Risk Bounds for Improper Prediction Procedures

Tomas Vaškevičius

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2021

To my wife

Contents

Acknowledgements

I would first like to thank my supervisors, Varun Kanade and Patrick Rebeschini, who have made my studies a truly wonderful experience by creating a relaxed, friendly, and stimulating environment. Varun and Patrick have encouraged me to explore a variety of topics of my interest while always being available, both on a technical and personal level. Through their guidance, advice, and many discussions, Varun and Patrick showed me what high-quality research is supposed to look like and what it really means to think deeply about a research problem. Although I am not there yet, my supervisors have instilled a benchmark that I will aim at for the years to come, regardless of what I end up doing.

I am grateful to Jared Tanner and Gábor Lugosi for agreeing to serve as my thesis examiners. Jared has dedicated his time to examine my work on two previous occasions and provided valuable feedback and advice. Books co-authored by Gábor, specifically on concentration and sequential prediction, have been invaluable resources, without which some of the contributions presented in this thesis would not have been possible.

Two chapters of this thesis are based on works performed in collaboration with Jaouad Mourtada and Nikita Zhivotovskiy. During our numerous discussions, I was impressed and inspired by Jaouad's and Nikita's mathematical proficiency, both of whom have immensely affected my own technical understanding of the topics investigated in this thesis.

My deepest gratitude goes to my parents and my brother, whose love, sacrifice, encouragement, support and advice had shaped and influenced my life in profound ways long before I started my studies at Oxford. I also thank my wife Ieva for being my source of inspiration, making sure I follow my dreams, and most importantly, for her company, which never ceases to bring me joy, happiness and meaning. I am grateful to the rest of my family for always making me feel welcome and believing in me.

Finally, I would like to thank the EPSRC and MRC for funding my studies through the OxWaSP CDT programme (EP/L016710/1).

Abstract

Statistical Learning Theory studies the problem of learning an unknown relationship between observed input-output pairs sampled independently from some unknown and arbitrary probability distribution. The quality of the inferred relationship is judged by its excess risk – a measure of prediction capability on unseen data compared to the best predictor in some predefined reference class of functions. A learning procedure is said to be improper if it is allowed to output a prediction rule outside the chosen reference class. This thesis presents four contributions to analyzing the prediction performance of improper learning algorithms.

Our first result contributes to developing the mathematical machinery suitable for the analysis of improper learning procedures. We obtain exponential-tail excess risk bounds in terms of offset Rademacher complexity for which only in-expectation guarantees were previously obtained.

Our second result shows that offset Rademacher complexity yields upper bounds on the excess risk of iterative regularization schemes characterized by mirror descent algorithms. Moreover, by providing a unified analysis, our proposed proof technique circumvents the limitations of some previous analyses tailored to exploit the exact form of specific iterative schemes.

Our third contribution concerns the analysis of the constrained linear least squares algorithm. We find that this classical and widely studied statistical estimator is suboptimal, with dimension-dependent excess risk improvements offered via known improper procedures.

In our fourth contribution, we investigate improperness in the linear regression problem with the squared loss, without imposing any assumptions on the distribution of the covariates and imposing a minimal assumption on the conditional distribution of the response variable. We first establish the in-expectation optimality of the truncated least squares estimator and then, we show that it can fail with constant probability. We conclude by proposing a deviation-optimal procedure. The considered setup admits heavy-tailed distributions while falling outside the scope of the typically studied procedures for heavy-tailed linear regression.

1 Introduction

This thesis presents some contributions to the analysis of improper statistical procedures. We provide a summary of our main results and outline the structure of this thesis in Section 1.7. Before turning to Section 1.7, in order to contextualize our results appropriately, we shall overview some historical developments of statistical learning and provide some motivation for the topics investigated in this thesis, keeping the discussion at a relatively high and informal level.

Given access to a finite collection of observations, represented as input-output pairs, a learning algorithm aims to construct a functional relationship linking the inputs to the outputs. The learning process succeeds if the inferred function *generalizes* beyond the seen example observations. Although we will stick to an abstract setting and we will not consider specific applications in this thesis, let us provide an example. The observations may consist of a collection of medical images, each image (input) associated with a binary label (output) denoting the presence or absence of abnormal tissues indicating the possibility of a disease. In this case, learning the unknown relationship between inputs and outputs would be of high practical interest since a successfully learned function that generalizes beyond the seen images may aid medical diagnoses in the future.

The learning process described above is inductive: given access to a specific finite sample of observations, we aim to extract general principles that will allow us to make predictions in the future. Arguably, whether such a process is at all possible is not self-evident, and several natural questions may be raised. For instance, consider the example of the medical images discussed above. What algorithm should be used to infer the underlying functional relationship? How many labelled images does the chosen algorithm need in order to guarantee, *with high confidence*, that the learned function does indeed generalize to previously unseen examples? More broadly, what does it mean to generalize, when is it possible, and what are the general algorithmic principles allowing to do so? Such questions are formalized and studied within the mathematical framework of *Statistical Learning Theory*. The contributions of this thesis are presented within this framework.

1.1 Statistical Learning Theory

Vapnik and Chervonenkis laid theoretical foundations of Statistical Learning Theory in a series of pioneering works initiated in the late 60s [205, 208, 206, 207]. Three

components constitute a general learning problem setup:

- an *unknown mechanism* that generates the data;
- a *loss function* used to measure the quality of different prediction rules;
- a *reference class* (also called hypothesis class) of functions, serving as a performance benchmark.

The interested reader will find an extensive background on the subject in the textbooks by Devroye, Györfi, and Lugosi [60], Vapnik [204], Hastie, Tibshirani, and Friedman [81], Györfi, Kohler, Krzyżak, and Walk [78], Wainwright [216], and in the references therein. We will now describe the three elements listed above in more detail.

For an inductive inference process to succeed – in the sense that inferences about the general population can be made from a specific sample – it is natural to require that the population and the sample are, in some way, coupled; otherwise, learning is impossible. The statistical learning approach of coupling the specific data with the general population proceeds by assuming the existence of an *unknown probability distribution*, from which the training data arises as independent identically distributed samples. The quality of any prediction rule – a functional relationship between inputs and outputs – is then measured by the quality of predictions made on new samples drawn independently from *the same* probability distribution.

Let us highlight that no specific form on the probability distribution is typically assumed in Statistical Learning Theory, apart from possibly constraining its support, for example, to constrain inputs and outputs to lie within some bounded sets. We call such a setup *distribution-free*. In contrast, the more classical statistical inference approaches tend to constrain the possible data-generating distributions to lie in some well-behaved family of distributions (e.g., assuming that the data is generated by some true linear model, with observations perturbed with independent Gaussian noise). Another distinctive characteristic from the classical statistical inference is that in Statistical Learning Theory, we typically do not aim to understand the underlying phenomenon (the distribution generating input-output pairs). Instead, we restrict ourselves to a more manageable problem of making predictions. The goal of making good predictions is easier than understanding the underlying distribution since once the distribution is understood, accurate predictions can be made, but not vice-versa. In the framework of Statistical Learning Theory, the assumption that the data samples are drawn independently from the same distribution allows bringing in various powerful probabilistic tools to study the problem. At the same time, the i.i.d. data assumption

is not always practical and justified, but nevertheless, it provides a rich enough model for theoretical explorations.

To provide formal comparisons between the quality of different input-output functional relationships, we need to assign quantitative estimates on their prediction performance. Hence, the *loss function* is introduced; taking the goals of the learner and the specifics of a given problem into account, it assigns a cost to the predictions (on some input variable) based on some notion of closeness to the actual output variable. For binary classification problems, the classical choice is the zero-one loss function, which assigns a cost of one for making an incorrect prediction and a zero cost for guessing the true output variable correctly. For regression problems, the typical choice is the quadratic loss, which assigns a cost of the squared difference between the prediction and the true output. Once the loss function is chosen, we assign a score called *risk* for each prediction function. It is defined as the *expected loss* incurred by the predictor applied to a fresh input drawn from the unknown data generating distribution.

Observe that the risk assigned to different prediction procedures is an absolute performance measure. For example, for noisy learning problems, there may not exist a function that incurs zero risk, and the best possible risk may depend on the noise level in the problem. As a form of normalization of the risk, the third component of a general learning problem is introduced: a *reference class* of functions. After choosing the reference class of functions, the quality of a learned prediction rule is judged by its *excess risk* – the difference between the risk of the selected function and the minimum risk among all the functions contained in the chosen reference class. Ideally, we would like the reference class to include all measurable functions; however, this is typically impossible. For example, under the zero-one classification loss function, it is only possible to match the performance (in a distribution-free manner) of reference classes of functions that satisfy a certain combinatorial condition, namely, their VC dimension needs to be finite.

Once the loss function and the reference class of predictors is fixed, the primary object of study in Statistical Learning Theory is the excess risk attained by a chosen learning algorithm. It is worth to one more time emphasize the point that in this thesis, we follow the classical approach of Statistical Learning Theory, where the underlying data generating distribution is unconstrained, except for possibly the support of inputs and outputs, ensuring certain boundedness conditions. In particular, in the results presented in this thesis, we shall never assume that the observed data consists of noisy predictions made by one of the reference functions, a setup frequently

studied within the more classical statistical approaches. Another fitting characteristic of the Statistical Learning Theory point of view is that we always seek to obtain *non-asymptotic* excess risk bounds, that is, bounds that hold for any finite sample size. This is not merely a technicality – conclusions drawn from non-asymptotic and asymptotic analyses can differ drastically in some settings. As a simple example, this difference can already be seen in a scalar mean estimation problem, where the observations are generated from an unknown real-valued distribution with bounded variance. For this problem, the sample mean estimator is optimal from the point of view of asymptotic performance bounds, yet it can have poor performance for finite sample problems, with significant gains offered by other statistical estimators in some cases; we refer to the work of Catoni [44] for details.

1.2 Empirical Risk Minimization

After formalizing a learning problem with appropriate choices of the loss function and the reference class of predictors, we are left with the question of how to exploit the observed data sample to construct a mapping between inputs and outputs that attains small excess risk (i.e., a mapping whose prediction performance is close to the best function in the chosen reference class). In Statistical Learning Theory, a central algorithmic principle for selecting a prediction function is called *empirical risk minimization* (ERM). An empirical risk minimization algorithm selects a function minimizing an empirical approximation of the risk computed on the observed data sample, among some predefined *hypothesis space* of admissible predictors. To distinguish between the true risk and its empirical approximation, we will sometimes call the former *population risk*, while the latter will be called *empirical risk*.

The ERM principle is fundamental to Statistical Learning. In his book [204], Vladimir Vapnik attributes Novikoff's [161] proof of convergence (obtained in the early 60s) for Rosenblatt's perceptron algorithm [181, 180] as the origin of the mathematical analysis of learning unknown functional relationships from finite data samples (in other words, it marks the beginning of Learning Theory). The perceptron algorithm attempts to learn parameters of a hyperplane that correctly separates the training dataset into two regions, provided that such a hyperplane exists. The procedure is performed by successively feeding the data to the algorithm, and upon each mistake, adjusting corresponding weights defining the hyperplane; thus, the perceptron algorithm is a particular instantiation of the ERM principle. Novikoff's proof shows that after a bounded number of corrections, depending on the Euclidean norms of the inputs and the margin of the optimal separating hyperplane, the training data will be correctly

separated. The bounded mistakes property implies that in a certain sense, the perceptron algorithm generalizes. Indeed, if there exists a hyperplane that perfectly separates the population points into two regions, after processing a finite stream of data, the perceptron algorithm will correctly classify all the future inputs (or misclassify only a tiny fraction of them).

The ERM principle was studied extensively in the landmark works of Vapnik and Chervonenkis, primarily in the binary classification setting of the zero-one loss, and when the hypothesis class over which the empirical risk minimizers are considered is taken to be equal to the reference class of functions, against which the prediction performance is measured. It was shown that in such settings, learnability is possible if and only if the reference class of predictors is a *Glivenko-Cantelli* class for the underlying data generating distribution, informally, meaning that uniformly for all functions of the class, their empirical and true means converge to zero with increasing sample size[1]. Naturally, Glivenko-Cantelli classes are learnable via the empirical risk minimization principle since uniform convergence ensures that empirical risks serve as good approximations for the population risks, simultaneously for all members of the reference class of predictors.

Vapnik and Chervonenkis showed [208] that a function class is Glivenko-Cantelli simultaneously for any data generating distribution if and only if it satisfies a certain combinatorial condition, specifically, if it is of finite VC dimension (named after Vapnik and Chervonenkis). Thus, distribution-free learnability in the binary classification setting is only possible for reference classes of functions that are not too complex, in a precise sense characterized by the notion of VC dimension. In addition, the VC dimension can be used to obtain sharp non-asymptotic distribution-free bounds of the excess risk of the ERM algorithm.

Moving beyond the binary classification setting, obtaining sharp excess risk bounds becomes a more subtle topic. If the loss function is convex and curved (e.g., the quadratic loss, as opposed to the flat zero-one classification loss) and in addition, if the reference class of functions is convex, then uniform convergence bounds computed over the whole reference class of functions does not capture the true non-asymptotic rate of convergence of the ERM algorithm. Instead, the sharpest known complexity measures in such regimes compute uniform convergence bounds of some *localized* subsets of

[1]See also the work of Shalev-Shwartz, Shamir, Srebro, and Sridharan [188], where a more general learning setting is considered. Learnability is shown there to be equivalent to a certain algorithmic stability property, which may hold without uniform convergence.

the reference class of the given reference class of functions; we provide a summary of localized complexity measures in Section 2.5.

The mathematical machinery of localization is necessary to take into account the following fact. When the loss function is curved and the reference class of predictors is convex, a variance compensation mechanism is automatically present for the ERM algorithm. However, as we will discuss below, over the developments of the past two decades, it became clear that convexity of the reference function class is not strictly needed – by resorting to more intelligent learning principles, it is possible to induce a similar variance compensation mechanism without the usual convexity assumption (we defer a brief explanation to Section 1.4). However, the existing localized complexity measures either do not apply to such algorithms or only provide distribution-free bounds that hold on average without providing any confidence estimates. One contribution of this thesis, presented in Chapter 3, demonstrates how to obtain excess risk bounds in terms of localized complexity measures that hold with high probability and apply to procedures beyond ERM, including examples not covered via the classical theory.

