D}} F(\lambda)$ is such that for some $A \subset \mathcal{H}$ with $\mu(A) > 0$

$$\lambda_\varepsilon(h) I_A(h) \equiv 0$$

where I_A stands for the indicator function of the set A. Take $\lambda \in \mathbb{D}$ which is positive

μ - a.e. and consider

$$\frac{d}{dt}_{|t=\tau} \int_{\mathcal{H}} (\lambda_\varepsilon + t(\lambda - \lambda_\varepsilon)) \log(\lambda_\varepsilon + t(\lambda - \lambda_\varepsilon)) d\mu =$$
$$= \int_{\mathcal{H}} \frac{d}{dt}_{|t=\tau} (\lambda_\varepsilon + t(\lambda - \lambda_\varepsilon)) \log(\lambda_\varepsilon + t(\lambda - \lambda_\varepsilon)) d\mu =$$
$$= \int_{\mathcal{H}} (\lambda - \lambda_\varepsilon) \log(\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon)) d\mu,$$

where the change of order of differentiation and integration is correct due to proposition 3.4.8 below. Let t_n be a monotone sequence with $t_0 < 1,\ t_n \to 0$. The function

$$(0, t_0] \ni \tau \mapsto (\lambda - \lambda_\varepsilon) \log(\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon))$$

is non-decreasing and $\int\limits_{\mathcal{H}} (\lambda - \lambda_\varepsilon) \log(\lambda_\varepsilon + t_0(\lambda - \lambda_\varepsilon)) d\mu < \infty$(by Proposition 3.4.8), hence, by the monotone convergence theorem and our assumption on the set A,

$$\int_{\mathcal{H}} (\lambda - \lambda_\varepsilon) \log(\lambda_\varepsilon + t_n(\lambda - \lambda_\varepsilon)) d\mu \to -\infty \text{ as } n \to \infty.$$

It remains to use the Mean Value theorem to conclude that there exists $\tau_0 > 0$ such that $F(\lambda_\varepsilon + \tau_0(\lambda - \lambda_\varepsilon)) < F(\lambda_\varepsilon)$ leading to contradiction. □

Typically, $\mathbb{D}$ would be a convex uniformly integrable subset of $L_1(\mu)$. Uniform integrability holds in particular when

$$\sup_{\lambda \in \mathbb{D}} \int |\lambda \log \lambda| d\mu < \infty,$$

(this is just the application of the well-known criterion of de la Vallee Poussin, see [14]). Another common example is the intersection of some $L_p(\mu)$-ball for $p > 1$ with the cone of probability densities, where μ is a finite measure.

3.4.6 Differentiability of the risk and of the entropy

To derive necessary conditions of the minima in the optimization problems (3.3.1), (3.3.2), we have to study differentiability properties of the functions involved in these

problems. For $F : \mathbb{D} \mapsto \mathbb{R}$, $\lambda \in \mathbb{D}$ and ν such that $\bar{\lambda} := \lambda + t_0\nu \in \mathbb{D}$ for some $t_0 > 0$, denote

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

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

First note that, under our assumptions on the loss function ℓ, both the true risk

$$\mathbb{D} \ni \lambda \mapsto P(\ell \bullet \lambda) =: L(\lambda)$$

and the empirical risk

$$\mathbb{D} \ni \lambda \mapsto P(\ell \bullet \lambda) := L_n(\lambda)$$

have directional derivatives at any point $\lambda \in \mathbb{D}$ in the direction of any other point $\bar{\lambda} = \lambda + t_0\nu \in \mathbb{D}, t_0 > 0$. Moreover, the following formulas hold:

$$DL(\lambda, \nu) = P(\ell' \bullet f_\lambda) f_\nu \tag{3.4.13}$$

and

$$DL_n(\lambda, \nu) = P_n(\ell' \bullet f_\lambda) f_\nu. \tag{3.4.14}$$

Indeed, this is directly implied by our assumptions on the uniform boundedness of the base class and differentiability of the loss function.

Proposition 3.4.8. *If $\bar{\lambda} \log \lambda \in L_1(\mu)$, where $\bar{\lambda} = \lambda + t_0\nu \in \mathbb{D}, t_0 > 0$, then the directional derivative $DH(\lambda; \nu)$ exists and*

$$DH(\lambda; \nu) = \int_{\mathcal{H}} \log(\lambda)\ \nu d\mu. \tag{3.4.15}$$

In particular, this is true if, for some $t_0 > 0$, $\bar{\lambda} = \lambda + t_0\nu \in \mathbb{D}$ and $\bar{\lambda}_1 = \lambda - t_0\nu \in \mathbb{D}$.

Proof. Due to the convexity of $\lambda \log \lambda$, the function

$$[0, t_0] \ni t \mapsto (\lambda + t\nu) \log(\lambda + t\nu)$$

is also convex. Therefore,

$$[0, t_0] \ni t \mapsto \frac{(\lambda + t\nu)\log(\lambda + t\nu) - \lambda \log \lambda}{t}$$

is a nondecreasing function. Take a decreasing sequence $\{t_n\}_{n\geq 0}$, $t_n \to 0$ as $n \to \infty$ and consider

$$\frac{H(\lambda + t_n\nu) - H(\lambda)}{t_n} - \int_{\mathcal{H}} \log(\lambda)\nu d\mu = \int_{\mathcal{H}} \left[\frac{(\lambda + t_n\nu)\log(\lambda + t_n\nu) - \lambda \log \lambda}{t_n} - (\log \lambda + 1)\nu \right] d\mu. \tag{3.4.16}$$

The sequence of integrands in the right hand side monotonically decreases to 0. Moreover, for $n = 0$, the integrand is integrable under the assumption $\bar{\lambda} \log \lambda \in L_1(\mu)$, and the first claim follows by monotone convergence.

If, for some $t_0 > 0$, $\bar{\lambda} = \lambda + t_0\nu \in \mathbb{D}$ and $\bar{\lambda}_1 = \lambda - t_0\nu \in \mathbb{D}$, then $\lambda = (\bar{\lambda} + \bar{\lambda}_1)/2$. Since $\lambda \log \lambda \in L_1(\mu)$ and $\bar{\lambda}, \bar{\lambda}_1$ are nonnegative functions, this easily implies that both $\bar{\lambda} \log \lambda \in L_1(\mu)$ and $\bar{\lambda}_1 \log \lambda \in L_1(\mu)$, and the last claim follows. □

3.4.7 Symmetrized Kullback-Leibler distance

For two densities $\lambda_1, \lambda_2 \in \mathbb{D}$, denote

$$K(\lambda_1|\lambda_2) := \int_{\mathcal{H}} \log \frac{\lambda_1}{\lambda_2} \lambda_1 d\mu$$

the Kullback-Leibler divergence between λ_1 and λ_2 and let

$$K(\lambda_1, \lambda_2) := K(\lambda_1|\lambda_2) + K(\lambda_2|\lambda_1)$$

be the symmetrized Lullback-Leibler distance. It is easy to check that

$$K(\lambda_1, \lambda_2) = \int_{\mathcal{H}} \log \frac{\lambda_1}{\lambda_2} (\lambda_1 - \lambda_2) d\mu.$$

We will also need the following proposition.

Proposition 3.4.9. *For all* $\lambda_1, \lambda_2 \in \mathbb{D}$,

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

Proof. Note that the function

$$[0,1] \ni t \mapsto ((1-t)\lambda_1 + t\lambda_2)\log((1-t)\lambda_1 + t\lambda_2)$$

is convex and, hence, its derivative

$$[0,1] \ni t \mapsto (\log((1-t)\lambda_1 + t\lambda_2) + 1)(\lambda_2 - \lambda_1)$$

is nondecreasing. Similarly, the function

$$[0,1] \ni t \mapsto (\log(t\lambda_1 + (1-t)\lambda_2) + 1)(\lambda_1 - \lambda_2)$$

is also nondecreasing. Therefore, the function

$$[0,1] \ni t \mapsto \log\frac{(1-t)\lambda_1 + t\lambda_2}{t\lambda_1 + (1-t)\lambda_2}(\lambda_1 - \lambda_2)$$

is nonincreasing. Take a sequence $t_n \in (0,1)$ such that t_n decreases monotonically to 0. Then the following sequence of functions is nondecreasing

$$\left\{\log\frac{(1-t_n)\lambda_1 + t_n\lambda_2}{t_n\lambda_1 + (1-t_n)\lambda_2}(\lambda_1 - \lambda_2)\right\}_{n\geq 1}$$

and it converges as $n \to \infty$ to the function $\log\frac{\lambda_1}{\lambda_2}(\lambda_1 - \lambda_2)$, which is nonnegative. Note also that for all n

$$\log\frac{(1-t_n)\lambda_1 + t_n\lambda_2}{t_n\lambda_1 + (1-t_n)\lambda_2}(\lambda_1 - \lambda_2) \in L_1(\mu).$$

Indeed,

$$((1-t_n)\lambda_1 + t_n\lambda_2)\log((1-t_n)\lambda_1 + t_n\lambda_2) \in L_1(\mu)$$

and, together with the fact that $\lambda_1, \lambda_2 \geq 0$, this implies that

$$\lambda_1\log((1-t_n)\lambda_1 + t_n\lambda_2) \in L_1(\mu),\ \lambda_2\log((1-t_n)\lambda_1 + t_n\lambda_2) \in L_1(\mu)$$

and similarly

$$\lambda_1\log(t_n\lambda_1 + (1-t_n)\lambda_2) \in L_1(\mu),\ \lambda_2\log(t_n\lambda_1 + (1-t_n)\lambda_2) \in L_1(\mu).$$

As a result, it is easy to conclude that

$$\lim_{n\to\infty}\int_{\mathcal{H}}\log\frac{(1-t_n)\lambda_1+t_n\lambda_2}{t_n\lambda_1+(1-t_n)\lambda_2}(\lambda_1-\lambda_2)d\mu=\int_{\mathcal{H}}\log\frac{\lambda_1}{\lambda_2}(\lambda_1-\lambda_2)d\mu=K(\lambda_1,\lambda_2).$$

□

Note that, in principle, the distance $K(\lambda_1,\lambda_2)$ can be infinite for some $\lambda_1,\lambda_2\in\mathbb{D}$. Lemma below provides the bounds showing that given a sparse density λ_1, any other density λ_2 such that $K(\lambda_1,\lambda_2)$ is small will follow the sparsity pattern of λ_1.

Lemma 3.4.10. *Let λ_1,λ_2 be two densities from $\mathbb{D}$ and Λ_1,Λ_2 the corresponding probability measures on $\mathcal{H}$. Then for any measurable $\mathcal{H}'\subset\mathcal{H}$*

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

Proof. By the well-known inequality between the Kullback-Leibler and Hellinger distances, for any $\mathcal{H}'\subset\mathcal{H}$

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

□

3.5 Main results for prediction problems

In this Section, we will obtain the bounds for approximation error and probabilistic estimates for the random error which together imply one of our main results – an oracle inequality for performance of $\hat{\lambda}_\varepsilon$ (see Corollary 3.5.6). We will apply similar techniques to get some results for the problem of density estimation in Section 3.6 (see Corollary 3.6.3).

3.5.1 Approximation error bounds

In this section, we study the properties of the solution λ_ε of problem (3.3.1). Namely, we are interested in the size of its excess risk $\mathcal{E}(f_{\lambda_\varepsilon})$ comparing with the excess risk $\mathcal{E}(f_\lambda)$ of *oracle solutions* $\lambda \in \mathbb{D}$. We will introduce a notion of *alignment coefficient* of an oracle λ with the dictionary $\mathcal{H}$. It turns out that, in special examples, this quantity is related to the degree of "sparsity" of λ as well as to its "regularity" in a proper sense. For oracles $\lambda \in \mathbb{D}$ that are "alligned" with the dictionary reasonably well (so that the alignment coefficient is not large) the size of excess risk $\mathcal{E}(f_{\lambda_\varepsilon})$ is controlled by the size of $\mathcal{E}(f_\lambda)$ up to an error term of the order ε^2. Similarly, the square of the $L_2(\Pi)$-approximation error $\|f_{\lambda_\varepsilon} - f_*\|^2_{L_2(\Pi)}$ is controlled by the same error for the oracle $\|f_\lambda - f_*\|^2_{L_2(\Pi)}$ up to an additional error term of the order ε^2.

For $w \in L_2(\mu)$, define *the alignment coefficient* $\gamma(w)$ to be

$$\gamma(w) := \sup\left\{\langle w, u\rangle_{L_2(\mu)} \ : \ \|f_u\|_{L_2(\Pi)} = 1, \langle u, 1\rangle_{L_2(\mu)} = 0\right\}.$$

It is easy to see that, for all constants $c \in \mathbb{R}$, $\gamma(w + c) = \gamma(w)$.

Similar quantities have been already used in the analysis of approximation error in the case of finite dictionaries in [49],[51] (actually, in this special case even more sophisticated definitions have been used that better take into account the geometry of sparse recovery problems). We will define *the Gram operator* of the dictionary as an integral operator $K : L_2(\mathcal{H}, \mu) \mapsto L_2(\mathcal{H}, \mu)$,

$$(Ku)(h) = \int\limits_{\mathcal{H}} \langle h, h'\rangle_{L_2(\Pi)} u(h')\mu(dh').$$

This is a bounded symmetric nonnegatively definite operator (at least, when μ is a finite measure), its square root is well defined and we have

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

Therefore,

$$\gamma(w) = \sup\left\{|\langle w, u\rangle_{L_2(\mu)}| \ : \ \|K^{\frac{1}{2}}u\|_{L_2(\mu)} = 1\right\}.$$

The last quantity often coincides with $\|K^{-\frac{1}{2}}w\|_{L_2(\mu)}$ for $w \in \text{Im } K^{1/2}$. Moreover, we can always formally define

$$\|K^{-\frac{1}{2}}w\|_{L_2(\mu)} := \inf\left\{\|u\|_{L_2(\mu)} : \ K^{\frac{1}{2}}u = w\right\}$$

when $w \in \text{Im } K^{\frac{1}{2}}$ and set it equal to infinity otherwise. Then we can write $\gamma(w) = \|K^{-1/2}w\|_{L_2(\mu)}$.

Remark: sometimes it might be more convenient to use a slightly different version of the alignment coefficient

$$\bar{\gamma}(w) := \sup\left\{\langle w, u\rangle_{L_2(\mu)} \ : \ \text{Var}(f_u(X)) = 1, \langle u, 1\rangle_{L_2(\mu)} = 0\right\},$$

along with the *covariance operator*

$$(\bar{K}u)(h) := \int_{\mathcal{H}} \text{cov}_{\Pi}(h, g)u(g)\mu(dg), \ h \in \mathcal{H},$$

where $\text{cov}_{\Pi}(h, g) := \Pi(hg) - \Pi(h)\Pi(g)$. Note that $\bar{\gamma}(w) \geq \gamma(w)$.

If the dictionary $\mathcal{H} = \{h_1, \ldots, h_N\}$ is finite, the Gram operator is represented by the Gram matrix, which in the simplest case of orthonormal dictionary is equal to identity matrix. In this case, we have $\gamma(w) = \|w\|_{\ell_2^N}$ and in the case of vectors w of small support there is a clear relationship between the size of $\gamma(w)$ and the degree of "sparsity" of the vector w. If the dictionary is not orthonormal, the size of the alignment coefficient $\gamma(w)$ is a measure of "alignment" of w with eigenspaces of K : roughly, if w belongs to the linear span of eigenspaces corresponding to the large eigenvalues of K, $\gamma(w)$ is not too large. In many examples of infinite dictionaries, the space of oracles for which the alignment coefficient is finite could be identified with the space of functions with some regularity properties, such as Sobolev space $\mathbb{H}^s := \mathbb{W}^{2,s}$, so that the alignment coefficient is bounded by a Sobolev-type norm of the corresponding function. If the function w consists of several well separated smooth "spikes", it happens that the size of $\gamma(w)$ can be controlled in terms of the number of "spikes" and, in this sense, it is related to sparsity.

Below we will be interested in those oracles $\lambda \in \Lambda$ for which $\gamma(\log \lambda)$ is not too large.

Theorem 3.5.1. *There exists a constant $C > 0$ depending only on the loss such that for all oracles $\lambda \in \mathbb{D}$*

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

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

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

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

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

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

The first bound of the theorem means that the true penalized solution λ_ε belongs to the Kullback-Leibler "ball" around arbitrary oracle $\lambda \in \mathbb{D}$ and, at the same time, the function f_{λ_ε} belongs to the $L_2(\Pi)$-ball around f_λ, the radius of both balls being, up to a constant, the maximum of the $L_2(\Pi)$-distance from f_λ to the target function f_* and $\gamma(\log \lambda)\varepsilon$. The second bound easily implies that

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

(the constant 2 in front of $\mathcal{E}(f_\lambda)$ can be replaced by $1 + \delta$, but the constant C then becomes of the order $1/\delta$). The last two bounds show that the solution λ_ε and oracles λ are concentrated on almost the same sets (for the oracles that approximate the target and are aligned with the dictionary reasonably well).