Despite its suboptimality in some regimes, the ERM principle remains a salient subject in ongoing research. Moreover, it serves as an important building block of more intricate algorithmic procedures. Chapter 5 of this thesis will present some contributions to the analysis of ERM. In particular, we will take a look at one of the most basic settings: learning a linear predictor with bounded Euclidean norm with the quadratic loss. Despite the benign problem structure exhibiting the desired curvature and convexity properties mentioned above, we will show that ERM incurs a suboptimal dimension-dependent factor on its excess risk in the distribution-free model of learning. We will also show that improper procedures can circumvent the shortcomings of ERM, and we will further develop this observation in Chapter 6.

1.3 Regularization

While algorithms minimizing the empirical risk among a given class of reference functions form a cornerstone of Learning Theory, a pertinent question remains unaddressed on selecting an appropriate the reference class. For example, suppose we must choose between two functions obtained as empirical risk minimizers of two nested hypothesis classes. The function that arises as an empirical risk minimizer of the larger of the two nested classes has at least as good data fit as the other function, where the data fit is reflected via the empirical risk. However, in the larger class, uniform convergence bounds become worse, and hence, we have less confidence that the obtained empirical

risk is close to the population risk. Consequently, a trade-off needs to be considered between the data fit and some notion of complexity of the hypothesis space, such as its VC dimension in the setting of binary classification.

A solution for appropriately balancing the complexity of the chosen function class and the obtained data fit, called *structural risk minimization*, was proposed by Vapnik and Chervonenkis [206]. The idea is to perform empirical risk minimization over nested classes of functions. Then, the quality of the obtained empirical risk minimizers is judged by the sum of their empirical risk (goodness of data fit, which gets better as the function class gets larger) and the uniform convergence bound expressed in terms of the complexity of the given function class (a measure of confidence in how close the empirical and population risks are, which gets worse as the function class gets larger). Observe that characterizing the performance of ERM for a given reference class in terms of its complexity plays a fundamental role in this scheme.

Ideas closely related to structural risk minimization – a *regularization* approach trading off empirical data fit with some notion of complexity – have been rediscovered across different fields. In fact, such ideas have already been put forward prior to the theoretical investigations in the context of Statistical Learning Theory. For example, already in the early 60s, similar regularization principles were proposed by Ivanov [90] and Tikhonov [195] in the literature on ill-posed problems; see [204, Chapter 4] for an overview of the history related to regularization.

While structural risk minimization described above relies on the ERM principle as its constituent part, other inductive principles of learning from finite samples of observations exist, a prominent example being *stochastic approximation* taking its roots in the early work of Robbins and Monro [176]. The idea of stochastic approximation is to use each data point only once by exploiting it to perform a step of some optimization procedure, traditionally the method of gradient descent. Consequently, the algorithm can be seen as directly minimizing the population risk instead of the empirical risk. If the data points are reused an infinite number of times, this principle reduces to empirical risk minimization since the optimization process would converge to a function with the best data fit (under additional assumptions). However, reusing the data points but stopping the optimization process before convergence results in another regularization principle, called *early stopping* or *iterative regularization*.

Compared to other regularization schemes such as the structural risk minimization discussed above, iterative regularization comes with built-in attractive computational features, as new models are brought into consideration at the cost of one step of an

optimization algorithm; this can be computationally cheaper than finding empirical risk minimizers in a sequence of nested hypothesis classes.

Chapter 4 of this thesis presents some contributions to the theory of iterative regularization. We will show how statistical analysis for a large family of early-stopped optimization algorithms can be carried out using tools known to yield sharp performance bounds for the ERM principle. In particular, through an adaptation of classical convergence arguments used in Convex Optimization literature to the statistical setting considered in this thesis, we will see that the notion of localized uniform convergence complexity measures appears naturally in the analysis of iterative regularization schemes induced by first-order gradient-based optimization methods.

1.4 Improper Prediction Procedures

We now turn to improperness – a central concept in this thesis that unifies all the contributions presented in Chapters 3 to 6.

We say that a prediction procedure is *improper* if, given a reference class of functions, the procedure is free to output functions outside this class. Thus, improperness is only defined within the context of having fixed some reference class of functions. For example, if one fixes all linear functions as a reference class, an improper procedure can make predictions using non-linear functions (e.g., truncated linear functions). In our discussions above, we considered algorithms that output a function from the given reference class minimizing the empirical risk. Such algorithms are proper.

Improperness has both computational and statistical consequences. While in this thesis, we will focus exclusively on the statistical aspects of learning, from the computational perspective, improperness allows escaping certain computational barriers. The key idea is that a successful statistical learning algorithm, restricted to selecting a hypothesis from a given class, in its solution could encode some combinatorial structure that is difficult to find without exhaustive search. A prototypical example is a binary classification setting, where the input space is a boolean hypercube, and the reference class of predictors is all boolean functions representable as disjunctions having three terms, each term being a conjunction of boolean variables. Then, a proper learning algorithm that successfully learns the underlying boolean formula for any data generating distribution can be transformed into an algorithm that, with non-zero probability, solves a problem (vertex coloring with three colors) known to be computationally intractable under widely believed computational complexity conditions (that the randomized polynomial time and nondeterministic polynomial time complexity classes are not equal). However, improper learning procedures – algorithms

not forced to represent the output formulas in the form prescribed above – can solve the problem efficiently from both computational and statistical perspectives. For this line of work, we refer to the seminal paper of Valiant [201], which introduced the field called Computational Learning Theory, where learnability is studied from the computational complexity theoretic point of view. See also the textbook by Kearns and Vazirani [97].

From the statistical perspective, improper learning algorithms allow us to obtain excess risk rates not achievable via proper algorithms in many problem setups. At first, this might seem somewhat counterintuitive because from the perspective of uniform convergence bounds, selecting a hypothesis from a larger class than the given reference class of functions can only deteriorate the algorithm's statistical performance. However, this reasoning turns out to be incorrect if the loss function possesses enough curvature. We have already discussed above that if the loss function is convex and curved, and the reference class is convex, the (proper) ERM estimators are considered to be good since their excess risk can be controlled by localized uniform convergence bounds known to be sharp in many instances (the applicability of such bounds will be extended to a class of improper estimators in Chapter 3 of this thesis). Furthermore, as we already remarked, the curvature and convexity conditions, together, result in a certain variance reduction mechanism for the ERM estimator. To briefly explain this mechanism, we will now informally argue that any two hypotheses in the reference class are automatically discarded from consideration, provided that they have comparable empirical risks and are far away from each other. Indeed, due to the convexity and curvature of the loss, some function in-between the two hypotheses will have smaller empirical risk, and hence, such a function will be a better choice for the ERM estimator since, due to the convexity assumption, this function is contained in the reference class of predictors. Thus, the effective size of the hypothesis space is automatically reduced, allowing us to take advantage of the localized uniform convergence machinery.

On the other hand, if the reference class of functions is non-convex, the above argument fails. However, the same variance reduction effect can be achieved by switching to improper statistical estimators that select functions from a carefully enlarged reference class. Let us remark that simple strategies such as returning any empirical risk minimizer in the convex hull of the original reference class of functions do not work. Optimal statistical performance is achieved by cleverly designed estimators that add just enough additional functions into consideration to induce the above-described variance compensation effect, but not too many functions so that a relatively small complexity of the original reference class is maintained. Perhaps the simplest

problem where improperness is necessary is that of *model selection aggregation* (cf. Nemirovski [155], Tsybakov [198]), where the reference class of functions is finite and the loss is quadratic. Due to the finite cardinality of the reference class, it is a non-convex class, and as a result, the ERM principle fails to achieve minimax optimal excess risk rates. The suboptimality of ERM can already be present for reference classes that contain only two functions. There are more known cases where improperness is needed to achieve optimal performance; we review the literature from the statistical perspective in Section 2.2.

In the context of model selection aggregation, one more phenomenon related to improperness was uncovered by Audibert [6]. Because the excess risk of the output hypothesis may take negative values in the improper learning settings, excess risk bounds that hold only on average may be suboptimal in high-probability regimes, where a confidence interval attached to the obtained bound is desired. More examples of this sort will be presented in Chapter 6 of this thesis, exhibiting two improper estimators that achieve optimal excess risk in expectation; however, both estimators incur a constant excess risk with a constant probability (exhibiting complete failure with constant probability).

Let us also remark that iterative regularization schemes investigated in Chapter 4 of this thesis can also be considered improper, and as such, we will find it fitting to analyze their statistical performance with mathematical tools suitable in improper learning settings. Finally, in Chapter 5, we will discuss that even in the classical problem of linear regression with the quadratic loss, with a bounded and convex reference class of linear predictors, the ERM principle still fails to achieve optimal distribution-free bounds, despite all the desirable problem characteristics discussed above. Following up on this observation, in Chapter 6, we will construct a statistically optimal (but computationally intractable) improper procedure for the problem of linear regression with the quadratic loss, under minimal distributional assumptions in a sense formalized in Section 6.4.3.

1.5 Why Study Risk Bounds?

The object of interest studied in this thesis is the *excess risk* – a random variable that measures the success of a statistical procedure with respect to the best performance achievable within some specified reference class. In particular, we aim to obtain *non-asymptotic* bounds on the excess risk that hold with *high probability* under *minimal assumptions* on the underlying data generating mechanism. Let us provide several reasons why studying this object in the setting outlined above is of interest.

First, in Statistics, the excess risk computed with respect to all measurable functions taken as a reference class is a very natural object. For example, letting the loss function be the squared loss, the excess risk, in this case, measures the squared L_2 distance (under the distribution of the inputs) to the conditional mean function of the output given the input, a function that minimizes the squared loss among all measurable functions. However, studying distribution-free non-asymptotic excess risk rates of convergence to zero of this object cannot be done without imposing further assumptions [78, Theorem 3.1]. At this point, a statistician is confronted with two equally valid choices. The first choice concerns imposing restrictions on the possible data generating mechanisms that ensure some regularity assumptions, such as some degree of smoothness or linearity of the conditional mean function. The second choice – adopted within this thesis and within the classical philosophy of Statistical Learning Theory [204]– is to modify the target performance benchmark from the best measurable function to the best function within some chosen reference class such as the class of sufficiently smooth functions or the class of all linear functions. This point of view allows us to refrain from imposing restrictive assumptions on the underlying data generating distribution.

Second, let us remark that abstract ideas studied within the excess risk framework, in some cases, have already led to successful practical applications. While the contributions presented within this thesis do not yield immediate practical consequences, let us discuss a few successful examples. The first example is the *support vector machines* algorithm, developed as a direct consequence of the structural risk minimization principle; see the discussions in [204] for historical remarks. The second example, originating within the Computational Learning Theory community, is the family of *boosting* algorithms, developed in an attempt to understand whether binary classification algorithms capable of performing marginally better than random guessing can be transformed into more accurate learning algorithms; we refer to the book by Schapire and Freund [185] for further details. The third and final example concerns the problem of *robust mean estimation* for heavy-tailed distributions. Recently, Catoni [44] showed that finite-sample performance for scalar mean estimation via the usual empirical mean estimator can be highly suboptimal if nothing is known about the distribution generating the samples, except that it has bounded variance. At the same time, Catoni [44] proposed a mean estimator that always performs nearly as well as the sample mean estimator for Gaussian data, a regime where the sample mean estimator is optimal. A remarkable aspect of his proposed estimator is that its creation relies on a clever argument, essentially involving only Markov's inequality;

from the point of view of the utilized mathematical machinery, Catoni's estimator, with the benefit of hindsight, could have been developed at least a century ago.

Finally, the study of excess risk in the Learning Theory setting provides natural opportunities for cross-fertilization between mathematical ideas originating from different disciplines. For example, the study of statistical guarantees for empirical risk minimizers via uniform convergence arguments has natural connections to the Empirical Processes Theory. The aim to obtain non-asymptotic excess risk calls for adaptations of the ideas developed in the literature concerning the concentration of measure phenomenon. At the same time, taking computational considerations into account, the study of statistical estimators builds bridges between Statistics, Optimization and Computational Complexity Theory. We believe that investigating such connections are of independent mathematical interest.

1.6 Authorship Note

Collaborators	**Publication Status**	**Thesis Chapter**
Varun Kanade Patrick Rebeschini	Published at *NeurIPS 2019* [210].	—
Varun Kanade Patrick Rebeschini	Working paper to be submitted for publication.	Chapter 3
Varun Kanade Patrick Rebeschini	Published at *NeurIPS 2020* [211]. A version of this work is currently under revision at *Information and Inference*.	Chapter 4
Nikita Zhivotovskiy	Under minor revision at *Bernoulli* [209].	Chapter 5
Jaouad Mourtada Nikita Zhivotovskiy	Accepted for publication at *Mathematical Statistics and Learning*. [154].	Chapter 6

Table 1.1: Summary of Completed Work.

This thesis is based on results obtained in collaboration with the co-authors listed in Table 1.1. The core of Chapters 4 to 6 are adapted from papers listed in Table 1.1, with parts of the text appearing verbatim as in the cited publications. However, the material present in the cited papers is reorganized in this thesis so that Chapters 3 to 6, containing the contributions of this thesis, follow the same structure. Some results appearing in the above-cited publication are omitted in this thesis, and some results

are reformulated. Such changes are highlighted in the bibliographic remarks sections present at the end of the respective chapters. The publication [210] cited in Table 1.1 is omitted in this thesis (it is only briefly mentioned in connection to an open problem discussed in Section 4.5) since it does not fit well within the general theme explored in this thesis. Some parts of the background material presented in Chapter 2 are also adapted from joint work with Varun Kanade and Patrick Rebeschini, to be submitted for publication (cf. Table 1.1). We now turn to Section 1.7, where we outline the structure of this thesis.

1.7 Thesis Structure and Outline of Main Results

The present chapter provided an introduction overviewing the high-level motivations for the choice of topics investigated in this thesis. Next, in Chapter 2, we present background material needed to properly understand the chapters to follow. The contributions of this thesis are contained in Chapters 3 to 6, each of which begins with pointers to relevant background material sections contained in Chapter 2. Each of the Chapters 3 to 6 follow similar structure and contain the following sections:

- **Introduction**, a section that motivates the problem and introduces the setting to be investigated;
- **Problem Formulation**, a section where the studied problem is formulated more precisely;
- **Summary of Contributions**, a section that summarizes the contributions present in the given thesis chapter;
- **Main Results**, a section containing exact theorem statements and discussions of the main results;
- **Limitations and Open Directions**, a section where we discuss potential extensions of the presented theory;
- **Bibliographic Remarks**, a section where we situate the chapter's results in a broader context and discuss related work.