Proof. Denote

$$F(\lambda) := P(\ell \bullet f_\lambda) + \varepsilon H(\lambda),\ \lambda \in \mathbb{D}.$$

It follows from (3.4.13), (3.4.14) and Proposition 3.4.8 that, for all $\lambda \in \mathbb{D}$ and $\tau \in (0,1)$, the directional derivative of F exists at the point $\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon)$ in the direction $\lambda - \lambda_\varepsilon$ and

$$DF(\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon); \lambda - \lambda_\varepsilon) =$$
$$P(\ell' \bullet f_{\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon)})(f_\lambda - f_{\lambda_\varepsilon}) + \varepsilon \int\limits_{\mathcal{H}} (\lambda - \lambda_\varepsilon) \log(\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon)) d\mu. \quad (3.5.1)$$

Moreover, since the function $[0,1] \ni \tau \mapsto F(\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon))$ is convex, its right derivative, which coincides with $DF(\lambda + \tau(\lambda - \lambda_\varepsilon); \lambda - \lambda_\varepsilon)$, is non-decreasing in $\tau \in [0,1]$. Since λ_ε is the minimum point of F, this implies that, for $\tau \in (0,1)$,

$$DF(\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon); \lambda - \lambda_\varepsilon) = \quad (3.5.2)$$
$$= P(\ell' \bullet f_{\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon)})(f_\lambda - f_{\lambda_\varepsilon}) + \varepsilon \int\limits_{\mathcal{H}} (\lambda - \lambda_\varepsilon) \log(\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon)) d\mu \geq 0.$$

We will subtract both sides of the last inequality from the expression

$$DF(\lambda + \tau(\lambda_\varepsilon - \lambda); \lambda - \lambda_\varepsilon) =$$
$$= P(\ell' \bullet f_{\lambda + \tau(\lambda_\varepsilon - \lambda)})(f_\lambda - f_{\lambda_\varepsilon}) + \varepsilon \int\limits_{\mathcal{H}} (\lambda - \lambda_\varepsilon) \log(\lambda + \tau(\lambda_\varepsilon - \lambda)) d\mu \quad (3.5.3)$$

to get

$$P\left(\ell' \bullet f_{\lambda + \tau(\lambda_\varepsilon - \lambda)} - \ell' \bullet f_{\lambda_\varepsilon + \tau(\lambda - \lambda_\varepsilon)}\right)(f_\lambda - f_{\lambda_\varepsilon}) + \varepsilon \int (\lambda - \lambda_\varepsilon) \log \frac{(1-\tau)\lambda + \tau\lambda_\varepsilon}{(1-\tau)\lambda_\varepsilon + \tau\lambda} d\mu \quad (3.5.4)$$
$$\leq P(\ell' \bullet f_{\lambda + \tau(\lambda_\varepsilon - \lambda)})(f_\lambda - f_{\lambda_\varepsilon}) + \varepsilon \int\limits_{\mathcal{H}} (\lambda - \lambda_\varepsilon) \log(\lambda + \tau(\lambda_\varepsilon - \lambda)) d\mu.$$

Under the assumptions on the loss, in particular, continuity of ℓ', passing to the limit as $\tau \to 0$, using the dominated convergence and Proposition 3.4.9, we get that the left hand side of (3.5.4) converges to

$$P(\ell' \bullet f_\lambda - \ell' \bullet f_{\lambda_\varepsilon})(f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}) + \varepsilon K(\lambda, \lambda_\varepsilon)$$

and the first term in the right hand side converges to $P(\ell' \bullet f_\lambda)(f_\lambda - f_{\lambda_\varepsilon})$. As for the second term in the right hand side of (3.5.4), note that

$$-(\lambda - \lambda_\varepsilon)(\log(\lambda + \tau(\lambda_\varepsilon - \lambda)) + 1) = \frac{d}{d\tau}(\lambda + \tau(\lambda_\varepsilon - \lambda))\log(\lambda + \tau(\lambda_\varepsilon - \lambda))$$

is a nondecreasing function of $\tau \in [0,1]$ (since it is the derivative of a convex function). Because of this, the second term in the right hand side of (3.5.4) is upper bounded by the integral

$$\int\limits_{\mathcal{H}} \log\lambda(\lambda - \lambda_\varepsilon)d\mu.$$

As a result, we get the following bound:

$$\begin{aligned} &P(\ell' \bullet f_\lambda - \ell' \bullet f_{\lambda_\varepsilon})(f_{\hat\lambda_\varepsilon} - f_{\lambda_\varepsilon}) + \varepsilon K(\lambda, \lambda_\varepsilon) \le \\ &P(\ell' \bullet f_\lambda)(f_\lambda - f_{\lambda_\varepsilon}) + \varepsilon \int\limits_{\mathcal{H}} \log\lambda(\lambda - \lambda_\varepsilon)d\mu \end{aligned} \tag{3.5.5}$$

Since ℓ is a loss of quadratic type, we have with some constants $C, c > 0$ depending only on ℓ that

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

and also

$$P(\ell' \bullet f_\lambda)(f_\lambda - f_{\lambda_\varepsilon}) = P((\ell' \bullet f_\lambda) - (\ell' \bullet f_*))(f_\lambda - f_{\lambda_\varepsilon}) =$$

$$\left\langle (\ell' \bullet f_\lambda) - (\ell' \bullet f_*), f_\lambda - f_{\lambda_\varepsilon} \right\rangle_{L_2(P)} \le C\|f_\lambda - f_*\|_{L_2(\Pi)}\|f_{\lambda_\varepsilon} - f_\lambda\|_{L_2(\Pi)}.$$

The definition of $\gamma(\log\lambda)$ implies that

$$\int\limits_{\mathcal{H}} \log\lambda(\lambda - \lambda_\varepsilon) \le \gamma(\log\lambda)\|f_{\lambda_\varepsilon} - f_\lambda\|_{L_2(\Pi)}.$$

Thus, it follows from (3.5.5) that with some constant $C > 0$ depending only on ℓ

$$\begin{aligned} &\|f_\lambda - f_{\lambda_\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon K(\lambda, \lambda_\varepsilon) \le \\ &C\left[\|f_\lambda - f_*\|_{L_2(\Pi)}\|f_{\lambda_\varepsilon} - f_\lambda\|_{L_2(\Pi)} + \varepsilon\gamma(\log\lambda)\|f_{\lambda_\varepsilon} - f_\lambda\|_{L_2(\Pi)}\right]. \end{aligned} \tag{3.5.6}$$

To obtain the first bound of the theorem it is enough to upper bound solutions of the resulting inequality with respect to $\|f_{\lambda_\varepsilon} - f_\lambda\|_{L_2(\Pi)}$.

To prove the second bound, note that by the definition of λ_ε, for all $\lambda \in \mathbb{D}$,

$$\mathcal{E}(f_{\lambda_\varepsilon}) + \varepsilon \int_{\mathcal{H}} \lambda_\varepsilon \log \lambda_\varepsilon d\mu \le \mathcal{E}(f_\lambda) + \varepsilon \int_{\mathcal{H}} \lambda \log \lambda d\mu,$$

which implies

$$\mathcal{E}(f_{\lambda_\varepsilon}) \le \mathcal{E}(f_\lambda) + \varepsilon \int_{\mathcal{H}} (\lambda \log \lambda - \lambda_\varepsilon \log \lambda_\varepsilon) d\mu \le$$

$$\mathcal{E}(f_\lambda) + \varepsilon \int_{\mathcal{H}} \log \lambda (\lambda - \lambda_\varepsilon) d\mu \le \mathcal{E}(f_\lambda) + \varepsilon \gamma(\log \lambda) \|f_\lambda - f_{\lambda_\varepsilon}\|_{L_2(\Pi)}, \quad (3.5.7)$$

and it is enough now to use the first bound on $\|f_\lambda - f_{\lambda_\varepsilon}\|_{L_2(\Pi)}$.

The last two inequalities follow from the bound on $K(\lambda, \lambda_\varepsilon)$ and Lemma 3.4.10. $\square$

We refer the reader to Section 3.7 for some common examples of base classes and expressions of the associated alignment coefficient.

3.5.2 Random error bounds

The purpose of this section is to develop exponential bounds on the random error

$$\left|\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) - \mathcal{E}(f_{\lambda_\varepsilon})\right|$$

that depend on the "approximate sparsity" of the true penalized solution λ_ε. Since we are dealing with a loss ℓ of quadratic type, bounding random error is essentially equivalent to bounding the norm $\|f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}\|_{L_2(\Pi)}$. At the same time, we provide upper bounds on the symmetrized Kullback-Leibler distance between $\hat{\lambda}_\varepsilon$ and λ_ε and show that the "approximate sparsity" of each of them is closely related to the "approximate sparsity" of another one.

We assumed that $\mathbb{D}$ is a convex set of probability densities such that $\lambda \log \lambda \in L_1(\mu),\ \lambda \in \mathbb{D}$, and solutions of the problems (3.3.1), (3.3.2) exist in $\mathbb{D}$ (see Theorem 3.4.7 for sufficient conditions of the existence of solutions).

Let $\mathcal{H}'$ be a measurable subset of $\mathcal{H}$. In the theorem below, it will be a subset of the base class (of the dictionary) $\mathcal{H}$ on which both $\hat{\lambda}_\varepsilon$ and λ_ε are approximately concentrated. Let L be a finite dimensional subspace of $L_2(\Pi)$ that will be used to approximate the functions from $\mathcal{H}'$. Let $d := \dim(L)$ and denote

$$U_L(x) := \sup_{g \in L, \|g\|_{L_2(\Pi)} \le 1} |g(x)|.$$

It is easy to check that $\|U_L\|_{L_2(\Pi)} = \sqrt{d}$. Indeed, if $\phi_1, \dots, \phi_d$ is an $L_2(\Pi)$ orthonormal basis of L, then

$$U_L(x) := \sup\left\{ \left| \sum_{k=1}^d c_k \phi_k(x) \right| : \sum_{k=1}^d c_k^2 \le 1 \right\} = \left(\sum_{k=1}^d \phi_k^2(x) \right)^{1/2},$$

which immediately implies that $\|U_L\|^2_{L_2(\Pi)} = d$. We will also use the following quantity:

$$U(L) := \|U_L\|_{L_\infty} + 1.$$

Note that $U(L)$ is of the order $\sqrt{d}$ if there exists an orthonormal basis $\phi_1, \dots, \phi_d$ of L such that the functions ϕ_j are uniformly bounded by a constant. Finally, denote

$$\rho(\mathcal{H}'; L) := \sup_{h \in \mathcal{H}'} \|P_{L^\perp} h\|_{L_2(\Pi)},$$

where $P_{L^\perp}$ stands for the orthogonal projection on $L^\perp$. We are interested in those subspaces L for which d and $U(L)$ are not very large and $\rho(\mathcal{H}'; L)$ is small enough, i.e., the space L provides a reasonably good $L_2(\Pi)$-approximation of the functions from $\mathcal{H}'$. A natural choice of L might be a subspace spanned by the centers of the balls of small enough radius δ covering $\mathcal{H}'$; in this case $\rho(\mathcal{H}'; L) = \delta$ and d is equal to the cardinality of such a δ-covering.

We do not try to get the exact values of the constants in the inequalities below. Moreover, such constants as C might have different values in different parts of the proof (although, its value always depends only on the loss function ℓ).

We are now ready to present the main results. For brevity, we will denote

$$\mathcal{Q}(s) := s\sqrt{\Omega(s/\sqrt{d})} \bigvee w_n(s/\sqrt{d}), \ s \in (0, 1].$$

Theorem 3.5.2. *Suppose that assumption on the dictionary (3.4.4) holds. There exist constants $C, D > 0$ depending only on ℓ such that for all measurable subsets $\mathcal{H}' \subset \mathcal{H}$, for all finite dimensional subspaces $L \subset L_2(\Pi)$ with $d := \dim(L)$ and*

$$\rho := \rho(\mathcal{H}'; L),$$

for all

$$\varepsilon \geq D \cdot \frac{\mathcal{Q}(1)}{\sqrt{n}}$$

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

$$\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \leq C \left[\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \bigvee \frac{d + t_n}{n\varepsilon} \bigvee \frac{1}{\varepsilon} \frac{\mathcal{Q}(\rho)}{\sqrt{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d}) + t_n}{n\varepsilon} \right], \quad (3.5.8)$$

$$\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \leq C \left[\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \bigvee \frac{d + t_n}{n\varepsilon} \bigvee \frac{1}{\varepsilon} \frac{\mathcal{Q}(\rho)}{\sqrt{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d}) + t_n}{n\varepsilon} \right] \quad (3.5.9)$$

and

$$c\|f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon K(\hat{\lambda}_\varepsilon, \lambda_\varepsilon) \leq C \Bigg[\frac{d + t_n}{n} \bigvee \frac{\mathcal{Q}(\rho)}{\sqrt{n}} \bigvee \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \frac{\mathcal{Q}(1)}{\sqrt{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d}) + t_n}{n} \Bigg], \quad (3.5.10)$$

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

Proof. Let λ_ε be the solution of (3.3.1) and $\hat{\lambda}_\varepsilon$ be the solution of (3.3.2). Denote

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

Also denote

$$F(\lambda) := P(\ell \bullet f_\lambda) + \varepsilon H(\lambda), \ \lambda \in \mathbb{D}$$

and

$$F_n(\lambda) := P_n(\ell \bullet f_\lambda) + \varepsilon H(\lambda), \ \lambda \in \mathbb{D}.$$

It follows from (3.4.13), (3.4.14) and Proposition 3.4.8 that, for all $\tau \in (0, 1)$, the directional derivative of F exists at the point $\lambda_\varepsilon + \tau(\hat{\lambda}_\varepsilon - \lambda_\varepsilon)$ in the direction $\hat{\lambda}_\varepsilon - \lambda_\varepsilon$

and

$$DF(\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon);\hat\lambda_\varepsilon-\lambda_\varepsilon)= \tag{3.5.11}$$
$$P(\ell'\bullet f_{\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon)})(f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon})+\varepsilon\int\limits_{\mathcal{H}}(\hat\lambda_\varepsilon-\lambda_\varepsilon)\log(\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon))d\mu.$$

Moreover, since the function $[0,1]\ni\tau\mapsto F(\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon))$ is convex, its right derivative, which coincides with $DF(\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon);\hat\lambda_\varepsilon-\lambda_\varepsilon)$, is nondecreasing in $\tau\in[0,1]$. Since λ_ε is the minimum point of F, this implies that, for $\tau\in(0,1)$,

$$DF(\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon);\hat\lambda_\varepsilon-\lambda_\varepsilon)=P(\ell'\bullet f_{\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon)})(f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon})+$$
$$+\varepsilon\int\limits_{\mathcal{H}}(\hat\lambda_\varepsilon-\lambda_\varepsilon)\log(\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon))d\mu\geq 0. \tag{3.5.12}$$

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

$$DF_n(\hat\lambda_\varepsilon+\tau(\lambda_\varepsilon-\hat\lambda_\varepsilon);\hat\lambda_\varepsilon-\lambda_\varepsilon)= \tag{3.5.13}$$
$$P_n(\ell'\bullet f_{\hat\lambda_\varepsilon+\tau(\lambda_\varepsilon-\hat\lambda_\varepsilon)})(f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon})+\varepsilon\int\limits_{\mathcal{H}}(\hat\lambda_\varepsilon-\lambda_\varepsilon)\log(\hat\lambda_\varepsilon+\tau(\lambda_\varepsilon-\hat\lambda_\varepsilon))d\mu\leq 0.$$

Subtracting (3.5.12) from (3.5.13) and rearranging the terms, we get

$$P\left(\ell'\bullet f_{\hat\lambda_\varepsilon+\tau(\lambda_\varepsilon-\hat\lambda_\varepsilon)}-\ell'\bullet f_{\lambda_\varepsilon+\tau(\hat\lambda_\varepsilon-\lambda_\varepsilon)}\right)(f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon})+ \tag{3.5.14}$$
$$+\varepsilon\int\left(\hat\lambda_\varepsilon-\lambda_\varepsilon\right)\log\frac{(1-\tau)\hat\lambda_\varepsilon+\tau\lambda_\varepsilon}{(1-\tau)\lambda_\varepsilon+\tau\hat\lambda_\varepsilon}d\mu\leq$$
$$\leq\left|(P-P_n)(\ell'\bullet f_{\hat\lambda_\varepsilon+\tau(\lambda_\varepsilon-\hat\lambda_\varepsilon)})(f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon})\right|.$$

Under the assumptions on the loss, in particular, continuity of ℓ', passing to the limit as $\tau\to 0$, using the dominated convergence and Proposition 3.4.9, we get

$$P(\ell'\bullet f_{\hat\lambda_\varepsilon}-\ell'\bullet f_{\lambda_\varepsilon})(f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon})+ \tag{3.5.15}$$
$$+\varepsilon K(\hat\lambda_\varepsilon,\lambda_\varepsilon)\leq\left|(P-P_n)(\ell'\bullet f_{\hat\lambda_\varepsilon})(f_{\hat\lambda_\varepsilon}-f_{\lambda_\varepsilon})\right|.$$

Since ℓ is a loss of quadratic type,

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

Our next step is to extract some information about the sparsity of $\hat{\lambda}_\varepsilon$ from these bounds. To this end, we use Lemma 3.4.10 which implies that for all $\mathcal{H}' \subset \mathcal{H}$

$$\varepsilon\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \le 2\varepsilon\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') + \varepsilon K(\hat{\lambda}_\varepsilon, \lambda_\varepsilon) \tag{3.5.16}$$

and

$$\varepsilon\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \le 2\varepsilon\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}') + \varepsilon K(\hat{\lambda}_\varepsilon, \lambda_\varepsilon). \tag{3.5.17}$$

To complete the proof of the theorem, it remains to bound

$$\left|(P - P_n)(\ell' \bullet f_{\hat{\lambda}_\varepsilon})(f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon})\right|.$$

Let

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

and

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

We need the following two lemmas.