Let us now briefly discuss the main results contained in Chapters 3 to 6.

The contributions presented in Chapter 3 concern developments of mathematical machinery for the analysis of improper statistical procedures. Among the most general tools for obtaining sharp excess risk bounds are the *local Rademacher averages* (see the

works by Bartlett, Bousquet, and Mendelson [19], Koltchinskii [105]). However, this machinery was primarily developed to cover the proper learning setting via the ERM principle, and in particular, it does not directly apply to improper procedures, such as optimal algorithms for the model selection aggregation problem. For some such procedures, an alternative notion of *localization via offset Rademacher averages* was proposed by Liang, Rakhlin, and Sridharan [120]. However, this notion of complexity has only been shown to yield *expected* excess risk bounds. By proving a one-sided Bernstein-type moment generating function bound for empirical processes with a negative offset term, our main result demonstrates that this complexity measure also provides *exponential tail* excess risk guarantees and in addition, these guarantees apply to some improper estimators not covered by the classical theory.

The *offset Rademacher averages* studied in Chapter 3 apply to statistical estimators satisfying a certain geometric condition. In Chapter 4, we show that this condition is satisfied by a large family of iterative regularization schemes. More precisely, we show that *mirror descent* algorithms introduced by Nemirovsky and Yudin [157] applied to *unregularized* empirical risk, along their optimization paths, visit a prediction function whose excess risk is controlled by offset Rademacher complexity of a function class depending on the algorithm's hyperparameters. An appealing feature of our analysis is that it applies simultaneously to a family of optimization algorithms that includes the typically studied gradient descent algorithm as a special case.

In Chapters 5 and 6, we study the fundamental setting of learning linear predictors with the quadratic loss.

Chapter 5 demonstrates that a reference class of linear predictors contained in a bounded Euclidean ball is not optimally learnable via empirical risk minimization in the distribution-free model of learning, where only boundedness of observations is assumed. The main result in this chapter establishes a *dimension-dependent suboptimality* of the classical constrained least squares estimator. This suboptimality is established despite the setting considered to be favorable for the ERM principle – the loss is quadratic, the reference class of functions is convex, and the covariates and response variables are almost surely bounded. Moreover, by combining known results in the literature, we discuss a more general (but weaker) separation of the statistical performance achievable via proper and improper statistical procedures in the context of learning linear predictors under boundedness assumptions only. This separation appears since there exist improper procedures satisfying performance bounds independent of the covariates' distribution, an attribute unachievable via any proper algorithm.

Motivated by the observation mentioned above regarding improper procedures, Chapter 6 studies a distribution-free learning setting, where the reference class of functions contains all linear predictors (without the boundedness assumption of Chapter 5) and *without any assumptions* on the marginal distribution of the covariates. In this setting, we formalize a minimal assumption on the conditional distribution of the response variable given the covariate vector. Under this assumption, we show the following three results. First, we prove that the (improper) truncated least squares estimator achieves the optimal expected excess risk rate. Next, we show that this estimator can fail with constant probability, thus establishing its deviation-suboptimality. Finally, we propose a procedure that achieves deviation-optimal performance. An important aspect of the setup investigated in this chapter is that our assumptions admit modelling heavy-tailed distributions of the covariates (since no assumptions are imposed on them) as well as on the response variable. The absence of assumptions on the covariates significantly departs from the literature concerning linear regression estimators robust to heavy-tailed distributions, where primarily proper estimators are considered, for which assumptions on the covariate vectors are unavoidable.

2 Background

2.1 Mathematical Framework of Statistical Learning

We denote the observed i.i.d. data sample distributed according to an *unknown distribution* P by $S_n = (X_i, Y_i)_{i=1}^n$, where $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$. We will typically (but not always) let $\mathcal{X}$ be a bounded subset of $\mathbb{R}^d$ and $\mathcal{Y}$ be a bounded subset of $\mathbb{R}$.

We denote the *loss function* by $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$. The data generating distribution P and the loss function ℓ, for any *predictor* $f : \mathcal{X} \to \mathcal{Y}$, define its *risk* $R(f)$ given by

$$R(f) = \mathbf{E}_{(X,Y)\sim P} \left[\ell(f(X), Y) \right]. \tag{2.1}$$

Minimizing the risk over all measurable functions is the ultimate goal in Statistical Learning Theory. However, as discussed in the introduction, this goal is infeasible without imposing restrictions on the unknown distribution P. Instead, we will measure the success of any prediction function $f : \mathcal{X} \to \mathcal{Y}$ by comparing its *excess risk* against some *reference class* $\mathcal{G}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. The excess risk $\mathcal{E}(f, \mathcal{G})$ is defined by

$$\mathcal{E}(f, \mathcal{G}) = R(f) - \inf_{g \in \mathcal{G}} R(g) = R(f) - R(g^\star),$$

where $g^\star$ denotes any function among $\mathcal{G}$ minimizing the population risk $R(g)$. We assume without loss of generality that such a function exists; otherwise, we could take any function in $\mathcal{G}$ whose attained risk is arbitrarily close to the infimum over $\mathcal{G}$.

The difficulty in minimizing the population risk (2.1) stems from the fact that the distribution P is unknown, and thus, the objective $R(f)$ cannot be computed exactly. Instead, a learning algorithm has access to an i.i.d. data sample $S_n = (X_i, Y_i)_{i=1}^n$ drawn from P. A *statistical estimator* $\widehat{f}$ is a procedure mapping observed data S_n to some (possibly random) function $\widehat{f}(S_n) \in \mathcal{F}$, where we call the class $\mathcal{F}$ the *range* of the estimator $\widehat{f}$. With a slight abuse of notation, we will denote by $\widehat{f}$ both the *procedure* mapping datasets to prediction functions, as well as the output prediction function $\widehat{f} = \widehat{f}(S_n)$.

The empirical counterpart to the population risk (2.1) is called the *empirical risk* and it is defined as follows:

$$R_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i).$$

We will sometimes find it convenient to switch to the Empirical Processes Theory notation. Letting P and P_n denote the true data generating measure and its empirical counterpart supported on S_n, respectively, for any function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ we let

$$Ph = \mathbf{E}_{(X,Y)\sim P}[h(X,Y)] \quad \text{and} \quad P_n h = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i).$$

To any predictor $f : \mathcal{X} \to \mathcal{Y}$ we associate the function $\ell_f : (X,Y) \mapsto \ell(f(X),Y)$. Then, $P\ell_f = R(f)$ and $P_n\ell_f = R_n(f)$. Finally, for any real-valued function f with domain $\mathcal{X}$, we denote

$$Pf = E_{X\sim P_X}[f(X)] \quad \text{and} \quad P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

where P_X denotes the marginal distribution of X.

In the following section, we discuss different types of bounds on the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$ for an estimator $\widehat{f}$ and a reference class $\mathcal{G}$.

2.1.1 Types of Excess Risk Bounds

Recall that the output predictor $\widehat{f} = \widehat{f}(S_n)$ depends on the random data sample S_n. Thus, the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$ is itself a *random variable*:

$$\mathcal{E}(\widehat{f}, \mathcal{G}) = \mathbf{E}_{(X,Y)\sim P}\left[\ell(\widehat{f}(X), Y)|S_n\right].$$

The randomness of the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$ splits the available types of bounds into those that hold on average (i.e., integrating over all data samples S_n distributed according to P), and those that hold with high probability (i.e., bounds that hold for most data samples).

Another dichotomy for the bounds available in the literature is related to their *sharpness*: some mathematical tools can only capture non-asymptotic excess risk rates that decay as $1/\sqrt{n}$ as a function of the sample size, which may be too loose in some circumstances. We provide more details below.

In-expectation vs in-deviation bounds. Fixing an estimator $\widehat{f}$, in-expectation excess risk bounds aim to find the smallest remainder term $\Delta_{\mathbf{E}}(n, \mathcal{G}, P)$, depending on some properties of the estimator $\widehat{f}$ such as its range $\mathcal{F}$, so that for some universal constant $c > 0$ the following holds:

$$\mathbf{E}_{S_n}\left[\mathcal{E}(\widehat{f}, \mathcal{G})\right] \leq c\Delta_{\mathbf{E}}(n, \mathcal{G}, P).$$

The presence of the universal constant c in the above bound is due to the fact that we are only interested in understanding non-asymptotic convergence *rates*, and particularly their dependence on the sample size n and structural properties of the reference class $\mathcal{G}$. In this thesis, we will never attempt to obtain sharp constants.

Similarly, bounds in deviation aim to find the smallest remainder term $\Delta_{\mathbf{Pr}}$ that depends on properties of the estimator $\widehat{f}$ so that the following holds for any $\delta \in (0, 1]$:

$$\mathbf{P}_{S_n}\Big(\mathcal{E}(\widehat{f}, \mathcal{G}) > c' \Delta_{\mathbf{Pr}}(n, \mathcal{G}, P, \delta)\Big) \leq \delta,$$

where $c' > 0$ is some universal constant. Observe that bounds of the above type can be transformed to in-expectation bounds via tail integration arguments; hence, obtaining sharp excess risk bounds that hold with high probability is typically a more challenging problem than obtaining in-expectation guarantees. If the remainder term $\Delta_{\mathbf{Pr}}(n, \mathcal{G}, P, \delta)$ is of order $\log(1/\delta)$ as a function of δ, we call such guarantees *exponential tail* bounds.

Obtaining sharp in-deviation guarantees is of particular interest in the improper learning settings, when the excess risk random variable can take negative values. In such regimes, statistical estimators satisfying expectation-optimal rates can be suboptimal in deviation, as first shown by Audibert [6]; see also Chapter 6 of this thesis.

Fast vs slow rates. Bounds on the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$, either in-expectation or in-deviation, decaying at most as fast as $O(1/\sqrt{n})$ are called *slow rate* bounds. Bounds on the excess risk that decay faster than $O(1/\sqrt{n})$, as a function of the sample size n, are called *fast rate* bounds. This thesis is primarily concerned with the fast rate bounds. Favourable problem structure is needed to attain fast rate bounds, and doing so requires using appropriate mathematical machinery. Let us illustrate the difference between the slow and fast rates with the following example.

Example 2.1 (Bounded Linear Regression). We define the problem setup as follows. Let the reference class of functions be all linear predictors whose Euclidean norms are at most one: $\mathcal{G} = \{\langle w, \cdot \rangle : w \in \mathbb{R}^d, \|w\|_2 \leq 1\}$. Fix the input space $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ and the output space $\mathcal{Y} = [-1, 1]$. Finally, let $\ell(y, y') = (y - y')^2$ be the quadratic loss and recall that the data generating distribution P can be any distribution supported on $\mathcal{X} \times \mathcal{Y}$.

Consider the empirical risk minimization estimator defined as any function in $\mathcal{G}$ (thus $\mathcal{F} = \mathcal{G}$) that minimizes the empirical risk:

$$\widehat{f}^{(\mathrm{ERM})} \in \arg\min_{f \in \mathcal{G}} R_n(f). \tag{2.2}$$

One way to bound the excess risk incurred by the ERM estimator (2.2) is via the classical symmetrization and contraction arguments applied to an empirical process indexed by $\mathcal{G}$:

$$
\begin{aligned}
&\mathbf{E}\mathcal{E}(\widehat{f}^{(\mathrm{ERM})}, \mathcal{G}) \\
&= \mathbf{E}\Bigg[(P - P_n)(\ell_{\widehat{f}^{(\mathrm{ERM})}} - \ell_{g^\star}) + \underbrace{P_n(\ell_{\widehat{f}^{(\mathrm{ERM})}} - \ell_{g^\star})}_{\leq 0}\Bigg] \\
&\leq \mathbf{E}\left[\sup_{f \in \mathcal{G}} (P - P_n)(\ell_f - \ell_{g^\star})\right] \\
&\leq 2\mathbf{E}\mathbf{E}_\sigma\left[\sup_{f \in \mathcal{G}} \frac{1}{n}\sum_{i=1} \sigma_i(\ell_f(X_i, Y_i) - \ell_{g^\star}(X_i, Y_i))\right] \qquad \text{(symmetrization)} \\
&= 2\mathbf{E}\mathbf{E}_\sigma\left[\sup_{f \in \mathcal{G}} \frac{1}{n}\sum_{i=1} \sigma_i \ell_f(X_i, Y_i)\right] \\
&\leq 8\mathbf{E}\mathbf{E}_\sigma\left[\sup_{f \in \mathcal{G}} \frac{1}{n}\sum_{i=1} \sigma_i f(X_i)\right], \qquad \text{(contraction)}
\end{aligned}
$$

where σ_i are i.i.d. Rademacher ($\{\pm 1\}$ valued and symmetric) random variables, and the last line in the above chain of equations is called the *(global) Rademacher complexity* [15] of the function class $\mathcal{G}$. The issue with the above classical proof technique is that the global Rademacher complexity can only decay at the rate $O(1/\sqrt{n})$. However, for the problem setting outlined in Example 2.1, fast excess risk rates of order $O(d/n)$ are achievable; this is possible by refining the above argument to compute the Rademacher complexity of an appropriately chosen subset of the class $\mathcal{G}$, taking into account that the ERM estimator gets better with the increasing sample size n (cf. Section 2.5). We will explore the above example problem setting in far greater detail in Chapter 5 of this thesis.

Let us conclude this chapter by making one final remark. Concerning in-deviation excess risk guarantees, obtaining fast rate bounds becomes an even bigger challenge. There are various proof techniques for obtaining sharp excess risk guarantees that hold only in expectation (e.g., via average stability arguments or online-to-batch conversions). Converting such guarantees to the ones that hold with high probability is a rather delicate problem because various concentration tools typically come with slow rate variance terms. Some sort of an additional mechanism needs to be at play in order to compensate for such variance terms; see Section 2.5 and Chapter 3 for extended discussions.

2.2 Improper Learning

Given a reference class of functions $\mathcal{G}$, a statistical estimator $\widehat{f}$ is called *improper* if its range $\mathcal{F}$ contains functions outside of the reference class $\mathcal{G}$. A principal example where improperness is needed to achieve statistical optimality is the problem of *model selection aggregation* [155] defined below.

Example 2.2 (Model Selection Aggregation). Let $\mathcal{X}$ be the input space, $\mathcal{Y} = [-1, 1]$, be the output space, and let $\ell(y, y') = (y - y')^2$ be the quadratic loss. In the model selection aggregation problem the reference class of functions is chosen to be a finite set $\mathcal{G} = \{g_1, \ldots, g_m\}$ of functions uniformly bounded by 1 over the input space $\mathcal{X}$.

The optimal excess risk rate of aggregation is the fast rate $\log(m)/n$ [198]. Any proper estimator can only achieve the slow rate $\sqrt{\log(m)/n}$. This limitation can readily be seen for reference classes of functions of size two. Indeed, let $\mathcal{X} = \{-1, 1\}^2$. For any $x \in \mathcal{X}$, define $g_1(x) = x_1$ and $g_2(x) = x_2$. Consider the data generating distribution P where the covariate vector X is distributed uniformly over $\mathcal{X}$ and where Y given $X = (x_1, x_2)$ is equal to x_1 with probability $1/2 + \Delta$ and x_2 otherwise, for some positive $\Delta > 0$. Setting, $\Delta = c/\sqrt{n}$ for some small enough universal constant c, any proper learning procedure that attains the fast excess risk rate $\log(m)/n$ can be transformed into a procedure that, using only n observations, can distinguish between two biased coins with bias of order $1/\sqrt{n}$. Choosing a small enough constant $c > 0$, this can be shown to be information theoretically impossible (cf. [198, 199]), thus establish suboptimality of any proper procedure. The suboptimality of proper procedures in the context of model selection aggregation have been known for a long time, and date back at least to the late 90s (see the work of Catoni [42]).

On the other hand, an improper procedure attaining the fast excess risk rate of order $\log(m)/n$, in its solution, does not necessarily encode structure that could be used to obtain a procedure for distinguishing between biased coins. Indeed, improper procedures can bypass the limitations of algorithms constrained to selecting a function from a given reference class $\mathcal{G}$. For obtaining optimal in-expectation performance in the statistical learning setting considered in this thesis, a progressive mixture approach dates back to the early work of Barron [14], with further explorations in the works of Catoni [42], Yang [222], Juditsky, Rigollet, and Tsybakov [94], among others. Progressive mixture rules can also be seen as an algorithm resulting from a standard online-to-batch conversion applied to the *exponential weights* aggregating algorithm in the context of sequential prediction of individual sequences. See the works by Vovk

[214, 215], Haussler, Kivinen, and Warmuth [83], Audibert [7] and the textbook by Cesa-Bianchi and Lugosi [47] for further details.

A noteworthy aspect of the expectation-optimal progressive mixture procedure for the model selection aggregation problem defined in Example 2.2 is that none of the proofs available in the literature follow traditional strategies based on empirical processes theory, such as those based on local Rademacher complexities (see Section 2.5). As it turns out, such proofs are impossible since they would imply corresponding exponential-tail deviation bounds, shown to not hold by Audibert [6], who also proposed a deviation-optimal method for model selection aggregation, called *star algorithm*. One of the key takeaways from Audibert's analysis is that the excess risk random variable $\mathcal{E}(\widehat{f}, \mathcal{G})$ can take *negative values* for improper estimators $\widehat{f}$. It follows that in-expectation guarantees for improper methods do not imply *any* bounds that hold with high probability since Markov's inequality does not apply. In Chapter 6 of this thesis, we will demonstrate the existence of two improper algorithms for distribution-free linear regression problem, both of the algorithms attaining optimal in-expectation guarantees, yet both algorithms failing with a constant probability.

Beyond the model selection aggregation problem (Example 2.2), improperness is also known to be necessary to achieve optimal statistical performance for logistic regression. When the reference class is taken to be all linear predictors with bounded Euclidean norm, proper algorithms attaining fast rate excess risk bounds necessarily incur exponential dependence on the diameter of the reference class as shown by Hazan, Koren, and Levy [85], answering an open problem posed by McMahan and Streeter [138]. Later, Foster, Kale, Luo, Mohri, and Sridharan [67] showed that improper procedures can attain the fast rate with polynomial dependence on the diameter, thus demonstrating a separation between statistical performance achievable via proper and improper estimators. Their algorithm builds on the ideas of Vovk's aggregating algorithm and the notion of mixable losses [214, 215]. See also the works by Mourtada and Gaïffas [153], Jézéquel, Gaillard, and Rudi [92], where computationally more efficient alternatives to the algorithm of [67] are investigated.

In Chapters 5 and 6, we will explore the necessity of improperness in the context of linear regression.

2.3 Convex Analysis

The notions of convex analysis presented in this section are primarily needed for a formal treatment of mirror descent optimization schemes described in Section 2.4.

This section is based on the textbooks by Cesa-Bianchi and Lugosi [47, Section 11] and Rockafellar [177, Section 26], where the interested reader may find further information.

Let $\mathcal{D}$ be a convex subset of $\mathbb{R}^d$. A function $f : \mathcal{D} \to \mathbb{R}$ is called convex if for any $x, y \in \mathcal{D}$ and any $\lambda \in [0, 1]$ it holds that

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

We now define the Legendre-Fenchel transform, also called the convex conjugate, a fundamental operation in convex analysis. It induces a one-to-one correspondence between lower semi-continuous convex functions.

Definition 2.1 (Legendre-Fenchel Transform)**.** Let $\mathcal{D} \subseteq \mathbb{R}^d$ and let f be a real-valued function with domain $\mathcal{D}$. The Legendre-Fenchel transform of f, denoted by f^*, is a real-valued function defined by

$$f^*(x^*) = \sup_{x \in \mathcal{D}} \{\langle x, x^\star \rangle - f(x)\}.$$

We define the domain of f^* by $\mathcal{D}^* = \{x^* \in \mathbb{R}^d : \sup_{x \in \mathcal{D}} \{\langle x, x^\star \rangle - f(x)\} < \infty\}$.

To each convex and differentiable function f, we may associate a distance function called *Bregman divergence*.

Definition 2.2 (Bregman Divergence)**.** Let $f : \mathcal{D} \to \mathbb{R}$ be a convex real-valued function, whose domain $\mathcal{D}$ has a non-empty interior $\text{int}(\mathcal{D})$. Assume that f is differentiable on $\text{int}(\mathcal{D})$. The *Bregman divergence* induced by f is a function $D_f : \mathcal{D} \times \text{int}(\mathcal{D}) \to \mathbb{R}$ defined by

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

In general, Bregman divergence is not a metric since it does not satisfy the triangle inequality, nor is it symmetric. However, observe that Bregman divergence induced by a convex function f is always non-negative.

If a differentiable convex function f satisfies some regularity conditions, its gradient mapping can be inverted by computing f^*, the Legendre-Fenchel transform of f. Such considerations are of interest to us since the mirror descent algorithm (see Section 2.4) uses gradients of convex functions to perform mappings between primal and dual spaces. In the context of mirror descent algorithms, we will restrict our attention to the mappings between primal and dual spaces induced by gradients of *mirror maps* – convex functions satisfying the properties defined below. We remark that mirror maps are sometimes called functions of Legendre type (e.g., in [177, 47]); however, we adopt the terminology that is sometimes used in the convex optimization literature (e.g., [38, Section 4.1]).

Definition 2.3 (Mirror Map)**.** Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a convex set with non-empty interior denoted by $\text{int}(\mathcal{D})$. A convex function $f : \mathcal{D} \to \mathbb{R}$ is called a *mirror map* provided that the following three conditions holds:

1. f is lower semi-continuous;
2. f is strictly convex and continuously differentiable on $\text{int}(\mathcal{D})$;
3. for any sequence $(x_k)_{k\geq 1}$ of points in $\mathcal{D}$ converging to a boundary point of $\mathcal{D}$, it holds that $\lim_{k\to\infty} \|\nabla f(x_k)\|_2 \to \infty$.

In Section 2.4.4, we provide some examples of mirror maps in connection with the mirror descent procedure. Let us now summarize the consequences of the above definitions.

Lemma 2.1. *Let $\mathcal{D} \subseteq \mathbb{R}^d$ and suppose that $f : \mathcal{D} \to \mathbb{R}$ is a mirror map. Let $f^* : \mathcal{D}^* \to \mathbb{R}$ be its Lagrange-Fenchel transform. Then, the following statements are true:*

1. *f^* is a mirror map;*
2. *the gradient mapping $(\nabla f) : \text{int}(\mathcal{D}) \to \mathbb{R}^d$ is a bijection whose range equals $\text{int}(\mathcal{D}^*)$. The inverse mapping $(\nabla f)^{-1} : \text{int}(\mathcal{D}^*) \to \text{int}(\mathcal{D})$ is given by $(\nabla f)^{-1} = \nabla f^*$;*
3. *for any $x, y \in \text{int}(\mathcal{D})$ we have $D_f(x, y) = D_{f^*}(\nabla f(y), \nabla f(x))$.*

The first two properties stated in the above lemma are proved in [177, Theorem 26.5]. The third property is proved in [47, Proposition 11.1]. We also remark that in the definition of the mirror map, we require that f is lower semi-continuous. This condition ensures that $(f^*)^* = f$ (see [177, Theorem 12.2]).

2.4 Convex Optimization

The purpose of this section is to provide background on convex optimization, particularly the mirror descent method of Nemirovsky and Yudin [157]. The material presented in this section is a prerequisite for the study of iterative regularization procedures discussed in Chapter 4. This section depends on the convex analysis background discussed in Section 2.3.

The general problem, for which an approximate solution is sought, is the minimization problem of the following form:

$$\min_{x \in \mathbb{R}^d} f(x),$$

where the function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable. We assume that the minimum value exists, and is attained by a point $x^\star \in \mathbb{R}^d$. It is possible to relax differentiability assumptions to sub-differentiability and also, to consider optimization problems over constrained sets. However, for the purposes of this thesis, we shall only consider unconstrained and differentiable problems.

The material presented in this section is classical. The interested reader will find extensive bibliographic remarks and further information on convex optimization in the textbooks [157, 35, 158]. All convergence proofs presented in this section are based on *potential function arguments*; an extensive survey for convergence proofs of mirror descent methods fitting within such a framework can be found in the recent survey by Bansal and Gupta [13]. The rest of this section is structured as follows.

1. In Section 2.4.1, we demonstrate a potential-based (or Lyapunov function-based) convergence analysis of the continuous-time gradient descent method.

2. Following Nemirovsky and Yudin [157, Section 3], Section 2.4.2 contains a discussion behind the ideas that led to the development of the mirror descent method. This section is restricted to the continuous-time setting.

3. Section 2.4.3 discusses the ideas needed for the discrete-time analysis.

4. Section 2.4.4 provides some example mirror maps, yielding different instantiations of the mirror descent algorithm.

2.4.1 Continuous-Time Gradient Descent

The continuous-time gradient descent method, initialized at a point $x_0 \in \mathbb{R}^d$, is characterized by the following equation:

$$\frac{d}{dt} x_t = -\nabla f(x_t),$$

where the time parameter t is non-negative. Let $x^\star \in \mathbb{R}^d$ denote a point minimizing the objective function f. We will show that the averaged iterate $\bar{x}_t = \int_0^t x_s ds$ approaches the optimal value $f(x^\star)$ at the rate $O(1/t)$.

Since the objective function f is convex and differentiable, the following inequality holds for any $x \in \mathbb{R}^d$:

$$\langle -\nabla f(x), x^\star - x \rangle \geq f(x) - f(x^\star) \geq 0,$$

where the last inequality follows since we have taken $x^\star$ to be a minimizer of the objective function f. The above inequality implies that along the gradient flow $(x_t)_{t\geq 0}$ the Euclidean distance $\|x^\star - x_t\|_2$ is *non-increasing*. In particular, it holds that

$$-\frac{d}{dt}\frac{1}{2}\|x^\star - x_t\|_2^2 = \left\langle \frac{d}{dt}x_t, x^\star - x_t \right\rangle = \langle -\nabla f(x_t), x^\star - x_t \rangle \geq f(x_t) - f(x^\star) \geq 0.$$

The fact that the averaged iterate $\bar{x}_t$, or that best iterate (in the sense of its objective value $f(x_t)$) along the path $(x_t)_{t\geq 0}$ approaches the optimal value $f(x^\star)$ is almost immediate from the above identity. The squared Euclidean distance in the above equation is called the *potential function*. Its role in convergence analysis can be described as follows. The potential function is non-negative and non-increasing along the trajectory of continuous-time gradient-descent. In fact, it is decreasing at the rate that depends on the sub-optimality gap $f(x_t) - f(x^\star)$. It follows that for any $\varepsilon > 0$, the sub-optimality gap $f(x_t) - f(x^\star)$ cannot stay larger than ε, for otherwise, the non-negative potential function would decrease below zero, yielding a contradiction.

More formally, for any $T > 0$, integrating both-sides of the equation displayed above yields

$$\frac{1}{2}\|x^\star - x_0\|_2^2 - \frac{1}{2}\|x^\star - x_T\|_2^2 = \int_0^T -\frac{d}{dt}\frac{1}{2}\|x^\star - x_t\|_2^2 dt \geq \int_0^T (f(x_t) - f(x^\star))dt.$$

Recalling that $\bar{x}_T = \int_0^T x_t dt$, it follows from the convexity of the function f that

$$f(\bar{x}_T) - f(x^\star) \leq \frac{1}{T}\left(\int_0^T (f(x_t) - f(x^\star))dt\right) \leq \frac{\|x^\star - x_0\|_2^2}{2T}.$$

Thus, the averaged point $\bar{x}_t$ approaches the optimal value $f(x^\star)$ at rate $O(1/t)$. Convergence rates for discrete-time gradient descent updates can be obtained by following along the lines of the above potential-based proof technique. The exact convergence rate depends on the imposed properties on f, such as the degree of smoothness of strong convexity.

One drawback of the gradient descent method is its intrinsic dependence on the Euclidean distances, as can be seen from the analysis sketched above. For example, in discrete-time non-smooth optimization by gradient descent, convergence speed depends on the Lipschitz constant, defined as the maximum possible Euclidean norm

of the gradients $\nabla f(x_t)$. If such gradients live, for example, in a d-dimensional hypercube $[-1,1]^d$, their worst-case Euclidean norms scale as $\sqrt{d}$, which may result in a prohibitively slow convergence speed in high-dimensional applications. In contrast, the ℓ_∞ norms of $\nabla f(x_t)$ admit a dimension-independent bound equal to 1 in such setups. It is thus desirable to develop flexible numeric optimization schemes applicable to non-Euclidean problem geometries. This is achieved by the method of mirror descent, discussed in the next section.