Lemma 3.5.3. *Let $\mathcal{H}$ be a class of functions on S uniformly bounded by 1 and let $L \subset L_2(\Pi)$ be a finite dimensional subspace with $d := \dim(L)$. Denote*

$$\rho := \rho(\mathcal{H}; L) := \sup_{h\in\mathcal{H}} \|P_{L^\perp} h\|_{L_2(\Pi)}.$$

Suppose that assumption (3.4.5) holds. Then with some constant $C > 0$

$$\mathbb{E}\sup_{h\in\mathcal{H}} |R_n(P_{L^\perp} h)| \le C\left[\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d})}{n} \bigvee \sqrt{\frac{1}{n}} w_n(\rho/\sqrt{d})\right].$$

Proof. First we will show that condition (3.4.4) implies the bound for the metric entropy $H(\mathcal{H}, L_2(\Pi), \delta)$. Indeed, since $\sqrt{n}G_n(h) \xrightarrow{\mathcal{L}} \mathcal{W}_\Pi$, where $\mathcal{W}_\Pi$ is the isonormal Gaussian process, we have that

$$\liminf_{n\to\infty} P\left(\sup_{\|h_1-h_2\|_{L_2(\Pi)}\le u} |G_n(h_1 - h_2)| > \varepsilon\right) \ge$$
$$P\left(\sup_{\|h_1-h_2\|_{L_2(\Pi)}\le u} |\mathcal{W}(h_1) - \mathcal{W}(h_2)| > \varepsilon\right),$$

which in turn implies by Fatou lemma and the formula $\mathbb{E}|X| = \int_0^\infty P(|X| \geq t)dt$ that

$$w(u) \geq \mathbb{E} \sup_{\|h_1 - h_2\|_2 \leq u} |\mathcal{W}(h_1) - \mathcal{W}(h_2)|.$$

It remains to recall that by Sudakov minorization (this bound is nontrivial, see [16], Lemma 1 for the proof)

$$\begin{aligned} H^{1/2}(\mathcal{H}, L_2(\Pi), \delta) &\leq K \int_\delta^1 \frac{1}{u^2} \mathbb{E} \sup_{\|h_1 - h_2\|_2 \leq u} |\mathcal{W}(h_1) - \mathcal{W}(h_2)| du \\ &\leq K \int_\delta^1 \frac{1}{u^2} w(u) du := \Omega^{1/2}(\delta). \end{aligned} \tag{3.5.18}$$

Let $\bar{\mathcal{H}} \subset \mathcal{H}$ be a minimal δ-net for $\mathcal{H}$ in $L_2(\Pi)$ (i.e., the set of centers of the $L_2(\Pi)$-balls of radius δ that form a minimal covering of $\mathcal{H}$). Clearly,

$$\log \operatorname{card}(\bar{\mathcal{H}}) \leq \Omega(\delta).$$

Also, we have

$$\begin{aligned} &\mathbb{E} \sup_{h \in \mathcal{H}} |R_n(P_{L^\perp} h)| \leq \\ &\leq \mathbb{E} \sup_{h \in \bar{\mathcal{H}}} |R_n(P_{L^\perp} h)| + \mathbb{E} \sup_{h_1, h_2 \in \mathcal{H}, \|h_1 - h_2\|_{L_2(\Pi)} \leq \delta} |R_n(P_{L^\perp}(h_1 - h_2))|. \end{aligned} \tag{3.5.19}$$

To bound the first term in the right hand side, we use an elementary inequality for a Rademacher process indexed by a finite class of functions (see, e.g., [51], proof of Lemma 2):

$$\mathbb{E} \sup_{h \in \bar{\mathcal{H}}} |R_n(P_{L^\perp} h)| \leq C \left[\rho \sqrt{\frac{\Omega(\delta)}{n}} \bigvee \frac{U(L)\Omega(\delta)}{n} \right], \tag{3.5.20}$$

where we also used the facts that

$$\max_{h \in \bar{\mathcal{H}}} \|P_{L^\perp} h\|_{L_2(\Pi)} \leq \rho$$

and

$$\max_{h \in \bar{\mathcal{H}}} \|P_{L^\perp} h\|_{L_\infty} \leq U(L).$$

To bound the second term, we will first use Gaussian multipliers inequality (see [62], Lemma 4.5) to get

$$\mathbb{E}\sup_{h_1,h_2\in\mathcal{H},\|h_1-h_2\|_{L_2(\Pi)}\le\delta}|R_n(P_{L^\perp}(h_1-h_2))|\le\sqrt{\frac{\pi}{2}}\mathbb{E}\sup_{f\in\mathcal{H}(\delta)}|G_n(P_{L^\perp}f)|,$$

where

$$\mathcal{H}(\delta):=\Big\{h_1-h_2:h_1,h_2\in\mathcal{H},\|h_1-h_2\|_{L_2(\Pi)}\le\delta\Big\}.$$

Note that $G_n(f),\ f\in\mathcal{H}(\delta)$ is a centered Gaussian process conditionally on $X_1,\dots,X_n$. Denote $\hat{\mathbb{E}}$ the conditional expectation given $X_1,\dots,X_n$. Then, for all $f_1,f_2\in\mathcal{H}(\delta)$,

$$\hat{\mathbb{E}}(\sqrt{n}G_n(P_{L^\perp}f_1-P_{L^\perp}f_2))^2=\Pi_n(P_{L^\perp}f_1-P_{L^\perp}f_2)^2\le 2\Pi_n(f_1-f_2)^2+2\Pi_n(P_L(f_1-f_2))^2.$$

By the definition of U_L, we have

$$|P_L(f_1-f_2)(x)|\le U_L(x)\|f_1-f_2\|_{L_2(\Pi)},$$

and we are getting the following bound:

$$\hat{\mathbb{E}}(\sqrt{n}G_n(P_{L^\perp}f_1-P_{L^\perp}f_2))^2\le 2\Pi_n(f_1-f_2)^2+2\Pi_n(U_L^2)\|f_1-f_2\|^2_{L_2(\Pi)}.$$

Recall that $\mathcal{W}_\Pi(f), f\in L_2(\Pi)$ denotes the isonormal Gaussian process (i.e., it is a centered Gaussian process with covariance $\hat{\mathbb{E}}\mathcal{W}_\Pi(f_1)\mathcal{W}_\Pi(f_2)=\langle f_1,f_2\rangle_{L_2(\Pi)}$) independent of $X_1,\dots,X_n$ and $g_1,\dots,g_n,$. Then

$$\hat{\mathbb{E}}(G_n(P_{L^\perp}f_1)-G_n(P_{L^\perp}f_2))^2\le 2\hat{\mathbb{E}}(G_n(f_1)-G_n(f_2))^2+$$

$$+\frac{2}{n}\Pi_n(U_L^2)\hat{\mathbb{E}}(W_\Pi(f_1)-W_\Pi(f_2))^2=:\mathbb{E}(Y(f_1)-Y(f_2))^2,$$

where

$$Y(f):=\sqrt{2}G_n(f)+\sqrt{\frac{2}{n}}\|U_L\|_{L_2(\Pi_n)}W_\Pi(f).$$

By Slepian lemma (see [62], Theorem 3.15), we conclude that

$$\hat{\mathbb{E}}\sup_{f\in\mathcal{H}(\delta)}|G_n(P_{L^\perp}f)|\le 2\hat{\mathbb{E}}\sup_{f\in\mathcal{H}(\delta)}|Y(f)|,$$

which also implies that

$$\mathbb{E}\sup_{h\in\mathcal{H}(\delta)}|R_n(P_{L^\perp}h)|\le$$

$$4\sqrt{\pi}\mathbb{E}\sup_{h\in\mathcal{H}(\delta)}|G_n(h)|+4\sqrt{\frac{\pi}{n}}\mathbb{E}\|U_L\|_{L_2(\Pi_n)}\hat{\mathbb{E}}\sup_{h\in\mathcal{H}(\delta)}|W_\Pi(h)|. \tag{3.5.21}$$

To bound the right hand side, observe that, by Dudley's entropy bound and (3.5.18),

$$\hat{\mathbb{E}}\sup_{h\in\mathcal{H}(\delta)}|W_\Pi(h)|\le C\int_0^\delta\sqrt{\log N(\mathcal{H}(\delta),L_2(\Pi),u)}du\le$$

$$C\int_0^\delta\sqrt{2\log N(\mathcal{H},L_2(\Pi),u/2)}du\le C\delta H^{1/2}(\delta)\le C\delta\Omega^{1/2}(\delta)$$

with some numerical constant $C>0$. We also have

$$\mathbb{E}\|U_L\|^2_{L_2(\Pi_n)}=\mathbb{E}U_L^2(X)=d.$$

Finally, our assumptions on the continuity modulus imply that

$$\mathbb{E}\sup_{h\in\mathcal{H}(\delta)}|G_n(h)|\le\frac{1}{\sqrt{n}}w_n(\delta).$$

Combining this with (3.5.19) and (3.5.20) yields

$$\mathbb{E}\sup_{h\in\mathcal{H}}|R_n(P_{L^\perp}h)|\le C\left[\rho\sqrt{\frac{\Omega(\delta)}{n}}\bigvee\delta\sqrt{d}\sqrt{\frac{\Omega(\delta)}{n}}\bigvee\frac{U(L)\Omega(\delta)}{n}\bigvee\sqrt{\frac{1}{n}}w_n(\delta)\right]$$

and it is enough to take $\delta:=\rho/\sqrt{d}$ to complete the proof.

□

Lemma 3.5.4. *Under assumptions of Theorem 3.5.2, there exists a constant $C>0$ depending only on the loss such that with probability $\ge 1-e^{-t}$ for all $\frac{1}{\sqrt{n}}\le\delta\le 1$, $\frac{1}{\sqrt{n}}\le\Delta\le 1$*

$$\alpha(\delta,\Delta)\le C\left[\delta\sqrt{\frac{d+t_n}{n}}\bigvee\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}}\bigvee\sqrt{\frac{1}{n}}w_n(\rho/\sqrt{d})\bigvee\right.$$

$$\bigvee\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\sqrt{\frac{1}{n}}w_n\left(\frac{1}{\sqrt{d}}\right)\bigvee\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\sqrt{\frac{\Omega(1/\sqrt{d})}{n}}\bigvee\Delta\sqrt{\frac{1}{n}}w_n\left(\frac{1}{\sqrt{d}}\right)$$

$$\left.\bigvee\Delta\sqrt{\frac{\Omega(1/\sqrt{d})}{n}}\bigvee\frac{U(L)\Omega(\rho/\sqrt{d})+t_n}{n}\right]=:\hat{\beta}_n(\delta,\Delta).$$

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

Proof. Recall that

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

and

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

The function $u \mapsto \ell'(y, f_{\lambda_\varepsilon} + u)u, \ |u| \le 2$ is Lipschitz with Lipschitz constant depending only on ℓ and M. Note that

$$\ell'(y, f_\lambda(\cdot))(f_\lambda(\cdot) - f_{\lambda_\varepsilon}(\cdot)) = \ell'(y, f_{\lambda_\varepsilon} + u)u|_{u=f_\lambda(\cdot)-f_{\lambda_\varepsilon}(\cdot)}$$

This allows us to apply the symmetrization and contraction inequalities which result in the following bound:

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

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

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

Hence, it is enough to bound separately the expected supremum of the Rademacher process R_n for each term in the sum. For the first term, the standard bound on Rademacher processes indexed by a finite dimensional subspace (see, e.g., [48], example 1) yields

$$\mathbb{E} \sup_{\lambda\in\Lambda(\delta,\Delta)} \left\{ \left| R_n\left(P_L(f_\lambda - f_{\lambda_\varepsilon})\right) \right| \right\} \le \delta\sqrt{\frac{d}{n}}. \tag{3.5.22}$$

To bound the remaining terms, we will use Lemma 3.5.3. First, note that

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}\left|R_n\left(\int\limits_{\mathcal{H}\setminus\mathcal{H}'}(\lambda-\lambda_\varepsilon)(h)P_{L^\perp}h\ d\mu(h)\right)\right| \tag{3.5.23}$$

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

$$\leq\sup_{\lambda\in\Lambda(\delta,\Delta)}\Big(\Lambda(\mathcal{H}\setminus\mathcal{H}')+\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\Big)\mathbb{E}\sup_{h\in\mathcal{H}\setminus\mathcal{H}'}|R_n(P_{L^\perp}h)|\leq$$

$$\Big(\Delta+\Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\Big)\mathbb{E}\sup_{h\in\mathcal{H}\setminus\mathcal{H}'}|R_n(P_{L^\perp}h)|.$$

We now use the bound of Lemma 3.5.3 with $\mathcal{H}\setminus\mathcal{H}'$ instead of $\mathcal{H}$ and with $\rho=1$ to get

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)}\left|R_n\left(\int\limits_{\mathcal{H}\setminus\mathcal{H}'}(\lambda-\lambda_\varepsilon)(h)P_{L^\perp}h\ d\mu(h)\right)\right|\leq \tag{3.5.24}$$

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

Similarly,

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

and using the bound of Lemma 3.5.3 with $\mathcal{H}'$ instead of $\mathcal{H}$ and with $\rho:=\rho(\mathcal{H}',L)$, we get

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

$$C\left[\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}}\bigvee\frac{U(L)\Omega(\rho/\sqrt{d})}{n}\bigvee\sqrt{\frac{1}{n}}w_n(\rho/\sqrt{d})\right]. \tag{3.5.25}$$

Combining (3.5.22)–(3.5.25) results in the following bound:

$$\mathbb{E}\alpha(\delta,\Delta) \le C\left[\delta\sqrt{\frac{d}{n}} \bigvee \rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} w_n(\rho/\sqrt{d}) \bigvee \right. \tag{3.5.26}$$

$$\bigvee \Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\sqrt{\frac{1}{n}} w_n\left(\frac{1}{\sqrt{d}}\right) \bigvee \Lambda_\varepsilon(\mathcal{H}\setminus\mathcal{H}')\sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \Delta\sqrt{\frac{1}{n}} w_n\left(\frac{1}{\sqrt{d}}\right)$$

$$\left. \bigvee \Delta\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(1/\sqrt{d})}{n}\right].$$

Talagrand's concentration inequality implies that with probability at least $1-e^{-s}$ and with a proper choice of numerical constant $C>0$

$$\alpha_n(\delta,\Delta) \le \beta(\delta,\Delta,s) := 2\left(\mathbb{E}\alpha_n(\delta,\Delta) + C\delta\sqrt{\frac{s}{n}} + C\frac{s}{n}\right) \tag{3.5.27}$$

We have to make the bound uniform with respect to

$$\frac{1}{\sqrt{n}} \le \delta \le 1,\ \frac{1}{\sqrt{n}} \le \Delta \le 1$$

To this end, let

$$\delta_j = \Delta_j = \frac{1}{2^j},$$

$$t_{i,j} = t + 2\log(i+1) + 2\log(j+1) + 2\log 2, i,j \ge 0.$$

Then, with probability at least

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

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

$$\alpha(\delta,\Delta) \le \beta(\delta_i,\Delta_j,t_{i,j}).$$

Note that

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

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

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

implying that

$$t_{i,j} \le t_n.$$

Thus, with probability at least $1 - e^{-t}$, for all $\delta, \Delta \in [n^{-1/2}, 1]$