2.4.2 Continuous-Time Mirror Descent

We will now explain how the convergence proof for continuous-time gradient descent presented in the previous section can be used to design a new method called *mirror descent*. The following presentation is based on [157, Chapter 3]. The principal idea therein is to replace the potential function – the squared Euclidean norm – used in the convergence analysis of gradient descent by an arbitrary function, satisfying certain properties needed to ensure convergence. The dynamics of the algorithm can then be deduced from the chosen potential function. It is noteworthy that this approach, undertaken by Nemirovsky and Yudin [157, Chapter 3], is of an opposite nature compared to the more traditional analysis, where a form of iterative procedure is known in advance, and a potential function is sought to establish its convergence.

Let us first describe how one may attempt to generalize the convergence argument presented in Section 2.4.1. We have shown that convergence of gradient descent follows from the potential function given by $V(x_t) = \frac{1}{2}\|x^\star - x_t\|_2^2$. Repeating the convergence argument presented in Section 2.4.1, we may not insist on the specific form of the potential function $V(x_t)$, but rather insist that the following identity is true:

$$-\frac{d}{dt}V(x_t) = \left\langle \frac{d}{dt}x_t, -\nabla V(x_t) \right\rangle = \langle -\nabla f(x_t), x^\star - x_t \rangle \,. \tag{2.3}$$

The first equality above follows from directly differentiating the potential function $V(\cdot)$. The second equality denotes what *we want* to hold; that is, we want to choose the dynamics $(x_t)_{t\geq 0}$ such that the second equality above is true. However, the only obvious way to make the second equality above hold is by taking $\frac{d}{dt}x_t = -\nabla f(x_t)$ and $V(x_t) = \frac{1}{2}\|x_t\|_2^2 - \langle x_t, x^\star \rangle$. This leads precisely to the gradient descent setup considered in the previous section, and thus, it appears that we have achieved nothing with the above derivations.

Let us now try to fix the above reasoning, which led us back to the gradient descent algorithm. Let $\psi^* : \mathbb{R}^d \to \mathbb{R}$ be some function and let $V(x_t) = \psi^\star(x_t) - \langle x_t, x^\star \rangle$. In an

attempt to make the equality $-\nabla V(x_t) = x^\star - x_t$ hold, let us observe that we, in fact, require that

$$x^\star - \nabla\psi^\star(x_t) = x^\star - x_t.$$

The above identity, however, lacks meaning, in a precise mathematical sense: its left hand side is adding $x^\star$, an element of the original vector space, with $\nabla\psi(x_t)$, an element of the dual space. This operation is only meaningful in the special case where the primal and dual spaces are isomorphic to one another. To make the above discussion more precise, let us now assume that the objective function f is defined on a subset of a d-dimensional real vector space E; let E^* denote the dual space, that is, the space of linear functionals on E. Thus, the domain of the function $\psi*$ should be E^*, in order that for x in the domain of ψ we have $\nabla\psi(x) \in E$. Let $(u_t)_{t\geq 0}$ denote a sequence of elements in E^* and rewrite the identity (2.3) via the following alternative expression:

$$\frac{d}{dt}V(u_t) = \left\langle \frac{d}{dt}u_t, -\nabla V(u_t) \right\rangle = \langle -\nabla f(x_t), x^\star - x_t \rangle\,, \tag{2.4}$$

where recall that $V(u_t) = \psi^*(u_t) - \langle u_t, x^\star\rangle$. The above identity is hence satisfied with the *mirror descent* dynamics given by:

$$\begin{aligned} \frac{d}{dt}u_t &= -\nabla f(x_t) \\ x_t &= \nabla\psi^*(u_t), \end{aligned} \tag{2.5}$$

where $t \geq 0$ and $x_0 \in E$ may be initialized arbitrarily. The main action is taken in the evolution of the dual sequence $(u_t)_{t\geq 0}$, while the sequence $(x_t)_{t\geq 0}$ may be considered as its shadow, thus explaining the name of the method. This is key insight behind the ideas presented by Nemirovsky and Yudin [157, Chapter 3].

Now observe that $x_t = \nabla\psi^*(u_t)$ gives a mapping $u_t \mapsto x_t$ from the dual space to the primal space. Assume now that ψ^* is a *mirror map* (in the sense of Definition 2.3). Then, let ψ be the Legendre-Fenchel transform of ψ^* (see Definition 2.1). From the properties of mirror maps summarized in Lemma 2.1, it follows that $\nabla\psi$ is the inverse mapping of $\nabla\psi^*$; in particular, $u_t = \nabla\psi(x_t)$. We may thus rewrite the mirror descent dynamics (2.5) by

$$\frac{d}{dt}\nabla\psi(x_t) = -\nabla f(x_t). \tag{2.6}$$

Recall that the success of the continuous-time gradient descent convergence analysis discussed in Section 2.4.1 was based on the non-negativity of the non-increasing

potential function. We will now normalize the potential function, without changing its time derivative, so that it is always non-negative:

$$\begin{aligned}\frac{d}{dt}\left(V(u_t)\right) &= \frac{d}{dt}\left(\psi^*(u_t) - \langle u_t, x^\star\rangle\right)\\ &= \frac{d}{dt}\left(\psi^*(\nabla\psi(x_t)) - \psi^*(\nabla\psi(x^\star))\left\langle\nabla\psi(x_t) - \nabla\psi(x^\star), \nabla\psi^*(\nabla\psi(x^\star))\right\rangle\right)\\ &= \frac{d}{dt}D_{\psi^*}(\nabla\psi(x_t), \nabla\psi(x^\star))\\ &= \frac{d}{dt}D_{\psi}(x^\star, x_t),\end{aligned}$$

where the last line follows via Lemma 2.1.

Let us summarize what we have achieved. First, for any mirror map ψ, we have associated the dynamics (2.6). These dynamics, via the above derivations and (2.4) satisfy the following identity:

$$-\frac{d}{dt}D_{\psi}(x^\star, x_t) = \langle -\nabla f(x_t), x^\star - x_t\rangle \geq f(x_t) - f(x^\star). \tag{2.7}$$

An identical argument to the one considered for continuous-time gradient descent (cf. Section 2.4.1) yields, for any $T > 0$:

$$f(\bar{x}_T) - f(x^\star) \leq \frac{D_\psi(x^\star, x_0)}{T}.$$

Thus, the intrinsic dependence on Euclidean distances present for the gradient flow was successfully replaced by another notion of distance, namely the Bregman divergence induced via the mirror map ψ.

In the next section, we discuss the discretization of the dynamics defined in (2.6).

2.4.3 Discrete-Time Mirror Descent

In the previous section, we have described some ideas behind the mirror descent procedure. We will now discretize the dynamics (2.6) and discuss how the convergence proof for the continuous-time updates can be adapted to the discrete-time analysis.

Let $\psi : \mathcal{D} \to \mathbb{R}$ be a mirror map. For $t = 0, 1, 2, \ldots$, let $(\eta_t)_{t\geq 0}$ be a sequence of positive real numbers called *step sizes*. Choosing $x_0 \in \text{int}(\mathcal{D})$ arbitrarily, we may discretize the dynamics (2.6) as follows:

$$\nabla\psi(x_{t+1}) = \nabla\psi(x_t) - \eta_t\nabla f(x_t), \tag{2.8}$$

where $t = 0, 1, 2, \ldots$. An alternative way to write the above updates is

$$x_{t+1} = \nabla\psi^*\left(\nabla\psi(x_t) - \eta_t\nabla f(x_t)\right).$$

Yet another formulation of mirror descent updates, discussed by Beck and Teboulle [21], is via the following proximal characterization:

$$x_{t+1} = \operatorname{argmin}_{x\in\mathbb{R}^d} \left\{ \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t\rangle}_{\text{linear approximation of } f(x)} + \underbrace{\frac{1}{\eta_t} D_\psi(x, x_t)}_{\text{penalty}} \right\}.$$

In the continuous-time convergence analysis (cf. Section 2.4.2), the key identity is given by the time derivative of the potential $D_\psi(x^\star, x_t)$ (cf. Equation (2.7)). In discrete-time analysis, the same proof strategy can be repeated. The discrete-time alternative to the time derivative of the potential is the difference $D_\psi(x^\star, x_t) - D_\psi(x^\star, x_{t+1})$. Indeed, the following well-known lemma will be useful.

Lemma 2.2. *Let $\psi : \mathcal{D} \to \mathbb{R}$ be a convex function, differentiable on $\operatorname{int}(\mathcal{D})$. Then, for any $x, y \in \operatorname{int}(\mathcal{D})$ and $z \in \mathcal{D}$ we have*

$$D_\psi(z, x) - D_\psi(z, y) = \langle \nabla\psi(y) - \nabla\psi(x), z - x\rangle - D_\psi(x, y).$$

Proof. The lemma can be proved directly by the definition of Bregman divergence. We have

$$\begin{aligned}
&D_\psi(z, x) - D_\psi(z, y) \\
&= (\psi(z) - \psi(x) - \langle \nabla\psi(x), z - x\rangle) - (\psi(z) - \psi(y) - \langle \nabla\psi(y), z - y\rangle) \\
&= \langle \nabla\psi(y) - \nabla\psi(x), z - x\rangle - (\psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle) \\
&= \langle \nabla\psi(y) - \nabla\psi(x), z - x\rangle - D_\psi(x, y).
\end{aligned}$$

□

Applying the above lemma with $x = x_t$, $y = x_{t+1}$ and $z = x^\star$, we obtain a discrete-time counterpart to the change of potential formula (2.7):

$$D_\psi(x^\star, x_t) - D_\psi(x^\star, x_{t+1}) = \langle -\eta_t \nabla f(x_t), x^\star - x_t\rangle - \underbrace{D_\psi(x_t, x_{t+1})}_{\text{discretization error}},$$

where we have also used the definition of discrete-time mirror decent updates (2.8). The only difference between the above identity and the continuous-time change of potential formula (2.7) is the presence of the discretization error term in the above equation. Once the above discrete-time change of potential identity is summed for $t = 0, 1, \ldots, T$, the convergence of $f(x_t)$ (or $f(\bar{x}_T)$, where $\bar{x}_T = \frac{1}{T+1}\sum_{t=0}^{T} x_t$) to the minimal value $f(x^\star)$ can be established similarly to the continuous-time case, provided

that the cumulative effect of the discretization error terms can be controlled. This control ultimately depends on the properties of the objective function f. See, e.g., [38, 13], for examples on how the discretization error terms can be bounded under different regularity assumptions on f.

2.4.4 Example Mirror Maps

In this section we provide some example mirror maps.

Squared Euclidean Norm. Let $\mathcal{D} = \mathbb{R}^d$ and define

$$\psi(x) = \frac{1}{2}\|x\|_2^2.$$

Then $\mathcal{D}^* = \mathbb{R}^d$ and $\psi^* = \psi$. The mirror descent updates (2.8) become the classical gradient descent updates

$$x_{t+1} = x_t - \eta_t \nabla f(x_t).$$

Negative Entropy. Let $\mathcal{D} = \{x \in \mathbb{R}^d : x_i \geq 0\}$ be the non-negative d-dimensional orthant. Let

$$\psi(x) = \sum_{i=1}^{d} (x_i \log(x_i) - x_i),$$

where we use the convention $0 \log 0 = 0$. Then, the Legendre-Fenchel transform of ψ is equal to $\psi^*(x) = e^x$, with the domain $\mathcal{D}^* = \mathbb{R}^d$. The mirror descent updates (2.8) become

$$x_{t+1} = x_t \odot \exp\left(-\eta_t \nabla f(x_t)\right), \tag{2.9}$$

where $\odot$ denotes componentwise multiplication, and exp is to be understood as componentwise exponentiation. The above updates can be recognized as the exponential gradient algorithm of Kivinen and Warmuth [98].

Hyperbolic Entropy. One limitation of the exponential gradient algorithm (2.9) is that the iterates $(x_t)_{t \geq 0}$ can only take positive values. This limitation is addressed via the hyperbolic entropy mirror map $\psi = \phi_\gamma$ (parametrized via a positive real number γ) defined by

$$\psi(x) = \phi_\gamma(x) = \sum_{i=1}^{d} \left(x_i \text{arcsinh}(x_i/\gamma) - \sqrt{x_i^2 + \gamma^2} \right),$$

where $\text{arcsinh}(x) = \log(\sqrt{x^2+1} + x)$. The domain of ϕ_γ is equal to $\mathcal{D} = \mathbb{R}^d$. For further information on the hyperbolic entropy mirror map we refer to [47, Example 11.5] and [69].

2.5 Local Rademacher Complexity Excess Risk Bounds

This section provides background on local Rademacher complexities, a complexity measure that yields sharp excess risk bounds in the bounded learning setting. This section forms an important background for the material presented in Chapter 3. The key concepts to be understood are the following:

1. how the classical theory of localization, presented in Section 2.5.1, relies on the so-called Bernstein condition, used to extinguish "slow rate" variance terms arising from an application of Talagrand's concentration inequality;

2. why Bernstein condition does not easily fit within improper learning setups;

3. how the more recently introduced approach to localization via offset processes provides an alternative avenue for obtaining sharp excess risk bounds without relying on the Bernstein condition. Instead, the offset Rademacher complexity approach exploits estimator-specific properties.

The background material presented in this section is structured as follows.

First, in Section 2.5.1, we introduce the classical approach to localization, developed in a series of works by Koltchinskii and Panchenko [107], Koltchinskii [104], Bartlett, Boucheron, and Lugosi [18], Lugosi and Wegkamp [130], Bartlett, Bousquet, and Mendelson [19], Koltchinskii [105], among others. Local Rademacher averages improve upon the global Rademacher averages – a complexity measure arising from direct symmetrization of the excess loss empirical process – by allowing us to compute the Rademacher complexity over a localized subset of the whole function class. This refinement of the global Rademacher complexity approach is necessary to capture convergence rates of the excess risk that decay faster than $1/\sqrt{n}$, where n is the sample size (see [16, Theorem 2.3]).

Next, in Section 2.5.2, we discuss a more recent approach of localization via offset Rademacher complexities, introduced in the statistical context with the quadratic loss by Liang, Rakhlin, and Sridharan [120], originating from prior work on online learning by Rakhlin and Sridharan [170].

2.5.1 The Classical Approach via Fixed Points

In this section, let us fix a loss function $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ such that for any $y \in [-b, b]$, the function $\ell(\cdot, y)$ is C_b-Lipschitz. Let $\mathcal{F}$ be the range of some estimator $\widehat{f}$ (i.e., for any observed dataset, the estimator $\widehat{f}$ selects a function inside

$\mathcal{F}$). Further, let $\mathcal{G}$ be some reference class of functions. We assume that the functions in $\mathcal{F}$ and $\mathcal{G}$ have domain $\mathcal{X}$ and range $[-b, b]$. Finally, let $g^\star$ denote any population risk minimizer over the class $\mathcal{G}$, i.e., $g^\star \in \operatorname{argmin}_{f\in\mathcal{G}} R(f)$. The aim of this section is to demonstrate how the local Rademacher complexity approach can upper bound the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G}) = R(\widehat{f}) - R(g^\star)$, and in particular, we aim to understand the imposed assumptions used to achieve the resulting bounds.

First, observe the classical decomposition

$$\begin{aligned}\mathcal{E}(\widehat{f}, \mathcal{G}) &= (R(\widehat{f}) - R(g^\star)) - (R_n(\widehat{f}) - R_n(g^\star)) + (R_n(\widehat{f}) - R_n(g^\star)) \\ &\leq \sup_{f\in\mathcal{F}} \{(R(f) - R(g^\star)) - (R_n(f) - R_n(g^\star))\} + (R_n(\widehat{f}) - R_n(g^\star))\end{aligned}$$

The term $R_n(\widehat{f}) - R_n(g^\star)$ is typically controlled by assuming that it is at most zero almost surely. This is true, for example, if $\widehat{f}$ is an empirical risk minimizer over $\mathcal{F}$ and $\mathcal{G} \subseteq \mathcal{F}$. Henceforth, it remains to control the supremum term.

In the approach of local Rademacher complexity bounds, the supremum term is controlled via Talagrand's concentration inequality for empirical processes [191], a functional Bernstein-type concentration inequality with variance proxy

$$\sigma^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left\{ \operatorname{Var}_{(X,Y)\sim P} \left[\ell_f(X, Y) - \ell_{g^\star}(X, Y) \right] \right\}.$$

In particular, denoting $Z = \sup_{f\in\mathcal{F}} \{(R(f) - R(g^\star)) - (R_n(f) - R_n(g^\star))\}$ and letting $c > 0$ be some large enough universal constant, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ the following deviation-type inequality holds:

$$Z \leq 2\mathbf{E}Z + c\sqrt{\frac{\sigma^2(\mathcal{F}) \log(1/\delta)}{n}} + c\frac{bC_b \log(1/\delta)}{n}, \tag{2.10}$$

Above, we have stated a version of Talagrand's inequality with absolute constants, as they are not of the primary importance in our setting. For sharper refinements see, e.g., [32, 99].

Let us now discuss the high-level ideas behind localization theory via Rademacher complexities. First, observe that the C_b-Lipschitzness assumption on the loss function ℓ allows us to control the variance of $\ell_f - \ell_{g^\star}$ as follows:

$$\operatorname{Var}(\ell_f - \ell_{g^\star}) \leq \mathbf{E}_{(X,Y)\sim P}[(\ell_f(X, Y) - \ell_{g^\star}(X, Y))^2] \leq C_b^2 \|f - g^\star\|^2_{L_2(P_X)}.$$

Hence, replacing $\mathcal{F}$ in Talagrand's inequality (2.10) by a localized subset $\mathcal{F}(r) = \{f \in \mathcal{F} : \|f - g^\star\|_{L_2(P_X)} \leq r\}$ for some radius $r > 0$, allows for an explicit control of the variance proxy term: $\sigma^2(\mathcal{F}(r)) \leq C_b^2 r^2$. Moreover, using a peeling argument, it is

possible to obtain a uniform Bernstein-type concentration bound on the excess risk over the full class $\mathcal{F}$, such that for each $f \in \mathcal{F}$, the variance-proxy is proportional to a "slow rate" term $\sqrt{\|f - g^\star\|^2_{L_2(P_X)}/n}$. See [216, Theorem 14.20, Equation 14.51] for more details and a precise quantification of the above statement.

We are now ready to discuss the final element used to obtain sharp excess risk bounds. Applying the obtained uniform Bernstein-type concentration bound (see above) to the estimator $\widehat{f}$ of interest, we obtain an upper bound on its excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$ in terms of three terms. The first term is the localized Rademacher complexity (see the definition below), which arises from an application of standard symmetrization and contraction arguments to the expected supremum of the excess loss empirical process computed over the localized set $\mathcal{F}(r)$, for some properly chosen radius $r > 0$. The second term is a term of order $\log(1/\delta)/n$, a confidence term appearing in Talagrand's inequality (2.10). The third term is a "slow rate" variance term of order $\sqrt{\|f - g^\star\|^2_{L_2(P_X)}/n}$. This term is problematic, as local Rademacher complexities aim to obtain the so-called "fast rate" bounds on the excess risk that decay faster than $1/\sqrt{n}$. In order to achieve this goal, a further assumption called the *Bernstein condition* is introduced that we state below. This condition ensures that the "slow rate" variance term can be compensated by the excess risk term itself. More specifically, the slow rate term disappears by subtracting half of the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$ from both sides of the resulting excess risk bound, applying the Bernstein condition on the right hand size, and optimizing the quadratic equation in the variable $\|\widehat{f} - g^\star\|_{L_2(P_X)}$.

Definition 2.4 (Bernstein Condition)**.** Let P be a distribution supported on $\mathcal{X} \times \mathcal{Y}$ and let ℓ be a loss function with domain $\mathcal{Y} \times \mathcal{Y}$. The tuple $(P, \ell, \mathcal{F}, g^\star)$ satisfies the Bernstein condition with parameter $\gamma > 0$ if the following holds for any $f \in \mathcal{F}$:

$$\mathbf{E}_{X \sim P_X}(f(X) - g^\star(X))^2 \leq \frac{1}{\gamma}\mathbf{E}_{(X,Y)\sim P}\left[\ell_f(X,Y) - \ell_{g^\star}(X,Y)\right].$$

An immediate observation regarding improperness is in order. The above condition essentially rules out improper learning settings, should one wish to obtain bounds that hold for any distribution P; recall that such distribution-free bounds are the central topic in this thesis (see the discussions in Chapter 1). Indeed, if $\mathcal{F}$ was a strict superset of the reference class $\mathcal{G}$, it would be possible to arrange for some distribution P, such that some element in $\mathcal{F} \backslash \mathcal{G}$ was a population risk minimizer over $\mathcal{F}$. If this was the case, the right-hand side in the equation defining the Bernstein condition above could be made negative, while the left-hand side is always non-negative, yielding a contradiction. It is the topic of Chapter 3 to obtain localized excess risk bounds

without relying on the Bernstein condition. A typical application domain where the Bernstein condition holds in a distribution-free sense (as opposed to, e.g., imposing low-noise assumptions on the underlying data-generating mechanism, as frequently done in the classification settings) is the proper learning setup with a convex class $\mathcal{F}$ and a strongly-convex loss function ℓ.

The Bernstein condition, together with Lipschitzness of the loss function, yields the following relation between the variance and expectation of the elements of the excess loss class $\{\ell_f - \ell_{g^\star} : f \in \mathcal{F}\}$:

$$\operatorname{Var}(\ell_f - \ell_{g^\star}) \leq C_b^2 \|f - g^\star\|_{L_2(P_X)}^2 \leq \frac{C_b^2}{\gamma} \mathbf{E}[\ell_f - \ell_{g^\star}] \quad \text{for any } f \in \mathcal{F}.$$

The above relationship ensures that any estimator based on the empirical risk minimization principle automatically selects low variance elements. Hence, intuitively, the variance terms arising from an application of Talagrand's concentration argument are automatically controlled. Such a linear variance-expectation relationship predates the developments in the line of work concerning local Rademacher averages. Indeed, it is a standard assumption used in the empirical processes analysis of M-estimators (see, e.g., the works by van de Geer [202], Massart [132]).

The remaining question is what is the smallest allowed value of the localization radius $r > 0$, such that Talagrand's inequality (2.10) applied to $\mathcal{F}(r)$ yields an upper bound on the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$. Using a peeling argument applied to a reweighted excess loss class (cf. Bartlett, Bousquet, and Mendelson [19, Section 3]), this value can be shown to equal a solution to a certain fixed-point equation, leading to the following definition.

Definition 2.5 (Local Rademacher Complexity)**.** Let P_X denote any distribution supported on $\mathcal{X}$ and let $\mathcal{H}$ denote any class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. For $r > 0$, let $\mathcal{H}(r) = \{h \in \mathcal{H} : \mathbf{E}_{X \sim P_X}[h(X)^2] \leq r\}$. Let $\sigma = (\sigma_i)_{i=1}^n$ be a sequence of i.i.d. Rademacher (i.e., symmetric and $\{\pm 1\}$-valued) random variables and let $S_n^X = (X_i)_{i=1}^n$ denote n independent random variables distributed according to P_X. Then, for any $\gamma > 0$, the local Rademacher complexity of the class $\mathcal{H}$ is defined by

$$\mathfrak{R}_n^{\mathrm{loc}}(P_X, \mathcal{H}, \gamma) = \inf \left\{ r > 0 : \mathbf{E}_{S_n^X, \sigma} \left[\sup_{h \in \mathcal{H}(\gamma^{-1} r)} \left\{ \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i) \right\} \right] \leq r \right\}.$$

Let us now state a precise result obtained by Bartlett, Bousquet, and Mendelson [19] (see also [216, Theorem 14.20]). In our notation, this result reads as follows.

Theorem 2.1 (Corollary 5.3 in [19]). *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $[-b, b]$ for some $b > 0$. Let P be a distribution supported on $\mathcal{X} \times [-b, b]$ and let $g^\star \in \operatorname{argmin}_{g \in \mathcal{G}} R(g)$, where $\mathcal{G}$ is some reference class of functions. Suppose that the following three conditions hold:*

1. *The loss function $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ is C_b-Lipschitz in its first argument;*
2. *The tuple $(P, \ell, \mathcal{F}, g^\star)$ satisfies the Bernstein condition with parameter $\gamma > 0$;*
3. *The function class $\mathcal{F} - g^\star = \{f - g^\star : f \in \mathcal{F}\}$ is star-shaped around 0, that is, $h \in \mathcal{F} - g^\star$ implies that for any $\lambda \in [0, 1]$ we have $\alpha h \in \mathcal{F} - g^\star$.*

Let $\widehat{f}$ be an estimator such that $R_n(\widehat{f}) - R_n(g^\star) \leq 0$ almost surely. Then, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$, we have

$$\mathcal{E}(\widehat{f}, \mathcal{G}) \leq c_1 C_b \mathfrak{R}_n^{loc}(P_X, \mathcal{F} - g^\star, C_b^{-1}\gamma) + c_2 \frac{(C_b b + C_b^2 \gamma^{-1}) \log(1/\delta)}{n},$$

where $c_1, c_2 > 0$ are universal constants.

In Chapter 3, we will see how the Bernstein condition can be replaced by an alternative estimator-dependent condition that is possible to satisfy in improper learning settings. We now turn to the next section, where we introduce an approach to localization on which the results of Chapter 3 build.

2.5.2 The Offset Rademacher Complexity Approach

We now describe the offset Rademacher complexity approach due to Liang, Rakhlin, and Sridharan [120], which originates in the prior work on online learning by Rakhlin and Sridharan [170]. This approach was introduced for the quadratic loss $\ell(y, y') = (y - y')^2$, although as we shall see in Chapter 3, there are no technical roadblocks for replacing the quadratic loss with any Lipschitz and strongly convex loss function.

Liang, Rakhlin, and Sridharan [120] use the star estimator algorithm of Audibert [6] to demonstrate the applicability of offset Rademacher complexities. Recall the model selection aggregation problem setting discussed in Section 2.2. We are given a finite dictionary of functions $\mathcal{G} = \{g_1, \ldots, g_m\}$, whose absolute values are uniformly bounded by some constant $b > 0$. The goal is to construct a new function whose excess risk is small compared to the best function in $\mathcal{G}$. From the work of Tsybakov [198], it is known that the optimal excess risk rate for model selection aggregation is of order $b^2 \log(m)/n$. Observe that due to the non-convexity of the class $\mathcal{G}$, the Bernstein condition introduced in the previous section does not hold. In fact, it cannot hold,

for otherwise, local Rademacher complexity bounds would imply that the empirical risk minimization estimator over $\mathcal{G}$ achieves the optimal rate. However, the optimal aggregation rate is unachievable by any procedure that outputs a member of the class $\mathcal{G}$ (in other words, a proper learning procedure) as shown in, for example, the works [42, 94, 175] and the references therein.

Let us now demonstrate the offset Rademacher complexity approach of Liang, Rakhlin, and Sridharan [120]. First, recall that $S_n = (X_i, Y_i)_{i=1}^n$ denotes the observed data sample of n input-output pairs, and R_n denotes the empirical risk functional. Audibert's star estimator is defined as follows:

$$\widehat{f}^{(\mathrm{star})} = \operatorname{argmin}_{f\in\mathcal{G},\lambda\in[0,1]} R_n(\lambda \widehat{f}^{(\mathrm{ERM})} + (1-\lambda)f), \text{ where } \widehat{f}^{(\mathrm{ERM})} = \operatorname{argmin}_{f\in\mathcal{G}} R_n(f).$$

The key observation [120, Lemma 1] is that the star estimator satisfies the following inequality for any data sample S_n:

$$R_n(\widehat{f}^{(\mathrm{star})}) - R_n(g^\star) \leq -\frac{1}{18}\sum_{i=1}^n (f^{(\mathrm{star})}(X_i) - g^\star(X_i))^2. \tag{2.11}$$

The obtained negative term on the right-hand side is intuitively the term that should compensate for the variance of the selected function. Recall that for the empirical risk minimization estimator, compensation for the variance terms appearing in the excess risk bounds is achieved via the assumed Bernstein condition (cf. Definition 2.4). However, as discussed above, the Bernstein condition does not hold in the setting of model selection aggregation. While the inequality obtained above is not explicitly stated in the work of Audibert [6], who introduced the star aggregation algorithm, such negative quadratic terms arising through the design of the star estimator play a crucial role in Audibert's original proof.