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

□

To complete the proof of the theorem, denote

$$\hat{\delta} := \|f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}\|_{L_2(\Pi)},$$

and

$$\hat{\Delta} := \hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}').$$

By lemma 3.5.4, the following inequalities hold with probability at least $1 - e^{-t}$ (uniformly for $n^{-1/2} \le \hat{\delta} \le 1, n^{-1/2} \le \hat{\Delta} \le 1$):

$$c\hat{\delta}^2 \le \hat{\beta}_n(\hat{\delta}, \hat{\Delta}), \tag{3.5.28}$$

$$\varepsilon\hat{\Delta} \le 2\varepsilon\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') + \frac{2}{\log(2)} \hat{\beta}_n(\hat{\delta}, \hat{\Delta}) \tag{3.5.29}$$

It remains to solve (3.5.28), (3.5.29) to get the bounds for $\hat{\delta}, \hat{\Delta}$. The cases when $\hat{\delta} < n^{-1/2}$ and/or $\hat{\Delta} < n^{-1/2}$ are simple: because $\alpha_n(\delta, \Delta)$ is non-decreasing in both variables, it is enough to substitute $n^{-1/2}$ instead of δ or Δ into the expression of it's upper bound $\hat{\beta}(\delta, \Delta)$ to get required inequalities. We proceed with the main case. If $\hat{\Delta} \le \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$, then (3.5.29) and the first bound of the theorem are trivially

satisfied. Moreover, (3.5.28) yields

$$\hat{\delta}^2 \leq C \left[\hat{\delta} \sqrt{\frac{d+t_n}{n}} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \right. \tag{3.5.30}$$
$$\left. \bigvee \sqrt{\frac{1}{n}} \left(w_n(\rho/\sqrt{d}) \bigvee \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') w_n(1/\sqrt{d}) \right) \bigvee \frac{U(L) H(\rho/\sqrt{d}) + t_n}{n} \right]$$

and the third bound of the theorem follows immediately. On the other hand, if $\hat{\Delta} > \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}')$, then (3.5.28) implies that with some constant $C > 0$

$$\hat{\delta}^2 \leq C \Big[\hat{\delta} \sqrt{\frac{d+t_n}{n}} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} \left(w_n(\rho/\sqrt{d}) \bigvee \hat{\Delta} w_n(1/\sqrt{d}) \right)$$
$$\bigvee \hat{\Delta} \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{U(L) H(\rho/\sqrt{d}) + t_n}{n} \Big]. \tag{3.5.31}$$

Solutions $\hat{\delta}$ of this inequality satisfy the bound

$$\hat{\delta}^2 \leq C \Big[\frac{d+t_n}{n} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} \left(w_n(\rho/\sqrt{d}) \bigvee \hat{\Delta} w_n(1/\sqrt{d}) \right) \bigvee \tag{3.5.32}$$
$$\bigvee \hat{\Delta} \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{U(L) \Omega(\rho/\sqrt{d}) + t_n}{n} \Big].$$

Substituting the resulting upper bound on $\hat{\delta}$ into inequality (3.5.29), with some elementary algebra, yields

$$\varepsilon \hat{\Delta} \leq 2\varepsilon \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') + C \Big[\frac{d+t_n}{n} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \tag{3.5.33}$$
$$\sqrt{\frac{1}{n}} \left(w_n(\rho/\sqrt{d}) \bigvee \hat{\Delta} w_n(1/\sqrt{d}) \right) \bigvee \hat{\Delta} \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{U(L) \Omega(\rho/\sqrt{d}) + t_n}{n} \Big].$$

If the condition on ε holds with constant $D > 2C$, this implies

$$\hat{\Delta} \leq 2\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') + C \Big[\frac{d+t_n}{n\varepsilon} \bigvee \frac{\rho}{\varepsilon} \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee$$
$$\sqrt{\frac{1}{n}} \frac{w_n(\rho/\sqrt{d})}{\varepsilon} \bigvee \frac{U(L) \Omega(\rho/\sqrt{d}) + t_n}{n\varepsilon} \Big]. \tag{3.5.34}$$

By (3.5.15) and Lemma 3.5.4, with probability at least $1 - e^{-t}$,

$$\varepsilon K(\hat{\lambda}_\varepsilon, \lambda_\varepsilon) \leq \frac{2}{\log 2} \alpha_n(\hat{\delta}, \hat{\Delta}) \leq \frac{2}{\log 2} \hat{\beta}_n(\hat{\delta}, \hat{\Delta}),$$

it is enough now to substitute the bounds on $\hat{\delta}, \hat{\Delta}$ into the expression for $\hat{\beta}_n(\hat{\delta}, \hat{\Delta})$ to see that with some constant $C > 0$

$$\varepsilon K(\hat{\lambda}_\varepsilon, \lambda_\varepsilon) \le C \Bigg[\frac{d + t_n}{n} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} \Big(w_n(\rho/\sqrt{d}) \bigvee \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') w_n(1/\sqrt{d}) \Big) \bigvee \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d}) + t_n}{n} \Bigg].$$

This and the bound of Lemma 3.4.10

$$\Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \le 2\hat{\Lambda}_\varepsilon(\mathcal{H} \setminus \mathcal{H}') + K(\hat{\lambda}_\varepsilon, \lambda_\varepsilon)$$

imply also the second bound of the theorem, which completes the proof.

□

To derive an upper bound on the random error $|\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) - \mathcal{E}(f_{\lambda_\varepsilon})|$ from the bound on the $L_2(\Pi)$-norm $\|f_{\hat{\lambda}_\varepsilon} - f_{\lambda_\varepsilon}\|_{L_2(\Pi)}$, it is enough to use Proposition 3.4.1 of Section 2.3. For arbitrary $\mathcal{H}' \subset \mathcal{H}$ and $L \subset L_2(\Pi)$ with $d := \dim(L)$ and $\rho := \rho(\mathcal{H}', L)$, denote

$$\Gamma_n(\mathcal{H}'; L; t) := \Bigg[\frac{d + t_n}{n} \bigvee \rho \sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \sqrt{\frac{\Omega(1/\sqrt{d})}{n}} \bigvee \frac{w_n(\rho/\sqrt{d})}{\sqrt{n}} \bigvee \Lambda_\varepsilon(\mathcal{H} \setminus \mathcal{H}') \frac{w_n(1/\sqrt{d})}{\sqrt{n}} \bigvee \frac{U(L)\Omega(\rho/\sqrt{d}) + t_n}{n} \Bigg]. \tag{3.5.35}$$

Corollary 3.5.5. *With probability at least* $1 - e^{-t}$,

$$|\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) - \mathcal{E}(f_{\lambda_\varepsilon})| \le C \Big[\Gamma_n(\mathcal{H}'; L; t) \bigvee \sqrt{\mathcal{E}(f_{\lambda_\varepsilon}) \Gamma_n(\mathcal{H}'; L; t)} \Big].$$

3.5.3 Oracle inequality for prediction problems

For clarity, we will assume that (3.4.5) is satisfied so that bounds of Proposition 3.4.2 hold for a suitable function $T(u)$. Combined, results of the previous two sections imply the following:

Corollary 3.5.6. *There exist constants $C, \bar{D} > 0$ depending only on ℓ such that for all measurable subsets $\mathcal{H}' \subset \mathcal{H}$, for all finite dimensional subspaces $L \subset L_2(\Pi)$ with $d := \dim(L)$ and*

$$\rho := \rho(\mathcal{H}'; L),$$

for all

$$\varepsilon \geq \bar{D}\sqrt{\frac{T(1/\sqrt{d})\log d}{n}}$$

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

$$\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) \leq \inf_{\lambda \in \mathbb{D}} \left(2\mathcal{E}(f_\lambda) + \varepsilon^2\gamma^2(\log\lambda) + C\left[\frac{d+t_n}{n} \bigvee \rho\sqrt{\frac{T(\rho/\sqrt{d})}{n}}\log(\sqrt{d}/\rho) \bigvee \right.\right.$$
$$\left.\left. \bigvee \int_{\mathcal{H}\setminus\mathcal{H}'} \lambda d\mu \sqrt{\frac{T(1/\sqrt{d})}{n}}\log d \bigvee \frac{U(L)T(\rho/\sqrt{d})}{n} \right] \right),$$

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

In particular, if $\mathcal{H}$ is a VC-subgraph class of VC-dimension V (for the definitions and examples, see [91], Section 2.6), then one can take $T(u) = (2V+1)\log\frac{A}{u}$. In this case, previous inequality can further be simplified to

$$\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) \leq \inf_{\lambda \in \mathbb{D}} \left(2\mathcal{E}(f_\lambda) + \varepsilon^2\gamma^2(\log\lambda) + C\left[\frac{d+t_n}{n} \bigvee \frac{\rho}{\sqrt{n}}\log^{3/2}\frac{\sqrt{d}}{\rho} \bigvee \right.\right.$$
$$\left.\left. \bigvee \int_{\mathcal{H}\setminus\mathcal{H}'} \lambda d\mu \frac{\log^{3/2} d}{\sqrt{n}} \bigvee \frac{U(L)\log\left(\sqrt{d}/\rho\right)}{n} \right] \right).$$

Constant 2 in front of $\mathcal{E}(f_\lambda)$ can be replaced by $1+\delta$ at the price of replacing C by $\frac{C}{\delta}$.

3.6 Density estimation

This section is devoted to other applications of techniques developed earlier in the chapter. We have seen that many prediction problems can be embedded and analyzed in the context of dictionary learning. In turns out that the problem of L_2 - density

estimation also naturally fits this framework (see [18], [51] for a similar approach). At the same time, we were able to relax the assumption on the uniform boundedness of the dictionary with the help of Proposition 3.4.4.

Suppose that $X_1, \dots, X_n \in S$ are iid observations from some distribution P that is absolutely continuous with respect to a given σ-finite measure ν such that $\frac{dP}{d\nu} = f_*$. Assume that $\|f_*\|_\infty = M < \infty$. As before, we want to estimate the unknown f_* (which in general might not belong to $\mathrm{co}(\mathcal{H})$) by a mixture of the elements of a dictionary $\mathcal{H}$, which in this case consists of probability densities with respect to ν. We will assume that

1. $\sup_{h \in \mathcal{H}} \|h\|_{L_2(\nu)} < \infty$;
2. the envelope $F(x) = F_{\mathcal{H}}(x) := \sup_{h \in \mathcal{H}} h(x)$ is such that $\|F\|_{\psi_2} := \Psi < \infty$.

In what follows, we will write $\|\cdot\|_2$ for $\|\cdot\|_{L_2(\nu)}$, and

$$\langle h_1, h_2 \rangle := \langle h_1, h_2 \rangle_{L_2(\nu)} = \int_S h_1(x) h_2(x) d\nu(x).$$

Let $\mathbb{D}$ be convex weakly compact subset $\mathbb{D}$ of $L_1(\mu)$, consisting of probability density functions with respect to μ. Consider the following minimization problem with L_2 - loss:

$$\lambda_\varepsilon = \mathrm{argmin}_{\lambda \in \mathbb{D}} \Big[\|f_\lambda - f_*\|_2^2 + \varepsilon \,\mathrm{pen}(\lambda) \Big]. \tag{3.6.1}$$

Since $\langle f_\lambda, f_* \rangle = P f_\lambda$, (3.6.1) is equivalent to

$$\lambda_\varepsilon = \mathrm{argmin}_{\lambda \in \mathbb{D}} \Big[\|f_\lambda\|_2^2 - 2P f_\lambda + \varepsilon \,\mathrm{pen}(\lambda) \Big]. \tag{3.6.2}$$

The empirical version of (3.6.2) is

$$\hat{\lambda}_\varepsilon = \mathrm{argmin}_{\lambda \in \mathbb{D}} \Big[\|f_\lambda\|_2^2 - 2P_n f_\lambda + \varepsilon \,\mathrm{pen}(\lambda) \Big]. \tag{3.6.3}$$

If we take $\mathrm{pen}(\lambda) = \int_{\mathcal{H}} \lambda \log \lambda d\mu$ to be the (negative) entropy penalty, then analogues of the previous results of this chapter hold. Note that the excess risk is simply given

by $\mathcal{E}(f_\lambda) = \|f_\lambda - f_*\|^2_{L_2(\nu)}$.

Remark: although we can not apply Theorems 3.5.2 and 3.5.1 directly, the analysis for the present problem follows the same path as before, even with several simplifications. We will outline some details below.

The uniform boundedness assumption on the dictionary seems to be too restrictive for the density estimation problem. Another idea that motivates subsequent results is based on an interesting observation made by V. Koltchinskii (see Corollaries 9.7, 9.8 in [53]): in the case of sparse regression problem with squared loss in a dictionary of cardinality N, the (non-penalized!) least-squares estimator over the unit simplex in $\mathbb{R}^N$ possesses sparsity properties. This phenomenon appears as a result of analysis of ℓ_1-penalized empirical risk minimization over an abstract closed convex set and the fact that the ℓ_1-norm is constant on the unit simplex. Similar conclusions can be made based on results in [18], although it is not mentioned explicitly in the paper. Motivated by this observation, we analyze problems (3.6.2) and (3.6.3) for

$$\mathrm{pen}(\lambda) = \|\lambda\|_1 := \int\limits_{\mathcal{H}} |\lambda(h)| d\mu$$

which is identically equal to 1 whenever λ is a probability density function (with respect to μ). We stress the fact that the penalty is introduced artificially as a method of theoretical analysis that allows us to obtain interesting results. Denote

$$F(\lambda) := \|f_\lambda\|_2^2 - 2Pf_\lambda + \varepsilon\|\lambda\|_1, \ \lambda \in \mathbb{D},$$
$$F_n(\lambda) := \|f_\lambda\|_2^2 - 2P_nf_\lambda + \varepsilon\|\lambda\|_1, \ \lambda \in \mathbb{D}.$$

As before, it is easy to see that both F and F_n are continuous in $L_1(\mu)$ and (not necessarily strictly) convex. Indeed, by Hölder and integral Minkowski (see [85],

Lemma A.1) inequalities

$$|F(\lambda_n) - F(\lambda_0)| = \left| \int_S (f_{\lambda_n} - f_{\lambda_0})(f_{\lambda_n} + f_{\lambda_0})\, d\nu - 2P(f_{\lambda_n} - f_{\lambda_0}) \right| \leq$$

$$\leq \|f_{\lambda_n} + f_{\lambda_0}\|_{L_2(\nu)} \|f_{\lambda_n} - f_{\lambda_0}\|_{L_2(\nu)} + 2\|\lambda_n - \lambda_0\|_1 \mathbb{E} F_{\mathcal{H}}(X) \leq$$

$$\leq 2 \sup_{h \in \mathcal{H}} \|h\|^2_{L_2(\nu)} \|\lambda_n - \lambda_0\|_1 + 2\|\lambda_n - \lambda_0\|_1 \mathbb{E} F_{\mathcal{H}}(X) \to 0$$

given that $\|\lambda_n - \lambda_0\|_{L_1(\mu)} \to 0$. Here, $F_{\mathcal{H}}(x) := \sup_{h \in \mathcal{H}} h(x)$ is the envelope of class $\mathcal{H}$. Now an argument of Theorem 3.4.7 implies that both problems (3.6.2) and (3.6.3) have (not necessarily unique) solutions in every convex weakly compact subset $\mathbb{D}$ of $L_1(\mu)$, consisting of probability density functions with respect to μ. Once again, note that for our choice $\text{pen}(\lambda) = \|\lambda\|_1$, solution of (3.6.3) coincides with a solution of the non-penalized $L_2(\nu)$-norm minimization problem, since the penalty term is just a constant on $\mathbb{D}$.

For $\lambda \in \mathbb{D}$, let

$$\partial \|\lambda\|_1 = \{w : \mathcal{H} \mapsto [-1, 1],\ w(t) = \text{sign } \lambda(t) \text{ for } t \in \text{supp}(\lambda)\} \tag{3.6.4}$$

be the subdifferential of $\|\cdot\|_1$ at point λ (we also assume that w is measurable).

Remark: equality in (3.6.4) follows from the general description of the subdifferential of a norm $\|\cdot\|$ in a Banach space $\mathfrak{X}$:

$$\partial \|x\| = \begin{cases} \{x^* \in \mathfrak{X}^* :\ \|x^*\| = 1,\ x^*(x) = \|x\|\}, & x \neq 0, \\ \{x^* \in \mathfrak{X}^* :\ \|x^*\| \leq 1\}, & x = 0, \end{cases}$$

where $\mathfrak{X}^*$ is the dual space. For details on our specific example, see [45], paragraph 4.5.1.

Next, we define a version of *the alignment coefficient* $\gamma(w)$ with respect to the $L_2(\nu)$-norm by

$$\gamma(w) := \sup \left\{ \langle w, u \rangle_{L_2(\mu)} \ :\ \|f_u\|_{L_2(\nu)} = 1, \langle u, 1 \rangle_{L_2(\mu)} = 0 \right\}.$$

We will be interested in those "oracles" $\lambda \in \mathbb{D}$ for which there exist $w \in \partial\|\lambda\|_1$ with $\gamma(w) < \infty$.