To bound the excess risk of the star estimator, the condition (2.11) is exploited by Liang, Rakhlin, and Sridharan [120] as follows:

$$\begin{aligned}
&\mathcal{E}(f^{(\mathrm{star})}, \mathcal{G}) \\
&= (R(f^{(\mathrm{star})}) - R(g^\star)) - (R_n(f^{(\mathrm{star})}) - R_n(g^\star)) + (R_n(f^{(\mathrm{star})}) - R_n(g^\star)) \\
&\leq (R(f^{(\mathrm{star})}) - R(g^\star)) - (R_n(f^{(\mathrm{star})}) - R_n(g^\star)) - \gamma\frac{1}{n}\sum_{i=1}^n (f^{(\mathrm{star})}(X_i) - g^\star(X_i))^2 \\
&\leq \sup_{f\in\mathcal{F}} \Big\{ (R(f) - R(g^\star)) - (R_n(f) - R_n(g^\star)) - \gamma\frac{1}{n}\sum_{i=1}^n (f(X_i) - g^\star(X_i))^2 \Big\}.
\end{aligned}$$

Taking expectations on both sides and applying classical symmetrization and contraction arguments (cf. [120, Theorem 3]), we obtain the following bound for some large

enough universal constants $c_1, c_2 > 0$:

$$\mathbf{E}_{S_n}\mathcal{E}(f^{(\mathrm{star})}, \mathcal{G}) \leq c_1 b \mathbf{E}_{S_n,\sigma}\left[\sup_{h\in\mathcal{F}-g^\star}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i) - \frac{\gamma}{b}h(X_i)^2\right\}\right],$$

where $\sigma = (\sigma_1)_{i=1}^n$ denotes a sequence of i.i.d. Rademacher random variables (i.e., $\{\pm 1\}$-valued symmetric random variables). The right-hand side of the above equation is called the *offset Rademacher complexity* of the class $\mathcal{F} - g^\star$. The negative quadratic terms produce a localization phenomenon similar to the one achieved via the classical localization approach via fixed-points, discussed in the previous section.

The primary limitation of the above excess risk bound is that it only holds in expectation. High probability bounds in terms of offset Rademacher averages will be developed in Chapter 3. However, let us remark that some high-probability excess risk bounds based on the approach discussed above have been obtained in terms of other notions of complexity. Liang, Rakhlin, and Sridharan [120, Theorem 4] obtain a high-probability excess risk bound in terms of a multiplier-type Rademacher process with an offset negative term; however, the obtained result does not hold in a distribution-free sense, and it upper bounds the excess risk in terms of another *random variable* (as opposed to a deterministic problem-dependent quantity, such as local Rademacher complexity) that is not necessarily easier to control than the excess risk itself. Also, the very recent work of Vijaykumar [213] extends the geometric inequality (2.11) to general loss functions. Some non-trivial "fast rate" excess risk bounds that hold with high probability are obtained therein. However, the bounds obtained by Vijaykumar [213] are expressed in terms of empirical covering numbers for *worst-case* deterministic data. This notion of complexity is not as sharp as local Rademacher averages discussed in the previous section and it already yields suboptimal bounds for the model selection aggregation problem, suffering from excess logarithmic terms. As we shall see in Chapter 3, it is possible to obtain the deviation-optimal excess risk rate for the star algorithm using the notion of offset Rademacher complexity.

2.6 Summary of Notation

We summarize the notation used throughout this thesis in Table 2.1.

Notation	Description
n	The number of observations.
$\mathcal{X}$	The input space.
$\mathcal{Y}$	The output space.
P	The data generating distribution supported on $\mathcal{X} \times \mathcal{Y}$.
P_X	The marginal distribution of the covariates X.
S_n	The observations $S_n = (X_i, Y_i)_{i=1}^n$ sampled i.i.d. from P.
P_n	The empirical distribution supported on S_n.
ℓ	The loss function that maps $\mathcal{Y} \times \mathcal{Y}$ to $[0, \infty)$.
ℓ_f	For $f : \mathcal{X} \to \mathbb{R}$ we define $\ell_f(X, Y) = \ell(f(X), Y)$.
$R(f)$	The population risk $\mathbf{E}_{(X,Y)\sim P}[\ell(f(X), Y)]$.
$P\ell_f$	Another way to write $R(f)$ (cf. Section 2.1).
$R_n(f)$	The empirical risk $\frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i)$.
$P_n\ell_f$	Another way to write $R_n(f)$ (cf. Section 2.1).
$\widehat{f}$	A procedure mapping datasets S_n to some function space $\mathcal{F}$.
$\mathcal{F}$	The set of possible values taken by $\widehat{f}(S_n)$ (i.e., the range of $\widehat{f}$).
f	An element of $\mathcal{F}$.
$\mathcal{G}$	A reference class of functions mapping $\mathcal{X}$ to $\mathcal{Y}$.
g	An element of $\mathcal{G}$.
$\mathcal{E}(\widehat{f}, \mathcal{G})$	The excess risk random variable $R(\widehat{f}) - \inf_{g\in\mathcal{G}} R(g)$.
$g^\star$	Any function $g \in \mathcal{G}$ that minimizes $R(g)$.
$\|f\|_{L_2(P)}^2$	Denotes the squared $L_2(P_X)$ norm $Pf^2 = \mathbf{E}_{X\sim P_X}[f(X)^2]$
$\|f\|_n^2$	Denotes the squared empirical L_2 norm $P_nf^2 = \frac{1}{n}\sum_{i=1}^n f(X_i)^2$.
$\mathfrak{R}_n^{\mathrm{loc}}(P_X, \mathcal{H}, \gamma)$	The local Rademacher complexity defined in Section 2.5.1.
$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma)$	The offset Rademacher complexity defined in Section 3.4.1.
ψ	A mirror map defined in Section 2.3.
D_f	Bregman divergence induced by a function f (see Section 2.3).
$R(w)$	Population risk of the linear function $\langle w, \cdot\rangle$.
$R_n(w)$	Empirical risk of the linear function $\langle w, \cdot\rangle$.
$a \lesssim b$	A shorthand for writing $a \le cb$ for some absolute constant $c > 0$.
$a \sim b$	A shorthand for the simultaneous inequalities $a \lesssim b$ and $b \lesssim a$.

Table 2.1: Table of notation

3 Exponential-Tail Excess Risk Bounds via Offset Rademacher Complexities

This chapter is based on results obtained in collaboration with Varun Kanade and Patrick Rebeschini, to be submitted for publication. The aim of this chapter is to obtain sharp excess risk bounds that hold with high probability and are applicable for possibly improper statistical estimators.

The results of this chapter assume familiarity with the mathematical framework of Statistical Learning (Section 2.1), the concept of improper learning (Section 2.2) and local Rademacher complexity measures (Section 2.5). For a summary of notation used throughout this thesis, we refer to Section 2.6.

3.1 Introduction

Local Rademacher averages serve as a principal tool for obtaining sharp performance guarantees for statistical estimators based on the principle of empirical risk minimization. However, recent years have witnessed several example problem settings where optimal statistical performance is only achievable via improper statistical estimators originating from aggregation theory. In this line of work, a fundamental problem not covered via the classical theory of local Rademacher complexities is model selection aggregation, where a statistician aims to construct a predictor that is as good as the best one in a given finite set of bounded functions.

In this chapter, we build on the approach to localization via offset Rademacher complexities (see Section 2.5.2), for which high probability theory has yet to be established. Our main result is an exponential-tail excess risk bound expressed in terms of the offset Rademacher complexity, which yields bounds at least as sharp as those obtainable via the classical theory (in the sense to be explained in Section 3.4.4). However, our bound applies under an estimator-dependent geometric condition, instead of the estimator-independent but distribution dependent Bernstein condition on which the classical localization theory is built. Consequently, the results obtained in this chapter apply to the improper prediction regimes not directly covered via the classical theory.

There exist several different general-purpose techniques for obtaining sharp excess risk bounds. We shall group them into two categories: those that obtain bounds in expectation and those that yield excess risk bounds that hold with high probability. We provide a brief review of the existing approaches for obtaining excess risk bounds

in Section 3.8. However, without going into too much detail, let us point out that high-probability bounds are of special interest in improper learning settings, where an obtained excess risk bound that holds in expectation does not imply any meaningful deviation-type bounds. The latter claim follows from the fact that in improper learning setups, the excess risk random variable can take negative values. In contrast, for proper estimators, an in-expectation excess risk bound can be converted (at least) to polynomial-tail deviation bounds via Markov's inequality. Beyond the difficulties related to improper learning, converting in-expectation guarantees to high-probability counterparts is challenging if one wishes to obtain bounds that decay faster than the "slow rate" $1/\sqrt{n}$, where n is the sample size. This is because applying many of the most popular concentration tools will result in a sub-Gaussian-type variance term of order $1/\sqrt{n}$ that may be difficult (but not impossible) to extinguish in general.

Among the most successful general-purpose tools for obtaining sharp excess risk upper bounds that hold with high probability is the *local Rademacher complexity* [19, 105, 106]. This approach automatically comes with exponential-tail guarantees due to the underlying mathematical machinery resting on a powerful concentration bound for controlling the supremum of empirical processes due to Talagrand [191, 192]. At the same time, local Rademacher averages are relatively simple to upper bound, with many settings of interest covered in the existing literature; for some examples, see the textbook by Wainwright [216, Chapters 13 and 14].

Due to technical reasons related to the so-called Bernstein condition (see Section 2.5.1 for a detailed discussion), local Rademacher complexity bounds are primarily suitable when two conditions hold: $\mathcal{G}$ is convex and $\mathcal{F} = \mathcal{G}$. A setup when $\mathcal{F} = \mathcal{G}$ is called *proper*. Soon after the development of local Rademacher complexities, it was noticed in the discussion paper by Tsybakov [197] that such restrictions fail to include a very natural problem called *model selection aggregation* [155, 198]. In this problem, the reference class of functions $\mathcal{G}$ is taken to be a finite set of bounded functions; particularly, it is a non-convex set, and local Rademacher complexity theory does not apply directly. Understanding how to optimally aggregate statistical models constructed from i.i.d. data (e.g., models arising from different tuning parameters or different statistical estimators) is a fundamental problem in Statistics. At the same time, deviation-optimal model selection aggregation procedures have been used to construct computable procedures (not necessarily computationally efficient) to demonstrate the achievability of some statistical minimax lower bounds (see, e.g., [171, 145, 154]).

The phenomenon concerning deviation-optimality of model selection aggregation estimators has generated a lot of attention in the Mathematical Statistics community; for example, see the works by Lecué and Mendelson [112], Rigollet [175], Dai, Rigollet, and Zhang [58], Lecué and Rigollet [115], Wintenberger [220], Bellec [22] for analysis of different model selection aggregation procedures. More broadly, the analysis of improper statistical estimators is becoming an increasingly important problem, as such procedures were shown to be necessary for optimal statistical performance in logistic regression, see [85, 67, 153], and linear regression, see the discussions in Chapters 5 and 6 of this thesis.

3.2 Problem Formulation

We consider a bounded problem setting. We assume that the unknown data-generating mechanism P is supported on $\mathcal{X} \times [-b, b]$ for some constant $b > 0$, but we do not impose further restrictions on P; that is, we consider the distribution-free setting under boundedness constraints only. Further, we assume that the loss function $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ is C_b-Lipschitz in its first argument. More specifically, we assume that for any $y \in [-b, b]$, the function $\ell(\cdot, y)$ is C_b-Lipschitz. Finally, we let $\mathcal{F}$ and $\mathcal{G}$ denote classes of functions mapping $\mathcal{X}$ to $[-b, b]$. As usual, $\mathcal{F}$ denotes the range of some estimator of interest, while $\mathcal{G}$ denotes some chosen reference class of functions against which the excess risk is computed. Finally, we let $g^\star$ be any function that minimizes the population risk over $\mathcal{G}$, assuming without loss of generality that such a function exists; otherwise, it could be replaced by any function in $\mathcal{G}$ whose excess risk is sufficiently close to the infimum over the whole class.

In the case of proper learning, the above setup is well covered via the classical local Rademacher complexity theory introduced in Section 2.5.1. Indeed, for an estimator $\widehat{f}$, under some technical conditions stated in Theorem 2.1, the following excess risk bound holds with probability at least $1 - \delta$:

$$\mathcal{E}(\widehat{f}, \mathcal{G}) \lesssim C_b \mathfrak{R}_n^{\mathrm{loc}}(P_X, \mathcal{F} - g^\star, C_b^{-1}\gamma) + \frac{(C_b b + C_b^2 \gamma^{-1}) \log(1/\delta)}{n}.$$

The parameter $\gamma > 0$ in the above equation is the one appearing in the *Bernstein condition* (cf. Definition 2.4), which asserts that the following is satisfied for any $f \in \mathcal{F}$:

$$\mathbf{E}_{X \sim P_X}(f(X) - g^\star(X))^2 \leq \frac{1}{\gamma} \mathbf{E}_{(X,Y) \sim P} \left[\ell_f(X, Y) - \ell_{g^\star}(X, Y) \right]. \tag{3.1}$$

As already discussed in Section 2.5.1, the above condition does not generally hold in distribution-free improper learning settings. Indeed, note that if the reference class

of functions $\mathcal{G}$ is a strict subset of $\mathcal{F}$, then for some distribution P, the right-hand side of the above equation can be made negative, while the left-hand side is always non-negative.

In Section 2.5.2, we have also discussed an alternative approach to localization, namely, the offset Rademacher complexity approach. We have seen that this approach provides an avenue for obtaining sharp *expected* excess risk bound without relying on the Bernstein condition (3.1).

> The problem addressed in the present chapter is obtaining *exponential-tail* excess risk bounds in terms of offset Rademacher averages, crucially, without imposing the Bernstein condition (3.1).

As a final remark, observe that whether the above goal is possible to achieve may not be a priori self-evident. Indeed, as discussed in Section 2.5.1, the Bernstein condition is (in part)[1] introduced in the classical local Rademacher complexities approach to compensate for the "slow rate" variance term arising from an application of Talagrand's inequality. Hence, should we try to base our approach on the same mathematical machinery, we would likely run into the same issues, leading to the introduction of the undesirable (in our context) Bernstein condition.