Given $w : \mathcal{H} \mapsto [-1, 1]$, let

$$\mathcal{H}(w) = \left\{ h \in \mathcal{H} : \ |w(h)| \geq \frac{1}{2} \right\}.$$

If $w \in \partial\|\lambda\|_1$, $\mathcal{H}(w)$ can be seen as the "smoothed" support of λ.

Next, note that directional derivative of the functional F_n at the point λ_1 in direction $u := \lambda_2 - \lambda_1$ is

$$DF_n(\lambda_1; u) := \lim_{t \downarrow 0} \frac{F_n(\lambda_1 + tu) - F_n(\lambda_1)}{t} = 2 \langle f_{\lambda_1}, f_{\lambda_2} - f_{\lambda_1} \rangle_{L_2(\nu)} - 2P_n(f_{\lambda_2} - f_{\lambda_1}),$$

where we used the fact that both $\lambda_1, \ \lambda_2$ are densities with respect to μ, hence

$$\|\lambda_1 + tu\|_1 = \|\lambda_1\|_1 = 1 \quad \forall t \in [0, 1].$$

Let $\hat{\lambda}$ be a solution to (3.6.3) (note that it is independent of ε when $\text{pen}(\lambda) = \|\lambda\|_1$). When $\lambda_1 = \hat{\lambda}$, the corresponding directional derivatives are nonnegative for any $\lambda_2 \in \mathbb{D}$. With these preliminary observations and previously developed techniques, we can prove the main result of this Section – an oracle inequality for performance of $\hat{\lambda}$.

First, recall some definitions: let L be a finite dimensional subspace of $L_2(\nu)$, $d = \dim(L)$ and

$$U_L(x) := \sup_{g \in L, \|g\|_{L_2(P)} \leq 1} |g(x)|.$$

Given $\mathcal{H}' \subseteq \mathcal{H}$, let

$$\rho(\mathcal{H}'; L) := \sup_{h \in \mathcal{H}'} \|P_{L^\perp} h\|_{L_2(P)} \leq \sqrt{M} \sup_{h \in \mathcal{H}'} \|P_{L^\perp} h\|_{L_2(\nu)}.$$

Moreover, denote

$$\mathcal{D} := \sup_{h \in \mathcal{H}} \|h\|_{L_2(P)};$$

$$U(L, \mathcal{H}) := \Psi + \|U_L\|_{\psi_2};$$

$$\Omega_n(\rho) := \Omega\left(\rho\sqrt{\frac{1}{d}}\right) \vee \log n;$$

$$t_n := t + c \log\log n.$$

For brevity, we will denote

$$\mathcal{Q}(s) := s\sqrt{\Omega(s/\sqrt{d})} \bigvee w_n(s/\sqrt{d}),\ s \in (0, \mathcal{D}].$$

For definitions of $\Omega(\cdot)$, $w_n(\cdot)$, see (3.4.4).

Theorem 3.6.1. *Let $\bar{\lambda} \in \mathbb{D}$ and $\bar{w} \in \partial\|\bar{\lambda}\|_1$. There exist numerical constants C and $\bar{D}$ large enough such that for any*

$$\varepsilon \geq \bar{D} \cdot \frac{\mathcal{Q}(\mathcal{D})}{\sqrt{n}}$$

and any subspace L with $d = \dim(L)$ and $\rho = \rho(\mathcal{H}(\bar{w}); L)$

$$\|f_{\hat{\lambda}} - f_{\lambda_*}\|_2^2 + \frac{\varepsilon}{4} \int\limits_{\mathcal{H}\setminus\mathcal{H}(\bar{w})} \hat{\lambda} d\mu \leq \|f_{\bar{\lambda}} - f_*\|_2^2 + \frac{1}{2}\varepsilon^2\gamma^2(\bar{w}) + C\left[\frac{dM + t_n}{n} \bigvee \frac{\mathcal{Q}(\rho)}{\sqrt{n}} \bigvee U(L, \mathcal{H})\frac{\Omega_n(\rho) + t_n \log n}{n}\right],$$

with probability $\geq 1 - e^{-t}$.

Remarks:

1. Note that the oracle inequality above is *exact*, meaning that it has factor 1 in front of $\|f_{\bar{\lambda}} - f_*\|_2^2$. We were not able to get constant 1 for prediction problems considered above.

2. Although $\mathcal{D}$ (or its nonrandom upper bound) are generally unknown, its value is not needed to obtain $\hat{\lambda}$ since $\hat{\lambda}$ does not depend on a particular choice of ε.

Proof. As we have already mentioned above, the necessary conditions for the minima in problem (3.6.3) can be written as follows: for any $\bar{\lambda} \in \mathbb{D}$

$$2\left\langle f_{\hat{\lambda}_\varepsilon}, f_{\hat{\lambda}_\varepsilon} - f_{\bar{\lambda}} \right\rangle - 2P_n(f_{\hat{\lambda}_\varepsilon} - f_{\bar{\lambda}}) \leq 0. \tag{3.6.5}$$

Next, let $\bar{w} \in \partial\|\bar{\lambda}\|_1$ and $\hat{w} \in \partial\|\hat{\lambda}\|_1$. Adding $2\left\langle f_*, f_{\bar{\lambda}} - f_{\hat{\lambda}_\varepsilon} \right\rangle + \varepsilon \left\langle \hat{w} - \bar{w}, \hat{\lambda} - \bar{\lambda} \right\rangle$ to both sides of the inequality and noting that $\left\langle f_*, f_{\bar{\lambda}} - f_{\hat{\lambda}_\varepsilon} \right\rangle = P(f_{\bar{\lambda}} - f_{\hat{\lambda}_\varepsilon})$, we get

$$\begin{aligned} 2\left\langle f_{\hat{\lambda}_\varepsilon} - f_*, f_{\hat{\lambda}_\varepsilon} - f_{\bar{\lambda}} \right\rangle + \varepsilon \int_{\mathcal{H}} (\hat{w} - \bar{w})(h)(\hat{\lambda} - \bar{\lambda})(h) d\mu(h) \leq \\ \leq 2(P - P_n)(f_{\bar{\lambda}} - f_{\hat{\lambda}_\varepsilon}) + \varepsilon \int_{\mathcal{H}} (\hat{w} - \bar{w})(h)(\hat{\lambda} - \bar{\lambda})(h) d\mu(h). \end{aligned} \tag{3.6.6}$$

It is easy to see that

$$\int_{\mathcal{H}} (\hat{w} - \bar{w})(h)(\hat{\lambda} - \bar{\lambda})(h) d\mu(h) \geq \frac{1}{2} \int_{\mathcal{H} \setminus \mathcal{H}(\bar{w})} \hat{\lambda}(h) d\mu(h) \tag{3.6.7}$$

for any choice of $\hat{w}$ and $\bar{w}$. At the same time, $\hat{w}(h) \equiv 1 \in \partial\|\hat{\lambda}\|_1$ by definition. For this choice of $\hat{w}$, we clearly have that for any $\bar{w} \in \partial\|\bar{\lambda}\|_1$

$$\int_{\mathcal{H}} (\hat{w} + \bar{w})\hat{\lambda} d\mu \leq \int_{\mathcal{H}} (\hat{w} + \bar{w})\bar{\lambda} d\mu,$$

which is equivalent to $\int_{\mathcal{H}} \hat{w}(\hat{\lambda} - \bar{\lambda}) d\mu \leq \int_{\mathcal{H}} \bar{w}(\bar{\lambda} - \hat{\lambda}) d\mu$, implying

$$\int_{\mathcal{H}} (\hat{w} - \bar{w})(\hat{\lambda} - \bar{\lambda}) d\mu \leq 2 \int_{\mathcal{H}} \bar{w}(\bar{\lambda} - \hat{\lambda}) d\mu. \tag{3.6.8}$$

Moreover,

$$2\left\langle f_{\hat{\lambda}_\varepsilon} - f_*, f_{\hat{\lambda}_\varepsilon} - f_{\bar{\lambda}} \right\rangle = \|f_{\hat{\lambda}_\varepsilon} - f_*\|_2^2 + \|f_{\hat{\lambda}_\varepsilon} - f_{\bar{\lambda}}\|_2^2 - \|f_{\bar{\lambda}} - f_*\|_2^2. \tag{3.6.9}$$

Together with (3.6.6), formulas (3.6.7),(3.6.8),(3.6.9) give

$$\begin{aligned} \|f_{\hat{\lambda}} - f_{\lambda_*}\|_2^2 + \|f_{\hat{\lambda}} - f_{\bar{\lambda}}\|_2^2 + \frac{\varepsilon}{2} \int_{\mathcal{H} \setminus \mathcal{H}(\bar{w})} \hat{\lambda}(h) d\mu(h) \leq \\ \leq \|f_{\bar{\lambda}} - f_*\|_2^2 + 2(P - P_n)(f_{\bar{\lambda}} - f_{\hat{\lambda}_\varepsilon}) + 2\varepsilon \int_{\mathcal{H}} \bar{w}(h)(\bar{\lambda} - \hat{\lambda})(h) d\mu(h). \end{aligned} \tag{3.6.10}$$

First, note that

$$2\varepsilon \int\limits_{\mathcal{H}} \bar{w}(h)(\bar{\lambda}-\hat{\lambda})(h)d\mu(h) \le 2\varepsilon \|f_{\hat{\lambda}} - f_{\bar{\lambda}}\|_2 \sup_{\|u\|_{L_2(\nu)}\le 1} \langle \bar{w}, u\rangle =$$

$$= 2\varepsilon\gamma(\bar{w})\|f_{\hat{\lambda}} - f_{\bar{\lambda}}\|_2 \le \frac{1}{2}\|f_{\hat{\lambda}} - f_{\bar{\lambda}}\|_2^2 + 2\varepsilon^2\gamma^2(\bar{w}). \tag{3.6.11}$$

To this end, we need to control the empirical process $(P-P_n)(f_{\bar{\lambda}} - f_{\hat{\lambda}_\varepsilon})$. We have all the necessary tools to obtain the required bound. As before, define

$$\Lambda(\delta,\Delta) := \left\{ \lambda\in\mathbb{D}:\ \|f_\lambda - f_{\bar{\lambda}}\|_{L_2(\nu)} \le \delta, \int\limits_{\mathcal{H}\setminus\mathcal{H}(\bar{w})} \lambda(h)d\mu(h) \le \Delta \right\}$$

and

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

Note that, since $f_* \le M$, $\|f_\lambda - f_{\bar{\lambda}}\|_{L_2(\nu)} \le \delta$ implies $\|f_\lambda - f_{\bar{\lambda}}\|_{L_2(P)} \le M^{1/2}\delta$.
By symmetrization inequality (Theorem 1.2.6)

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

Let L be a d-dimensional subspace of $L_2(\nu)$; we will use the following representation and separately bound each term in the sum below:

$$f_\lambda - f_{\lambda_\varepsilon} = P_L\left(f_\lambda - f_{\lambda_\varepsilon}\right) + \int\limits_{\mathcal{H}(\bar{w})} P_{L^\perp}(h)\left(\lambda(h)-\lambda_\varepsilon(h)\right)d\mu(h) +$$

$$+ \int\limits_{\mathcal{H}\setminus\mathcal{H}(\bar{w})} P_{L^\perp}(h)\left(\lambda(h)-\lambda_\varepsilon(h)\right)d\mu(h). \tag{3.6.12}$$

We will follow the steps of Lemma 3.5.4, emphasizing the necessary modifications. First, note that

$$\mathbb{E}\sup_{\lambda\in\Lambda(\delta,\Delta)} \left\{ |R_n\left(P_L(f_\lambda - f_{\bar{\lambda}})\right)| \right\} \le \delta\sqrt{\frac{dM}{n}}. \tag{3.6.13}$$

The remaining bounds are based on the following modification of Lemma 3.5.3:

Lemma 3.6.2. *Let $\mathcal{H}$ be a class of functions on S with an envelope F such that $\|F\|_{\psi_2} := \Psi < \infty$ and let $L \subset L_2(\nu)$ be a finite dimensional subspace with $d := \dim(L)$ and $U_L(x) := \sup_{h \in L, \|h\|_{L_2(P)} \leq 1} |h(x)|$. Denote*

$$\mathcal{D} := \sup_{h \in \mathcal{H}} \|h\|_{L_2(P)};$$

$$\rho := \rho(\mathcal{H}; L) := \sup_{h \in \mathcal{H}} \|P_{L^\perp} h\|_{L_2(P)};$$

$$U(L, \mathcal{H}) := \Psi + \|U_L\|_{\psi_2};$$

$$\Omega_n(\rho) := \Omega\left(\rho\sqrt{\frac{1}{d}}\right) \vee \log n.$$

Suppose that assumption (3.4.5) holds. Then with some constant $C > 0$

$$\mathbb{E} \sup_{h \in \mathcal{H}} |R_n(P_{L^\perp} h)| \leq C\left[\rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee U(L, \mathcal{H}) \frac{\Omega_n(\rho)}{n} \bigvee \sqrt{\frac{1}{n}} w_n\left(\rho\sqrt{\frac{1}{d}}\right)\right].$$

Proof. The proof is identical to the argument behind Lemma 3.5.3, the only difference being the bound on the Rademacher process indexed by finitely many functions. In this case, we use the estimate provided by Corollary 3.4.5, part (b). More precisely, using the notations of Lemma 3.5.3, we get from Corollary 3.4.5

$$\mathbb{E} \sup_{h \in \tilde{\mathcal{H}}} |R_n(P_{L^\perp} h)| \leq C\left[\rho\sqrt{\frac{\Omega(\delta)}{n}} \bigvee \frac{\sqrt{\Psi + \|U_L\|_{\psi_2}}}{n} (\Omega(\delta) \vee \log n)\right]. \tag{3.6.14}$$

The rest of the proof goes without changes, so that we get the desired bound by setting $\delta := \rho\sqrt{\frac{1}{d}}$. □

We will apply Lemma 3.6.2 to bound the second and third terms in (3.6.12) in the same way as Lemma 3.5.3 was applied to prove Lemma 3.5.4. Resulting bound

takes the following form:

$$\mathbb{E}\alpha(\delta,\Delta) \le C\left[\delta\sqrt{\frac{dM}{n}} \bigvee \rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} w_n\left(\rho\sqrt{\frac{1}{d}}\right) \bigvee \right. \tag{3.6.15}$$
$$\bigvee \Delta\left(\mathcal{D}\sqrt{\frac{\Omega(\mathcal{D}/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} w_n\left(\mathcal{D}\sqrt{\frac{1}{d}}\right)\right) \bigvee$$
$$\left. \bigvee U(L,\mathcal{H})\frac{\Omega_n(\rho)}{n}\right].$$

The final step consists in applying Adamczak's version of Talagrand's concentration inequality (see Theorem 1.2.5) and making the bound uniform with respect to δ, Δ in a way similar to Lemma 3.5.4. For fixed δ, Δ we have

$$\alpha(\delta,\Delta) \le K\left[\mathbb{E}\alpha(\delta,\Delta) + \delta\sqrt{\frac{tM}{n}} + C\Psi\frac{t\log n}{n}\right], \tag{3.6.16}$$

where we used the fact that the envelope for the class $\{f_\lambda - f_{\bar\lambda},\ \lambda \in \Lambda(\delta,\Delta)\}$ satisfies

$$\sup_{\lambda\in\Lambda(\delta,\Delta)} |(f_\lambda - f_{\bar\lambda})(X)| \le \|\lambda - \bar\lambda\|_1 F(X) \le 2F(X),$$

and $\left\| \max_{1\le i\le n} F(X_i)\right\|_{\psi_1} \le C\log n\Psi$ by the properties of Orlicz norms. The uniform version of (3.6.16) looks as follows: for all $n^{-1/2} \le \delta \le \sqrt{2}\mathcal{D}$ and $n^{-1/2} \le \Delta \le 1$ simultaneously

$$\alpha(\delta,\Delta) \le \beta(\delta,\Delta) := \bar{C}\left[\delta\sqrt{\frac{dM + t_n}{n}} \bigvee \rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} w_n\left(\rho\sqrt{\frac{1}{d}}\right) \bigvee \right.$$
$$\bigvee \Delta\left(\mathcal{D}\sqrt{\frac{\Omega(\mathcal{D}\sqrt{M}/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} w_n\left(\mathcal{D}\sqrt{\frac{1}{d}}\right)\right) \bigvee \tag{3.6.17}$$
$$\left. \bigvee U(L,\mathcal{H})\frac{\Omega_n(\rho) + t_n\log n}{n}\right],$$

with probability $\ge 1 - e^{-t}$; here, $t_n := t + c\log\log n$.

Finally, set $\hat\delta := \|f_{\hat\lambda} - f_{\bar\lambda}\|_{L_2(\nu)}$ and $\hat\Delta := \int\limits_{\mathcal{H}\setminus\mathcal{H}(\bar w)} \hat\lambda d\mu$. As before, the cases when $\hat\delta < n^{-1/2}$ or $\hat\Delta < n^{-1/2}$ have to be handled separately by replacing $\hat\delta$ or $\hat\Delta$ with

its upper bound. To complete the proof in the main case $n^{-1/2} \leq \hat{\delta} \leq \sqrt{2}\mathcal{D}$ and $n^{-1/2} \leq \hat{\Delta} \leq 1$, consider the inequality (obtained from (3.6.10), (3.6.11) and (3.6.17)):

$$\begin{aligned} \|f_{\hat{\lambda}} - f_{\lambda_*}\|_2^2 + \frac{1}{2}\hat{\delta}^2 + \frac{\varepsilon}{2}\hat{\Delta} \leq \\ \leq \|f_{\bar{\lambda}} - f_*\|_2^2 + \frac{1}{2}\varepsilon^2\gamma^2(\bar{w}) + 2\beta(\hat{\delta}, \hat{\Delta}). \end{aligned} \tag{3.6.18}$$

Assume that $\bar{D}$ from the assumptions of the Theorem satisfies

$$\bar{D} \geq \bar{4}C$$

where $\bar{C}$ is the constant from (3.6.17). Using the inequality

$$C\hat{\delta}\sqrt{\frac{dM + t_n}{n}} \leq \frac{1}{2}\hat{\delta}^2 + C_1\frac{dM + t_n}{n},$$

we deduce from (3.6.18) that

$$\begin{aligned} \|f_{\hat{\lambda}} - f_{\lambda_*}\|_2^2 + \frac{\varepsilon}{4}\hat{\Delta} \leq \|f_{\bar{\lambda}} - f_*\|_2^2 + \frac{1}{2}\varepsilon^2\gamma^2(\bar{w}) + \\ C\left[\frac{dM + t_n}{n} \bigvee U(L, \mathcal{H})\frac{\Omega_n(\rho) + t_n \log n}{n} \bigvee \right. \\ \left. \bigvee \rho\sqrt{\frac{\Omega(\rho/\sqrt{d})}{n}} \bigvee \sqrt{\frac{1}{n}} w_n\left(\rho\sqrt{\frac{1}{d}}\right)\right], \end{aligned} \tag{3.6.19}$$

concluding the proof. □

Assume that complexity assumption (3.4.5) is satisfied so that bounds of Proposition 3.4.4 hold for a suitable function $T(u)$. In this case, our previous result implies the following:

Corollary 3.6.3. *Let $\bar{\lambda} \in \mathbb{D}$ and $\bar{w} \in \partial\|\bar{\lambda}\|_1$. There exist numerical constants C and $\bar{D}$ large enough such that for any*

$$\varepsilon \geq \bar{D}\left[\mathcal{D}\sqrt{\frac{T(2/\sqrt{d})}{n}}\right]$$

and any subspace L with $d = \dim(L)$ and $\rho = \rho(\mathcal{H}(\bar{w}); L)$ the following holds with probability $\geq 1 - e^{-t}$:

$$\|f_{\hat{\lambda}} - f_{\lambda_*}\|_2^2 + \frac{\varepsilon}{4} \int\limits_{\mathcal{H}\setminus\mathcal{H}(\bar{w})} \hat{\lambda} d\mu \leq \|f_{\bar{\lambda}} - f_*\|_2^2 + \frac{1}{2}\varepsilon^2\gamma^2(\bar{w}) + C\left[\frac{dM + t_n}{n} \bigvee \rho\sqrt{\frac{T(\rho/\Psi\sqrt{d})}{n}} \log\frac{\Psi\sqrt{d}}{\rho} \bigvee \bigvee U(L,\mathcal{H})\frac{T_n + t_n \log n}{n}\right],$$

where

$$T_n := T\left(\frac{\rho}{\Psi\sqrt{d}}\right)\log^2\left(\frac{\Psi\sqrt{d}}{\rho}\right) \bigvee \Psi^2 T\left(\frac{1}{\sqrt{n}}\right) \bigvee T\left(\sqrt{\frac{T(1/\sqrt{n})\log n}{\Psi n}}\right) \bigvee \log n$$

and $t_n = t + c\log\log n$.

In particular, if $T(u) = V\log\frac{A}{u}$, then $T_n \leq \log^3\left(\frac{\Psi\sqrt{d}}{\rho}\right) \bigvee \Psi^2\log(\Psi n)$. In many typical situations, $\rho \gtrsim \frac{1}{n}$. In this case, previous inequality can be further simplified to

$$\|f_{\hat{\lambda}} - f_{\lambda_*}\|_2^2 + \frac{\varepsilon}{4} \int\limits_{\mathcal{H}\setminus\mathcal{H}(\bar{w})} \hat{\lambda} d\mu \leq \|f_{\bar{\lambda}} - f_*\|_2^2 + \frac{1}{2}\varepsilon^2\gamma^2(\bar{w}) + C\left[\frac{dM + t_n}{n} \bigvee \frac{\rho}{\sqrt{n}}\log^{3/2}\frac{\Psi\sqrt{d}}{\rho} \bigvee U(L,\mathcal{H})\frac{\Psi^2\log^3\frac{\Psi\sqrt{d}}{\rho} + t_n\log n}{n}\right].$$

The bound of Corollary 3.6.3 has a clear intuitive meaning: if there exists $\bar{\lambda}$ such that $\|f_{\bar{\lambda}} - f_*\|_2^2$ is small, $\|\cdot\|_1$ has a "smooth" subgradient at the point $\bar{\lambda}$, and at the same time its support can be well approximated by a linear subspace of small dimension d, then $\|f_{\hat{\lambda}} - f_{\lambda_*}\|_2^2$ is also small, and most of the "weight" of $\hat{\lambda}$ is distributed over $\mathrm{supp}(\bar{\lambda})$. For examples and more details, see Subsection 3.7.4 below.

3.7 Examples

Below we consider few common examples of the base classes and show how to obtain upper bounds for the alignment coefficient.

3.7.1 Weakly correlated partitions

Let $\mathcal{H}_j,\ j = 1, \dots, N$ be a measurable partition of $\mathcal{H}$. We are interested in the situation when the number N of function classes $\mathcal{H}_j$ is large and they are "weakly correlated". As a concrete examples of such a partition, one can consider the case when $S = [0,1]^N$ and, for each $j = 1, \dots, N$, $\mathcal{H}_j$ is a class of functions depending on the j-th variable. For the problem of density estimation, a natural example is a situation when $h_i \in \mathcal{H}_i$ and $h_j \in \mathcal{H}_j,\ i \neq j$ implies $\mathrm{supp}(h_i) \cap \mathrm{supp}(h_j) = \emptyset$ (or, more generally, when the measure of intersection is "small"). This might be viewed as an extension to the case of infinite dictionaries of usual notions of "almost orthogonality" (such as, for instance, restricted isometry property mentioned in the introduction) frequently used in the literature on sparse recovery. It is also close to "sparse additive models" and "sparse multiple kernel learning", [70], [58]. Suppose there exist oracles $\lambda \in \mathbb{D}$ such that f_λ provides a good approximation of the target f_* and, at the same time, λ is "sparse" in the sense that it is concentrated mostly on a small number of sets $\mathcal{H}_j$. For technical purposes, we will provide slightly different constructions for prediction problems and for density estimation.

3.7.1.1 The case of prediction problems

For each set $\mathcal{H}_j$, let $K_j : L_2(\mathcal{H}_j; \mu) \mapsto L_2(\mathcal{H}_j, \mu)$ be the integral operator (self-adjoint and nonnegatively definite) defined by

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

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

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

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

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

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

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

Denote $\mathcal{L}_J = \text{l.s.}\left(\bigcup_{j\in J} \mathcal{L}_j\right)$ and let $\rho(J) := \sup\Big\{\rho_\Pi(f, g) : f \in \mathcal{L}_J, g \in \mathcal{L}_{J^c}\Big\}$. The quantity $\rho(J)$ should be compared with the notion of *canonical correlation* often used in the multivariate statistical analysis. It is easy to check (see [53], Proposition 7.1) that

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

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

Proposition 3.7.1. *For all $J \subset \{1, \dots, N\}$ and all $w = \sum_{j\in J} w_j$ with $w_j \in \mathrm{Im}(K_j^{1/2})$,*

$$\gamma(w) \leq \beta_2(J)\left(\sum_{j\in J} \bar{\gamma}^2(w_j)\right)^{1/2} \leq \beta_2(J)\left(\sum_{j\in J} \|K_j^{-1/2} w_j\|^2_{L_2(\mathcal{H}_j,\mu)}\right)^{1/2}. \tag{3.7.1}$$

Proof. Let u_j be such that $u = \sum_j u_j$ with $\mathrm{supp}(u_j) \subset \mathcal{H}_j$. Then

$$\begin{aligned}
\bar{\gamma}(w) &\leq \sup_{u:\sigma^2_\Pi(f_u)=1} \sum_j \langle w_j, u\rangle \leq \sup\left\{\sum_{j\in J} \langle w_j, u_j\rangle :\ \sum_{j\in J} \sigma^2_\Pi(f_{u_j}) \leq \beta_2^2(J)\right\} \leq \\
&\leq \sup\left\{\sum_{j\in J} \sigma_\Pi(f_{u_j})\bar{\gamma}(w_j) :\ \sum_{j\in J} \sigma^2_\Pi(f_{u_j}) \leq \beta_2^2(J)\right\} \leq \\
&\leq \beta_2(J)\left(\sum_{j\in J} \bar{\gamma}^2(w_j)\right)^{1/2}.
\end{aligned}$$

□

It is often convenient to assume that an "oracle density" is bounded below by a constant δ. If $\lambda := \sum_{j\in J} \lambda_j + \delta$, where $\delta \in (0, 1)$, λ_j are nonnegative functions defined

on $\mathcal{H}_j$ and

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

then $\log \lambda = \sum\limits_{j \in J} w_j I_{\mathcal{H}_j} + \log \delta$, where $w_j := \log(\lambda_j + \delta) - \log \delta$. Therefore, (3.7.1) implies

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

3.7.1.2 The case of density estimation

The notion of "almost orthogonality" for density functions $h_1,\ h_2$ can be interpreted in the following sense: there exist measurable disjoint sets $\mathcal{H}_1 \subset \mathrm{supp}(h_1),\ \mathcal{H}_2 \subset \mathrm{supp}(h_2)$ such that $\left| \int\limits_{\mathcal{H}_i} h_i d\nu - 1 \right|$ is small for $i = 1, 2$. This should be compared to the notion of *mutual singularity* of measures. Consequently, the definition of weakly correlated partitions is slightly different from the construction above, so we will outline the main differences. For each $\mathcal{H}_j$, define the Gram operator $K_j : L_2(\mathcal{H}_j, \mu) \mapsto L_2(\mathcal{H}_j, \mu)$ by

$$(K_j u)(h) := \int_{\mathcal{H}_j} \langle h, g \rangle_{L_2(\nu)} u(g) \mu(dg),\ h \in \mathcal{H}_j,$$

where we use notations of Section 3.6. We also define

$$\rho_\nu(h, g) := \frac{\langle h, g \rangle_{L_2(\nu)}}{\|f\|_{L_2(\nu)} \|g\|_{L_2(\nu)}}.$$

Let $\mathcal{L}_j$ be the subspace of $L_2(\nu)$ spanned by $\mathcal{H}_j$, and, for $J \subset \{1, \dots, N\}$, denote $\mathcal{L}_J = \mathrm{l.s.} \left(\bigcup\limits_{j \in J} \mathcal{L}_j \right)$ and

$$\beta_2(J) := \inf \Bigg\{ \beta > 0 : \forall f_j \in \mathcal{L}_j, j = 1, \dots, N \quad \sum_{j \in J} \|f_j\|^2_{L_2(\nu)} \le \beta^2 \left\| \sum_{j=1}^{N} f_j \right\|^2_{L_2(\nu)} \Bigg\}. \tag{3.7.2}$$

Given $h_j \in \mathcal{L}_j,\ j = 1, \dots, N$, denote by $\kappa(\{h_j : j \in J\})$ the minimal eigenvalue of the Gram matrix $\Big(\langle h_i, h_j \rangle_{L_2(\nu)} \Big)_{i,j \in J}$, and let

$$\kappa(J) := \inf \Big\{ \kappa\left(\{h_j : j \in J\}\right) : h_j \in \mathcal{L}_j, \|h_j\|_{L_2(\nu)} = 1 \Big\}.$$

If moreover

$$\rho(J) := \sup\Big\{\rho_\nu(f,g) : f \in \mathcal{L}_J, g \in \mathcal{L}_{J^c}\Big\},$$

then

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

The following analogue of Proposition 3.7.1 is straightforward:

Proposition 3.7.2. *For all $J \subset \{1,\dots,N\}$ and all $w = \sum_{j\in J} w_j$ with $\gamma(w_j) < \infty$ for every $j \in J$,*

$$\gamma(w) \le \beta_2(J)\Big(\sum_{j\in J}\gamma^2(w_j)\Big)^{1/2} \le \beta_2(J)\Big(\sum_{j\in J}\|K_j^{-1/2}w_j\|^2_{L_2(\mathcal{H}_j,\mu)}\Big)^{1/2} \tag{3.7.3}$$

3.7.2 Monotone functions dictionary and decision stumps

3.7.2.1 Monotone functions dictionary

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

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

Hence

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

3.7.2.2 Decision stumps and applications

Next, we will outline another approach that allows to obtain bounds for alignment coefficient in many cases, including the example above. This approach is based on direct correspondence between the kernels of gaussian processes with covariance operator K and the spaces of functions with finite alignment coefficient.

A well-known technique to obtain representations for the kernels of a gaussian process is related to the so-called Factorization theorem (see [65]). We just mention its important corollary. Assume that we are given a dictionary $\{h_t\}$ indexed by $t \in [0,1]$, where $h_t : S \mapsto \mathbb{R}$, S is a compact subset of $\mathbb{R}^k$, $d\Pi(x) = p(x)dx$ is such that $0 < c_1 \le p(x) \le c_2 < \infty$ for all $x \in S$. Finally, assume that $K(s,t) := \langle f_s, f_t \rangle_{L_2(\Pi)}$ is the covariance function of a gaussian process with continuos sample paths. Then

$$\gamma(w) < \infty \iff \exists v \in L_2(S, dx) : \ w(t) = \int_0^1 h_t(s)v(s)d\Pi(s). \tag{3.7.4}$$

Moreover, in this case $\gamma(w) = \|v\|_{L_2[0,1]}$.

We will use this technique to obtain the bound for alignment coefficient when

$$\left\{h_t(x) = I_{[0,t]}(x) - I_{(t,1]}(x), \ t \in [0,1]\right\},$$

where $x \in [0,1] := S$ and μ is the Lebesgue measure on $[0,1]$. This is a variant of so-called "decision stumps" used in binary classification. Applying (3.7.4) to the centered family $\left\{g_t(x) = h_t(x) - \int_0^1 h_t(s)p(s)ds, \ t \in [0,1]\right\}$, we get

$$\bar{\gamma}(w) < \infty \iff w(t) = \int_0^t v(s)p(s)ds - \int_t^1 v(s)p(s)ds,$$

or $w'(t) = 2v(t)p(t)$, where v is chosen such that $\int_0^1 v(s)p(s)ds = 0$ (this can always be done for the centered family). In particular, $w(0) = w(1) = 0$. Since we assumed that p is bounded away from 0 and ∞,

$$\bar{\gamma}^2(w) = \int_0^1 v^2(s)\frac{p^2(s)}{p^2(s)}ds \le c\int_0^1 \left(w'(s)\right)^2 ds \le c\|w\|^2_{\mathbb{W}^{2,1}[0,1]}.$$

Now we will apply this bound in the context of weakly correlated partitions.

Due to their simplicity, decision stumps constitute a rather poor family of "threshold classifiers", usually too small to contain a good prediction rule. However, their high-dimensional analogue is a popular choice for boosting - type algorithms, such as AdaBoost [34]. These algorithms combine *weak learners* (e.g., decision stumps) with properly chosen weights, and the resulting prediction rule often has very strong generalization properties. Our approach can be seen as a version of "regularized boosting", previously considered in [13].

Suppose that $X = (X_1, \dots, X_N) \in S := [0,1]^N$ and let $\mathcal{H}_j := \left\{h_t^{(j)}\right\}$, with

$$h_t^{(j)}(x) = I_{[0,t]}(x_j) - I_{(t,1]}(x_j),\ x = (x_1, \dots, x_N).$$

Assume that coordinate projections of X are independent, so that our "weak correlation" assumption holds, in particular, $\beta_2(J) = 1$ for any $J \subset \{1, \dots, N\}$. It is well known that the Vapnik-Chervonenkis dimension of decision stumps dictionary $\mathcal{H} := \bigcup_{j=1}^{N} \mathcal{H}_j$ is bounded by $V := 2(\log_2 N + 1)$. An implication of this fact is that complexity assumption (3.4.5) holds with $T(u) = (2V+1)\log\frac{1}{u}$ [91]. Moreover, assume that λ is a d-sparse oracle, meaning that $\lambda = \sum_{j=1}^{d} \lambda_j$ with $\mathrm{supp}(\lambda_j) \subseteq \mathcal{H}_{i_j}$, $1 \le i_1 < \ldots < i_d \le N$. Set $J := \{i_1, \dots, i_d\}$.

Let L_j be the subspace spanned by $\left\{h_{\frac{i}{M}}^{(j)},\ i = 1, \dots, M-1\right\}$, and define

$$L := \text{l.s.} \left(\bigcup_{j=1}^{d} L_{i_j} \right)$$

of dimension $\dim(L) = d(M-1)$. Clearly, $\|h_t^{(j)} - h_s^{(j)}\|_{L_2(\Pi)}^2 \le c|t-s|$, which implies

$$\rho\left(\mathrm{supp}(\lambda), L\right) \le c\frac{1}{\sqrt{M}}.$$

At the same time, note that, since L consists of piecewise-constant functions,

$$\inf_{f \in L, \|f\|_\infty = 1} \|f\|_{L_2(\Pi)}^2 \ge \frac{C(\Pi)}{M},$$

implying $U(L) \le C(\Pi)\sqrt{M}$.

Let $0 < \delta < \frac{1}{N}$ and define

$$\lambda_\delta := \delta + \sum_{j \in J} c_\delta \lambda_j,$$

so that $\log \lambda_\delta = \log \delta + \sum_{j \in J} \log\left(1 + \frac{c_\delta \lambda_j}{\delta}\right) I_{\mathcal{H}_j}$, where $c_\delta = 1 - N\delta$ is chosen such that $\int_{\mathcal{H}} \lambda_\delta d\mu = 1$. Note that $\int_{\mathcal{H} \setminus \bigcup_{j \in J} \mathcal{H}_j} \lambda_\delta \, d\mu = (N - d)\delta$ and

$$\bar{\gamma}^2(\log \lambda_\delta) \le \sum_{j \in J} \|w_{j,\delta}\|^2_{\mathbb{W}^{2,1}[0,1]} \le d \max_{1 \le j \le d} \|w_{j,\delta}\|^2_{\mathbb{W}^{2,1}[0,1]},$$

where $w_{j,\delta} = \log\left(1 + \frac{c_\delta \lambda_j}{\delta}\right)$. Finally, set $\varepsilon := D \log d \sqrt{\frac{\log N}{n}}$. It remains to apply the general inequality (3.5.6) to λ_δ and optimize over M. In particular, choosing

$$M_* = C \log(nd) \, (\log N)^{1/3} \, \frac{n^{1/3}}{d^{2/3}},$$

we get

Corollary 3.7.3. *With probability* $\ge 1 - e^{-t}$,

$$\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) \le \inf_{\delta > 0} \left[2\mathcal{E}(f_{\lambda_\delta}) + C \left(\frac{(d \log N)^{1/3}}{n^{2/3}} \log(nd) + \frac{d \log^2 d \log N}{n} \max_{1 \le j \le d} \|w_{j,\delta}\|^2_{\mathbb{W}^{2,1}} + \right.\right.$$
$$\left.\left. + \delta(N - d) \sqrt{\frac{\log N}{n}} + \frac{t}{n} \right) \right],$$

whenever $\log N < \frac{d^{2/3} n^{1/6}}{\sqrt{\log nd}}$.

If $N \ll e^n$, this inequality becomes meaningful when the oracle λ is such that $\max_{1 \le j \le d} \|w_{j,\delta}\|^2_{\mathbb{W}^{2,1}} \lesssim \log^\tau\left(\frac{1}{\delta}\right)$ for some $\tau > 0$. If this is the case, one can choose $\delta_{n,N} \asymp \frac{1}{(n \vee N)^2}$, which yields $|\mathcal{E}(f_\lambda) - \mathcal{E}(f_{\lambda_{1/N^2}})| \le C(N \vee n)^{-1}$ by Proposition 3.4.1, and the inequality can be further simplified to

$$\mathcal{E}(f_{\hat{\lambda}_\varepsilon}) \le 2\mathcal{E}(f_\lambda) + C \left[\frac{(d \log N)^{1/3}}{n^{2/3}} \log(nd) + \frac{d \log^2 d \log N}{n} \log^\tau(N \vee n) + \frac{t}{n} \right].$$

3.7.3 Fourier dictionary

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

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

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

$$\langle w, \lambda\rangle_{L_2(\mu)} = \langle w, \lambda m\rangle_{L_2(\mathbb{R}^d)} = \langle \widehat{w}, \widehat{\lambda m}\rangle_{L_2(\mathbb{R}^d)} = \langle \widehat{w}, f_\lambda\rangle_{L_2(\mathbb{R}^d)} = \left\langle \frac{\widehat{w}}{p^{1/2}}, f_\lambda p^{1/2}\right\rangle_{L_2(\mathbb{R}^d)},$$

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

3.7.4 Location families and generalizations

3.7.4.1 Location families on a torus

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

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

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

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

3.7.4.2 Dictionaries with stationary covariance functions

We will describe another similar example. Assume that $\mathcal{H} = \{h_t,\ t \in \mathbb{R}\}$ and μ is the Lebesgue measure. Moreover, assume that

(a) $\mathbb{E}h_t(X)h_s(X) = K(t-s)$ in the case of prediction problems, or

(b) $\langle h_t, h_s\rangle_{L_2(\nu)} = K(t-s)$ in the case of $L_2(\nu)$-density estimation,

and that K is continuous. By Bochner's theorem,

$$K(z) = \int\limits_{\mathbb{R}} e^{izx} d\gamma(x)$$

for some Borel measure γ. This gives $\|f_u\|^2_{L_2(\Pi)} = \int_{\mathbb{R}} |\hat{u}(x)|^2 d\gamma(x)$ for prediction problems or, similarly, $\|f_u\|^2_{L_2(\nu)} = \int_{\mathbb{R}} |\hat{u}(x)|^2 d\gamma(x)$ in the case of density estimation. If, moreover, $d\gamma(x) = v(x)dx$ and $v(x)$ is positive on $\mathrm{supp}(\hat{w})$, then

$$\langle w, u\rangle_{L_2(dx)} = \langle \hat{w}, \hat{u}\rangle_{L_2(dx)} = \left\langle \frac{\hat{w}}{\sqrt{v}}, \hat{u}\sqrt{v}\right\rangle_{L_2(dx)} \leq \|f_u\|_{L_2(dx)} \left\|\frac{\hat{w}}{\sqrt{v}}\right\|_{L_2(dx)},$$

implying $\gamma(w) \leq \left\|\frac{\hat{w}}{\sqrt{v}}\right\|_{L_2(dx)}$. If $v(x) \geq L(1+|x|^2)^{-\alpha}$, we get $\gamma(w) \leq A\|w\|_{\mathbb{W}^{2,\alpha}(dx)}$. When $h(\cdot)$ is a density with respect to Lebesgue measure and $K(\cdot)$ is generated by location family $\{h(\cdot - t),\ t \in \mathbb{T} \subset \mathbb{R}\}$, one can take $v(x) = c|\hat{h}(x)|^2$.

The following example is related to the density estimation problem. Let f_* be the unknown density of X with bounded support and such that $\|f_*\|_\infty < \infty$. Assume that $\mathcal{H} = \bigcup\limits_{j=1}^{N} \mathcal{H}_j$, where

$$\mathcal{H}_j = \{h_j(\cdot - \theta),\ \theta \in [0,1]\}$$

and h_j is a probability density function on a bounded interval $T_j \subset \mathbb{R}$ with respect to $\nu(dx) = dx$, and μ is the Lebesgue measure. We will also assume for simplicity

that supports of $h_i(\cdot - \theta)$ and $h_j(\cdot - \theta)$ are disjoint for any $i \neq j$ and $\theta \in [0,1]$ (in other words, $\mathrm{dist}(T_i, T_j) > 1$). In this case, $\beta_2(J) = 1$ for any $J \subset \{1, \dots, N\}$ (here we use the version of $\beta_2(J)$ adapted to density estimation, see (3.7.2)). The sparsity assumption can be understood in the following sense: suppose that there exists a good approximation of the unknown density f_* by the convex mixture of the elements of $\mathcal{H}_{i_j},\ j = 1 \dots d$ for an integer $d \ll N$. In other words, there exists an good oracle $\lambda = \sum\limits_{j=1}^{d} \lambda_j$ with $\mathrm{supp}(\lambda_j) \subset \mathcal{H}_{i_j}$. It is well-known [74] that for each $\mathcal{H}_j$ complexity assumption (3.4.5) is satisfied with $T_j(u) = V_j \log \frac{R}{u}$, hence it is also satisfied for $\mathcal{H}$ with

$$T(u) = \log N \vee \max_j T_j(u).$$

Additionally, we will make the following smoothness assumption on h_j:

$$h_j \in \Sigma(\beta, B, T_j)$$

for some $\beta > 0$, where $\Sigma(\beta, B, T_j)$ is the Hölder smoothness class, see Definition 2.1 in Chapter 2. Let $t_j^{(1)}, \dots, t_j^{(R)}$ be the uniform grid on T_j of mesh size $\tau_R = \frac{c_j}{R}$, and define L_j as a linear subspace spanned by a basis of piecewise-polynomial functions of degree at most $\lfloor \beta \rfloor + 1$ (see (2.3.1)), with $\dim(L) \leq R(\lfloor \beta \rfloor + 1)$. Alternatively, one can take the space spanned by B-splines with knots $t_j^{(0)}, \dots, t_j^{(R)}$ of dimension $\dim(L_j) = R$. We set

$$L := \text{l.s.} \left(\bigcup_{j=1}^{d} L_{i_j} \right)$$

of dimension $\dim(L) \leq C(\beta) \cdot dR$. By approximation properties of such spaces (see Section 2.5 for piecewise-polynomials and [29], Chapter 13 for B-splines),

$$\rho\left(\mathrm{supp}(\lambda), L\right) \leq C_2 \sqrt{\|f_*\|_\infty} \frac{1}{R^\beta}.$$

To get an upper bound on $U(L)$, we use the fact that both piecewise-polynomial spaces and B-splines satisfy

$$\|\phi_i\|_\infty \leq C(\beta)\sqrt{R},$$

where $\{\phi_i\}$ is the $L_2(dx)$-orthonormal basis of L_j. Together with an observation that at most $\lfloor\beta\rfloor+1$ of basis functions are nonzero at any given $x \in T_j$, we get from Hölder inequality

$$\sup_{g\in L, \|g\|_{L_2(dx)}=1} |g(x)| = \sup_{\sum \alpha_j^2 \le 1} \left|\sum \alpha_j \phi_j(x)\right| \le C(\beta)\sqrt{R}.$$

Consequently, if $r_{f_*} := \inf_{x\in \mathrm{supp}(f_*)} f_*(x) > 0$ (recall that by our assumptions f_* has bounded support), we have

$$U(L) = \sup_{g\in L, \|g\|_{L_2(\Pi)}=1} |g(x)| \le \frac{1}{\sqrt{r_{f_*}}} \sup_{g\in L, \|g\|_{L_2(dx)}=1} |g(x)| \le \frac{C(\beta)}{\sqrt{r_{f_*}}}\sqrt{R}.$$

Remark: quantity $\left(\inf_{x\in\mathrm{supp}(f_*)} f_*(x)\right)^{-1/2}$ can be replaced by any upper bound on $\sup_{f\in L, \|f\|_{L_2(\Pi)}=1} \|f\|_{L_2(dx)}$ which in general might depend on R.

Let $\varepsilon := D\sqrt{\frac{\log N \vee \log(nd)}{n}}$. It remains to substitute obtained expressions into the general inequality (3.6.3) and optimize over R. In particular, choosing

$$R_* = C\left(\log N \vee \log \frac{nd}{\|f_*\|_\infty}\right)^{\frac{3\beta}{2\beta+1}} \frac{(n\|f_*\|_\infty)^{1/2(1+\beta)}}{d^{1/(1+\beta)}},$$

we get an inequality with the leading error term of order $\frac{d^{\frac{\beta}{\beta+1}}}{n^{\frac{1+2\beta}{2+2\beta}}}$:

Corollary 3.7.4. *With probability $\ge 1-e^{-t}$, for any d-sparse λ*

$$\|f_{\hat\lambda} - f_*\|_2^2 + \frac{\varepsilon}{4}\int\limits_{\mathcal{H}\setminus\mathcal{H}(w)} \hat\lambda(x)dx \le \|f_\lambda - f_*\|_2^2 + C\left[\frac{\left(d^\beta \|f_*\|_\infty^{1/2}\right)^{1/(\beta+1)}}{n^{\frac{1+2\beta}{2+2\beta}}}\left(\log N \vee \log\frac{dn}{\|f_*\|_\infty}\right)^{\frac{3\beta}{2\beta+1}}\right.$$
$$\left. + \frac{d\,(\log N \vee \log nd)}{n}\max_{1\le j\le d}\left\|\frac{\hat w_j}{|\hat h_j|}\right\|_{L_2(dx)}^2 + \Theta\frac{(t+\log\log n)\log n}{n^{\frac{4\beta+3}{4\beta+4}}}\right],$$

where $w_j \in \partial\|\lambda_j\|_1$ are suitable smooth elements of subdifferentials such that $w = \sum_{j\in J} w_j \in \partial\|\lambda\|_1$, and

$$\Theta = \sqrt{\frac{R_*}{r_{f_*}}} = \sqrt{\frac{\left(\log N \vee \log\frac{nd}{\|f_*\|_\infty}\right)^{\frac{3\beta}{2\beta+1}} \frac{\|f_*\|_\infty^{1/2(1+\beta)}}{d^{1/(1+\beta)}}}{r_{f_*}}}.$$

depends on n and N only logarithmically.

The following modification of the previous example is also of some interest. Assume that

$$\mathcal{H} = \{h(\cdot - \theta),\ \theta \in [-T, T]\}$$

is the location family generated by a single density function $h \in \Sigma(\beta, B, \mathbb{R})$. As before, μ is the Lebesgue measure on $[-T, T]$, ν is the Lebesgue measure on $\mathbb{R}$, and the unknown density f_* has bounded support and satisfies $0 < r_{f_*} \le f_*(x) \le R_{f_*} < \infty$. In this case, sparsity is naturally understood in the following sense: assume that there exists an oracle λ such that $\mathcal{E}(f_\lambda) = \|f_\lambda - f_*\|_2^2$ is small, and moreover

$$\lambda = \sum_{j=1}^{d} \lambda_j,$$

where "spikes" λ_j have disjoint connected supports, and $\nu\left(\mathrm{supp}(\lambda)\right) \le Cd \ll T$. For such "$d$-sparse" oracles λ, the alignment coefficient of the subgradient $w \in \partial\|\lambda\|_1$ is often controlled by $\sigma = \min\limits_{i \ne j} [\mathrm{dist}(\mathrm{supp}(\lambda_i), \mathrm{supp}(\lambda_j))]$. In particular, if $\gamma(w) \le C\|w\|_{\mathbb{W}^{2,1}(\mathbb{R})}$, the latter norm does not exceed $C\sqrt{\frac{d}{\sigma}}$ for a properly chosen w.
The subspaces L_j are constructed as before, with interpolation knots spread uniformly over $\mathrm{supp}(\lambda_j)$. The analogue of Corollary 3.7.4 is the following statement:

Corollary 3.7.5. *With probability $\ge 1 - e^{-t}$, for any d-sparse λ and $w \in \partial\|\lambda\|_1$*

$$\|f_{\hat\lambda} - f_*\|_2^2 + \frac{\varepsilon}{4} \int\limits_{\mathcal{H}\setminus\mathcal{H}(w)} \hat\lambda(x)dx \le \|f_\lambda - f_*\|_2^2 + C\left[\frac{\left(d^\beta R_{f_*}\right)^{1/(\beta+1)}}{n^{\frac{1+2\beta}{2+2\beta}}} (\log dn)^{\frac{3\beta}{2\beta+1}} + \frac{d\log(nd)}{n}\left\|\frac{\hat w}{|\hat h|}\right\|_{L_2(dx)}^2 + \Theta\frac{(t+\log\log n)\log n}{n^{\frac{4\beta+3}{4\beta+4}}}\right],$$

where

$$\Theta = \sqrt{\frac{\left(\log\frac{nd}{R_{f_*}}\right)^{\frac{3\beta}{2\beta+1}} \frac{R_{f_*}^{1/2(1+\beta)}}{d^{1/(1+\beta)}}}{r_{f_*}}}.$$

depends on n only logarithmically.

3.8 Concluding remarks

Let us mention that techniques used to obtain oracle inequalities for the solutions of entropy penalized problems can be applied to obtain similar bounds for other types of penalization, such as L_1-norm penalty. A particular example of this type of results was presented in the previous section on density estimation.

One of our recent joint projects with V. Koltchinskii makes an attempt to obtain results for other types of prediction problems, in particular, for the functional regression model with subgaussian design. We proceed with a brief description of this model.

Let $\mathbb{T}$ be an index set equipped with the σ-algebra and a measure μ. Consider the following functional linear model:

$$Y = \int_{\mathbb{T}} X(t)\lambda_*(t)d\mu + \xi, \tag{3.8.1}$$

where $X := \{X(\omega, t),\ t \in \mathbb{T}\}$ is a *subgaussian* random process indexed by $\mathbb{T}$, $\lambda_* \in L_1(\mu)$ and ξ is a zero-mean random variable with $\sigma_\xi^2 := \mathbb{E}\xi^2 < \infty$ and independent of X. The goal is to estimate unknown λ_* based on a given sample from P. Clearly, (3.8.1) can be viewed as a special case of *dictionary learning*, with the dictionary consisting of point evaluation functionals $\mathcal{H} = \{\delta_t(\cdot),\ t \in \mathbb{T}\}$. Let $\mathbb{D}$ be a convex weakly compact subset of $L_1(\mu)$ and consider the following penalized risk minimization problem:

$$\hat{\lambda}_\varepsilon := \mathrm{argmin}_{\lambda\in\mathbb{D}}\Big[P_n(y - f_\lambda(x))^2 + \varepsilon\|\lambda\|_1\Big], \tag{3.8.2}$$

Assume $\mathbb{T} \subset \mathbb{R}^2$ and $X(t)$ represents an image. Often, the "relevant information" about this image will be concentrated on a small subset of $\mathbb{T}$. In this case, a reasonable approach is to characterize properties of $\hat{\lambda}_\varepsilon$ when λ_* is *sparse*, in a sense that $\{X(t) :\ t \in \mathrm{supp}(\lambda_*)\}$ can be well approximated by a linear subspace of small dimension. If this is the case, the performance of $\hat{\lambda}$ should be controlled by dimension of this subspace and other parameters describing the quality of approximation. In addition,

$\hat{\lambda}_\varepsilon$ should inherit the sparsity pattern of λ_*. This work is currently in progress, and we postpone the details now. Another interesting question is to consider sparse prediction problems (for example, usual regression) in the dictionaries that do not admit uniform upper bounds in sup-norm. This requires further extensions of techniques applied for density estimation with L_2 loss in Section 3.6.

REFERENCES

[1] Adamczak, R., "A tail inequality for suprema of unbounded empirical processes with applications to markov chains," *Electron. J. Probab*, vol. 13, pp. 1000–1034, 2008.

[2] Angluin, D., "Queries and concept learning," *Machine learning*, vol. 2, no. 4, pp. 319–342, 1988.

[3] Angluin, D., "Queries revisited," in *Algorithmic Learning Theory*, pp. 12–31, Springer, 2001.

[4] Audibert, J.-Y. and Tsybakov, A. B., "Fast learning rates for plug-in classifiers," *Preprint*, 2005. Available at: `http://imagine.enpc.fr/publications/papers/05preprint_AudTsy.pdf`.

[5] Audibert, J.-Y. and Tsybakov, A. B., "Fast learning rates for plug-in classifiers," *Ann. Statist.*, vol. 35, no. 2, pp. 608–633, 2007.

[6] Balcan, M.-F., Beygelzimer, A., and Langford, J., "Agnostic active learning," *J. Comput. System Sci.*, vol. 75, no. 1, pp. 78–89, 2009.

[7] Balcan, M.-F., Hanneke, S., and Wortman, J., "The true sample complexity of active learning," in *Proceedings of the Conference on Learning Theory*, pp. 45–56, 2008.

[8] Balcan, M., Broder, A., and Zhang, T., "Margin based active learning," *Learning Theory*, pp. 35–50, 2007.

[9] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D., "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.

[10] Bartlett, P., Mendelson, S., and Neeman, J., "ℓ_1-regularized linear regression: persistence and oracle inequalities," *Probability theory and related fields*, pp. 1–32, 2009.

[11] Bickel, P., Ritov, Y., and Tsybakov, A., "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[12] Birgé, L., "Model selection for density estimation with L_2-loss," *Arxiv preprint arXiv:0808.1416*, 2008.

[13] Blanchard, G., Lugosi, G., and Vayatis, N., "On the rate of convergence of regularized boosting classifiers," *The Journal of Machine Learning Research*, vol. 4, pp. 861–894, 2003.

[14] Bogachev, V. I., *Measure theory. Vol. I, II.* Berlin: Springer-Verlag, 2007.

[15] Bousquet, O., "A Bennett concentration inequality and its application to suprema of empirical processes," *C. R. Math. Acad. Sci. Paris*, vol. 334, no. 6, pp. 495–500, 2002.

[16] Bousquet, O., Koltchinskii, V., and Panchenko, D., "Some local measures of complexity of convex hulls and generalization bounds," in *Computational learning theory (Sydney, 2002)*, vol. 2375 of *Lecture Notes in Comput. Sci.*, pp. 59–73, Berlin: Springer, 2002.

[17] Bunea, F., Tsybakov, A., and Wegkamp, M., "Sparsity oracle inequalities for the Lasso," *Electronic Journal of Statistics*, vol. 1, pp. 169–194, 2007.

[18] Bunea, F., Tsybakov, A., and Wegkamp, M., "Sparse density estimation with ℓ_1 penalties," in *Proceedings of the 20th annual conference on Learning theory*, pp. 530–543, Springer-Verlag, 2007.

[19] Cai, T. and Low, M., "An adaptation theory for nonparametric confidence intervals," *The Annals of Statistics*, vol. 32, no. 5, pp. 1805–1840, 2004.

[20] Candes, E. and Tao, T., "The Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[21] Candès, E., "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, 2008.

[22] Candes, E., Romberg, J., and Tao, T., "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[23] Castro, R. M. and Nowak, R. D., "Minimax bounds for active learning," *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 2339–2353, 2008.

[24] Castro, R. M., Nowak, R. D., and Willett, R., "Faster rates in regression via active learning," *Technical Report ECE-05-3, University of Wisconsin - Madison*, 2005.

[25] Cohn, D., Atlas, L., and Ladner, R., "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[26] D. Cohn, L. A., Ladner, R., El-Sharkawi, M., Marks, R., Aggoune, M., and Park, D., "Training connectionist networks with queries and selective sampling," *Advances in Neural Information Processing Systems (NIPS)*, 1990.

[27] Dalalyan, A. and Tsybakov, A., "Aggregation by exponential weighting and sharp oracle inequalities," in *Proceedings of the 20th annual conference on Learning theory*, pp. 97–111, Springer-Verlag, 2007.

[28] Dasgupta, S., Hsu, D., and Monteleoni, C., "A general agnostic active learning algorithm," in *Advances in Neural Information Processing Systems 20*, pp. 353–360, Cambridge, MA: MIT Press, 2008.

[29] DeVore, R. and Lorentz, G., *Constructive approximation*, vol. 303. Springer, 1993.

[30] Devroye, L., Györfi, L., and Lugosi, G., *A probabilistic theory of pattern recognition*, vol. 31. New York: Springer-Verlag, 1996.

[31] Donoho, D., "Neighborly polytopes and sparse solution of underdetermined linear equations," *Technical Report, Stanford University*, 2005. Available at `http://www-stat.stanford.edu/~donoho/Reports/2005/NPaSSULE-01-28-05.pdf`.

[32] Donoho, D., "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*, vol. 59, no. 6, pp. 797–829, 2006.

[33] Dudley, R. M., *Uniform central limit theorems*, vol. 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press, 1999.

[34] Freund, Y. and Schapire, R., "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*, pp. 23–37, Springer, 1995.

[35] Freund, Y., Seung, H. S., Shamir, E., and Tishby, N., "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, pp. 133–168, 1997.

[36] Friedman, E., "Active learning for smooth problems," in *Proceedings of the 22nd Conference on Learning Theory*, vol. 1, pp. 3–2, 2009.

[37] Gaïffas, S., "Sharp estimation in sup norm with random design," *Statist. Probab. Lett.*, vol. 77, no. 8, pp. 782–794, 2007.

[38] Genovese, C. and Wasserman, L., "Adaptive confidence bands," *The Annals of Statistics*, vol. 36, no. 2, pp. 875–905, 2008.

[39] Giné, E. and Koltchinskii, V., "Concentration inequalities and asymptotic results for ratio type empirical processes," *Ann. Probab.*, vol. 34, no. 3, pp. 1143–1216, 2006.

[40] Giné, E. and Nickl, R., "Confidence bands in density estimation," *Ann. Statist.*, vol. 38, no. 2, pp. 1122–1170, 2010.

[41] Hanneke, S., "A bound on the label complexity of agnostic active learning," in *Proceedings of the 24th international conference on Machine learning*, pp. 353–360, ACM, 2007.

[42] Hanneke, S., "Rates of convergence in active learning," *Ann. Statist.*, vol. 39, no. 1, pp. 333–361, 2011.

[43] Hanneke, S., "Activized learning: Transforming passive to active with improved label complexity,"

[44] Hoffmann, M. and Nickl, R., "On adaptive inference and confidence bands," *The Annals of Statistics*, to appear.

[45] Ioffe, A. and Tikhomirov, V., "Theory of external problems," 1974.

[46] Kearns, M. J. and Vazirani, U. V., *An introduction to computational learning theory*. Cambridge, MA: MIT Press, 1994.

[47] Klein, T. and Rio, E., "Concentration around the mean for maxima of empirical processes," *Ann. Probab.*, vol. 33, no. 3, pp. 1060–1077, 2005.

[48] Koltchinskii, V., "Local rademacher complexities and oracle inequalities in risk minimization," *Ann. Statist.*, vol. 34, no. 6, pp. 2593–2656, 2006.

[49] Koltchinskii, V., "Sparsity in penalized empirical risk minimization," *Annales Inst. H. Poincare, Probabilites et Statistique*, 2007.

[50] Koltchinskii, V., "The Dantzig selector and sparsity oracle inequalities," *Bernoulli*, vol. 15, no. 3, pp. 799–828, 2009.

[51] Koltchinskii, V., "Sparse recovery in convex hulls via entropy penalization," *Ann. Statist.*, vol. 37, no. 3, pp. 1332–1359, 2009.

[52] Koltchinskii, V., "Rademacher complexities and bounding the excess risk in active learning," *J. Mach. Learn. Res.*, vol. 11, pp. 2457–2485, 2010.

[53] Koltchinskii, V., *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Springer, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour.

[54] Koltchinskii, V. and Minsker, S., "Sparse recovery in convex hulls of infinite dictionaries," COLT, 2010.

[55] Koltchinskii, V. and Panchenko, D., "Empirical margin distributions and bounding the generalization error of combined classifiers," *The Annals of Statistics*, vol. 30, no. 1, pp. 1–50, 2002.

[56] Koltchinskii, V., Panchenko, D., and Lozano, F., "Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins," *The Annals of Applied Probability*, vol. 13, no. 1, pp. 213–252, 2003.

[57] Koltchinskii, V. and Yuan, M., "Sparse recovery in large ensembles of kernel machines," in *21st Annual Conference on Learning Theory*, pp. 229–238, 2008.

[58] Koltchinskii, V. and Yuan, M., "Sparsity in multiple kernel learning," *The Annals of Statistics*, vol. 38, no. 6, pp. 3660–3695, 2010.

[59] Koltchinskii, V. and Panchenko, D., "Rademacher processes and bounding the risk of function learning," in *High dimensional probability, II (Seattle, WA, 1999)*, vol. 47 of *Progr. Probab.*, pp. 443–457, Boston, MA: Birkhäuser Boston, 2000.

[60] Koltchinskii, V. and Panchenko, D., "Complexities of convex combinations and bounding the generalization error in classification," *Ann. Statist.*, vol. 33, no. 4, pp. 1455–1496, 2005.

[61] Lang, S., *Real and functional analysis*, vol. 142. Springer, 1993.

[62] Ledoux, M. and Talagrand, M., *Probability in Banach spaces*, vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Berlin: Springer-Verlag, 1991. Isoperimetry and processes.

[63] Lepski, O., Mammen, E., and Spokoiny, V., "Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors," *The Annals of Statistics*, vol. 25, no. 3, pp. 929–947, 1997.

[64] Lewis, D. and Gale, W., "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–12, Springer-Verlag New York, Inc., 1994.

[65] Lifshits, M., *Gaussian random functions*, vol. 322. Springer, 1995.

[66] Lounici, K., "Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators," *Electronic Journal of statistics*, vol. 2, pp. 90–102, 2008.

[67] Lounici, K., *Estimation Statistique En Grande Dimension, Parcimonie Et Inégalités D'Oracle*. PhD thesis, 2009.

[68] Low, M. G., "On nonparametric confidence intervals," *Ann. Statist.*, vol. 25, no. 6, pp. 2547–2554, 1997.

[69] Mammen, E. and Tsybakov, A., "Smooth discrimination analysis," *The Annals of Statistics*, vol. 27, no. 6, pp. 1808–1829, 1999.

[70] Meier, L., Van De Geer, S., and Bühlmann, P., "High-dimensional additive modeling," *The Annals of Statistics*, vol. 37, no. 6B, pp. 3779–3821, 2009.

[71] Mendelson, S., "Empirical processes with a bounded ψ_1 diameter," *Geometric and Functional Analysis*, vol. 20, no. 4, pp. 988–1027, 2010.

[72] Mendelson, S. and Neeman, J., "Regularization in kernel learning," *The Annals of Statistics*, vol. 38, no. 1, pp. 526–565, 2010.

[73] Minsker, S., "Plug-in approach to active learning," *J. Mach. Learn. Res.*, pp. 67–90, 2012.

[74] Nolan, D. and Pollard, D., "U-processes: rates of convergence," *The Annals of Statistics*, vol. 15, no. 2, pp. 780–799, 1987.

[75] Resnick, S., *Heavy-tail phenomena: probabilistic and statistical modeling*, vol. 10. Springer Verlag, 2007.

[76] Rosset, S., Swirszcz, G., Srebro, N., and Zhu, J., "ℓ_1-regularization in infinite dimensional feature spaces," *Learning theory, Lecture Notes in Comput. Sci.*, vol. 4539, pp. 544–558, 2007.

[77] Schapire, R., Freund, Y., Bartlett, P., and Lee, W., "Boosting the margin: A new explanation for the effectiveness of voting methods," *The annals of statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[78] Settles, B., "Active learning literature survey," *Computer Sciences Technical Report 1648, University of Wisconsin–Madison*, 2010.

[79] Settles, B. and Craven, M., "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1070–1079, Association for Computational Linguistics, 2008.

[80] Talagrand, M., *The generic chaining*. Springer, 2005.

[81] Talagrand, M., "New concentration inequalities in product spaces," *Invent. Math.*, vol. 126, no. 3, pp. 505–563, 1996.

[82] Talagrand, M., "A new look at independence," *Ann. Probab.*, vol. 24, no. 1, pp. 1–34, 1996.

[83] Tibshirani, R., "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[84] Tong, S. and Koller, D., "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[85] Tsybakov, A. B., *Introduction to Nonparametric Estimation*. Springer, 2009.

[86] Tsybakov, A., "Optimal aggregation of classifiers in statistical learning," *The Annals of Statistics*, vol. 32, no. 1, pp. 135–166, 2004.

[87] Valiant, L. G., "A theory of the learnable," *Commun. ACM*, vol. 27, pp. 1134–1142, November 1984.

[88] van de Geer, S., "The deterministic Lasso," Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007.

[89] Van De Geer, S., "High-dimensional generalized linear models and the Lasso," *The Annals of Statistics*, vol. 36, no. 2, pp. 614–645, 2008.

[90] Van De Geer, S. and Bühlmann, P., "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[91] van der Vaart, A. W. and Wellner, J. A., *Weak convergence and empirical processes*. Springer Series in Statistics, New York: Springer-Verlag, 1996.

[92] Vapnik, V. N. and Chervonenkis, A. Y., *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow, 1974.

[93] Vapnik, V. N., *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control, New York: John Wiley & Sons Inc., 1998. A Wiley-Interscience Publication.

[94] Villani, C., *Optimal transport: old and new*, vol. 338. Springer Verlag, 2009.

[95] Wapnik, W. N. and Tscherwonenkis, A. J., *Theorie der Zeichenerkennung*, vol. 28 of *Elektronisches Rechnen und Regeln, Sonderband [Electronic Computing and Control, Special Issue]*. Berlin: Akademie-Verlag, 1979. Translated from the Russian by Klaus-Günter Stöckel and Barbara Schneider, Translation edited by Siegfried Unger and Klaus Fritzsch.

[96] Yang, Y., "Aggregating regression procedures to improve performance," *Bernoulli*, pp. 25–47, 2004.