3.3 Summary of Contributions

In this chapter, we obtain *exponential-tail* excess risk upper bounds that hold for a *general class* of estimators satisfying a certain geometric condition that we call *offset condition* (see Definition 3.1). This geometric condition can serve as a design principle for statistical estimators that satisfy sharp excess risk guarantees with high probability and in particular, arguments based on convex geometry can be used to establish that such a condition holds for a broad class of known estimators (see the examples in Section 3.5). The class of estimators satisfying the geometric condition includes improper learning settings that are not covered by the classical theory of local Rademacher complexities. In the classical setting of empirical risk minimization performed over a convex class under boundedness assumptions, our complexity measure yields results *at least as sharp* as those obtainable by the classical theory of local Rademacher complexities (this is made more precise in Section 3.4.4). The starting point of our analysis is the work of Liang, Rakhlin, and Sridharan [120], who were the

[1] The Bernstein condition would also play a role in expectation-only bounds using fixed-point local Rademacher complexity approach.

first to provide an *in-expectation* analysis of the star aggregation algorithm based on *offset Rademacher complexity*, a modified notion of classical localization that arises from the analysis of *offset empirical processes*.

The main contribution of the current chapter is obtaining results analogous to the ones achievable via the classical local Rademacher complexity theory, yet applicable under a different set of assumptions. In particular, the main element of the classical theory is an *estimator-independent* Bernstein condition (see Section 2.5.1 for details) that ensures a linear relationship between the variance and expectation of the excess loss class. In contrast, our results build on an *estimator-dependent* geometric condition, called offset condition. The theory developed in this chapter shows that the offset condition is sufficient to ensure sharp excess risk guarantees for possibly improper estimators. For example, as discussed in Section 3.5, any estimator that satisfies the offset condition while outputting a sparse combination of a given finite dictionary of functions, attains deviation-optimal excess risk rate for the problem of model selection aggregation, where improperness is necessary for optimality.

The main results are presented in Section 3.4.

- Section 3.4.1 contains the definition of the geometric condition (called offset condition) that serves as our replacement of the Bernstein condition and the definition of offset Rademacher complexity, which is slightly modified from the one appearing in prior work by Liang, Rakhlin, and Sridharan [120]. Specifically, we include additional negative terms, which play an important role in our concentration arguments and in proving that our notion of complexity is never worse than the classical notion of local Rademacher complexities (cf. Lemma 3.1).

- Section 3.4.2 contains a moment generating function bound for shifted multiplier empirical processes (Proposition 3.1), which is the main technical contribution of the present chapter. This result serves as our replacement for Talagrand's concentration inequality, on which the classical theory of localization is built. The key feature of our concentration result is the fact that the variance of the supremum of shifted multiplier processes is automatically controlled by a linear function of their expectations due to the presence of the negative quadratic terms inside the supremum. In contrast, the classical theory of localization needs to *assume* that a certain variance-expectation relationship holds, as elaborated in Section 2.5.1. We prove Proposition 3.1 via an application of an exponential Efron-Stein inequality as discussed in greater detail in Section 3.6.2.

- In Section 3.4.3, we present our main theorem – an exponential-tail excess risk bound stated in terms of the offset Rademacher complexity (cf. Theorem 3.1). The key difference from the usual theory of localization is that the estimator-independent Bernstein condition appearing in Theorem 2.1 is replaced via the estimator-dependent offset condition. We prove Theorem 3.1 by bounding Laplace transform of the shifted empirical processes (arising through the geometric condition imposed on an estimator) in terms of a Laplace transform of a related shifted multiplier empirical process. We then complete the proof via an application of Proposition 3.1.

- Further connections between the classical theory and the theory developed in this chapter are discussed in Section 3.4.4. In Lemma 3.1, we show that the offset Rademacher complexity is at most as large as the classical local Rademacher complexity. Thus, the bounds obtained in this chapter, when they apply, are at least as sharp as those obtainable via the classical theory (cf. Corollary 3.1). Finally, we discuss the sense in which the Bernstein condition and the offset condition can be considered as dual to one another, particularly when the roles of empirical and population quantities are interchanged (cf. Lemma 3.2).

Section 3.5 contains example applications of the theory developed in this chapter. In Lemma 3.3, we bound the offset Rademacher complexity of sparse linear classes; in Corollary 3.2, we show how this bound can be applied for non-linear classes via a change-of-basis argument. As a direct consequence, we show how our theory can yield deviation-optimal bounds for two different model selection aggregation procedures, both of which output a sparse combination of dictionary elements and satisfy the offset condition. Such applications are outside the scope of the classical theory of localization, due to the necessary improperness of optimal estimators, as discussed in the introduction.

The proofs are deferred to Section 3.6.

3.4 Main Results

The main results of this chapter are presented in this section. In Section 3.4.1, we introduce the geometric condition (called offset condition) used to replace the Bernstein condition; further, we define the offset Rademacher complexity (slightly modified from the one appearing in prior works) used to replace the classical notion of local Rademacher complexity. Section 3.4.2 contains a moment generating function bound for shifted multiplier empirical processes. This result serves as our replacement

for Talagrand's concentration inequality, the foundation of the classical theory of localization. Section 3.4.3 contains a high probability excess risk bound in terms of the offset Rademacher complexity; this result applies in the settings where the Bernstein condition does not hold. Finally, in Section 3.4.4, we provide a comparison between the offset and Bernstein conditions and demonstrate that the theory presented in this chapter can recover the classical distribution-free bounds overviewed in Section 2.5.1.

3.4.1 Definitions

We begin with the definition of the *offset condition*. Observe that this condition is *estimator-dependent*, as opposed to the Bernstein condition (cf. Definition 2.4).

Definition 3.1 (Offset Condition)**.** Let $\mathcal{G}$ be a class of functions mapping $\mathcal{X}$ to $[-b, b]$ for some $b > 0$. Fix a loss function $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ and recall that R_n denotes the induced empirical risk functional. Let $\varepsilon : [0, 1] \to \mathbb{R}$ be some function and let $\gamma > 0$ be some positive real number. Let P be a distribution supported on $\mathcal{X} \times \mathcal{Y}$. An estimator $\widehat{f}$ satisfies the *offset condition* with respect to $(\mathcal{G}, \ell, \varepsilon, \gamma)$ for the distribution P, if for any any $\delta \in [0, 1]$ the following holds:

$$\mathbf{P}_{S_n}\left(R_n(\widehat{f}) - R_n(g^\star) \leq -\gamma \sum_{i=1}^{n} (\widehat{f}(X_i) - g^\star(X_i))^2 + \varepsilon(\delta)\right) \geq 1 - \delta,$$

where $S_n = (X_i, Y_i)_{i=1}^n$ is an i.i.d. sample drawn from the distribution P and $g^\star = g^\star(\mathcal{G}, P, \ell)$ denotes any population risk minimizer in the class $\mathcal{G}$.

Whenever the following deterministic inequality holds for any sample $S_n = (X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$:

$$R_n(\widehat{f}) - R_n(g^\star) \leq -\gamma \sum_{i=1}^{n} (\widehat{f}(X_i) - g^\star(X_i))^2 + \varepsilon,$$

we say that the estimator $\widehat{f} = \widehat{f}(S_n)$ satisfies the *deterministic offset condition* with respect to $(\mathcal{G}, \ell, \varepsilon, \gamma)$.

In the above definition the function $\varepsilon(\cdot)$ allows for the offset condition to fail with probability δ, while incurring a penalty $\varepsilon(\delta)$. As we shall see in Section 3.5, such a condition naturally enters the analysis of some improper estimators. Also, we will discuss some example estimators that satisfy the deterministic offset condition.

In Section 2.5.1, we described how Bernstein condition implies local Rademacher complexity excess risk bounds for empirical risk minimization estimators. Likewise, we shall see that offset condition implies excess risk bounds expressed in terms of the offset Rademacher complexity defined below.

Definition 3.2 (Offset Rademacher Complexity)**.** Let P_X be any distribution supported on $\mathcal{X}$ and let $\mathcal{H}$ be any class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Let $\sigma = (\sigma_i)_{i=1}^n$ denote a sequence of i.i.d. Rademacher (i.e., symmetric and $\{\pm 1\}$-valued) random variables and let $S_n^X = (X_i)_{i=1}^n$ denote n independent random variables distributed according to P_X. Then, for any $\gamma > 0$, the offset Rademacher complexity of the class $\mathcal{H}$ is defined by

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma) = \mathbf{E}_{S_n^X, \sigma}\left[\sup_{h \in \mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i) - \gamma h(X_i)^2 - \gamma \mathbf{E}_{X \sim P_X}[h(X)^2]\right\}\right].$$

Let us remark that our definition above differs from the one presented in Section 2.5.2 since we include extra negative terms $-\gamma \mathbf{E}_{X \sim P_X}[h(X)^2]$ inside the above supremum. This refinement is necessary for our concentration argument to work, since we establish moment bounds for shifted multiplier processes that contain negative population terms (cf. Section 3.4.2). At the same time, the inclusion of the negative quadratic population terms allows us to show that the above notion of complexity is at least as sharp as the classical one introduced in Definition 2.5 (see Lemma 3.1 in Section 3.4.4 for details).

3.4.2 Concentration of Shifted Multiplier Processes

The primary technical tool in this chapter is the following proposition, which proves a Bernstein-type one-sided concentration bound for the supremum of shifted multiplier processes (defined below in Equation (3.2)). This proposition plays a crucial role in establishing our main result, Theorem 3.1 presented in the next section. In particular, provided that an estimator of interest satisfies the offset condition, we will show that the moment generating function of its excess risk can be controlled by the moment generating function of a certain shifted multiplier process. We defer the proof of the below proposition Section 3.6.2.

Proposition 3.1. *Let $\mathcal{H}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Further, let $P_{(X,\zeta)}$ be a joint distribution on $\mathcal{X} \times \mathbb{R}$ with marginal distributions P_X and P_ζ, and let $S_n = (X_i, \zeta_i)_{i=1}^n$ be a set of n i.i.d. samples from $P_{(X,\zeta)}$. Fix any positive constant $\gamma > 0$ and define a random variable $U = U(S_n)$ to be the supremum of the offset multiplier process as follows:*

$$U = \sup_{h \in \mathrm{star}(\mathcal{H})}\left\{\sum_{i=1}^n \zeta_i h(X_i) - \mathbf{E}_{(X,\zeta) \sim P_{(X,\zeta)}}[\zeta h(X)] - \gamma h(X_i)^2 - \gamma \mathbf{E}_{X \sim P_X}[h(X)^2]\right\}. \tag{3.2}$$

Suppose that there exist positive constants κ and σ such that $\sup_{h\in\mathcal{H}} \|h\|_{L_\infty(P_X)} \leq \kappa$ and $\|\zeta\|_{L_\infty(P_\zeta)} \leq \sigma$. Then, for $\eta = 8(\sigma^2\gamma^{-1} + \gamma\kappa^2)$ and any $\lambda \in (0, 1/\eta)$ the following holds:

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \leq \frac{\lambda^2 \eta \mathbf{E}U}{2(1-\eta\lambda)}.$$

Before turning to the offset Rademacher complexity upper bounds, let us remark that in the above moment bound, the variance proxy/variance factor (in the sense of [31, Section 2.4]) is equal to $\eta\mathbf{E}U$; thus the variance of the random variable U is automatically controlled by its expectation. In particular, the above bound can be transformed into deviation bounds of the form $U \leq 2\mathbf{E}[U] + c\eta\log(1/\delta)$, where $\delta > 0$ is the confidence parameter. In contrast, recall that the variance proxy in Talagrand's concentration inequality (2.10) is not controlled by the expectation of the corresponding empirical process, which in turn leads to the localization machinery where Rademacher averages need to be computed over explicitly constrained subsets of the function class of interest. In contrast, using the above concentration result, our theory allows us to obtain high probability bounds in terms of the offset Rademacher complexity, as we show in the following section.

3.4.3 Exponential-Tail Offset Rademacher Complexity Bound

We now present the main result of this chapter, the proof of which can be found in Section 3.6.1. The following theorem provides an alternative to Theorem 2.1, but with Bernstein condition replaced via offset condition. As a consequence, the below theorem is applicable to potentially improper estimators; see the examples in Section 3.5.

Theorem 3.1. *Let $\widehat{f}$ be an estimator with range $\mathcal{F}$, where $\mathcal{F}$ denotes a class of functions mapping $\mathcal{X}$ to $[-b, b]$ for some $b > 0$. Let P be any distribution supported on $\mathcal{X} \times [-b, b]$ and denote $g^\star \in \operatorname{argmin}_{g\in\mathcal{G}} R(g)$, where $\mathcal{G}$ is some reference class of functions. Suppose that the following two conditions hold:*

1. *The loss function $\ell : [-b, b]\times[-b, b] \to [0, \infty)$ is C_b-Lipschitz in its first argument;*
2. *The estimator $\widehat{f}$ satisfies the offset condition with respect to $(\mathcal{G}, \ell, \varepsilon, \gamma)$ for the distribution P, where ε is some function mapping $[0, 1]$ to $\mathbb{R}$ and $\gamma > 0$ is some positive real number.*

Then, for any $\delta_1, \delta_2 \in (0, 1)$ with probability at least $1 - \delta_1 - \delta_2$, we have

$$\mathcal{E}(\widehat{f}, \mathcal{G}) \leq c_1 C_b' \mathfrak{R}_n^{off}(P_X, \operatorname{star}(\mathcal{F} - g^\star), (C_b')^{-1}\gamma) + c_2 \frac{\gamma^{-1}(C_b')^2 \log(1/\delta_1)}{n} + \varepsilon(\delta_2),$$

where $c_1, c_2 > 0$ are some universal constants and $C_b' = C_b + \gamma b$.

Remark 3.1. In comparison with Theorem 2.1, the above result replaces C_b with a worse constant $C_b' = C_b + \gamma b$. However, the primary application domain where above theorems hold is the setting where for any $y \in [-b, b]$, the function $\ell(\cdot, y)$ is C_b-Lipschitz and γ-strongly convex in the fist argument (see section 3.5 for examples). In such a setting it can be shown that $\gamma b \leq C_b$ and hence $C_b' \leq 2C_b$.

3.4.4 Recovering Local Rademacher Complexity Results Without Bernstein Condition

In this section we discuss that Theorem 3.1 yields excess risk bounds that are no worse than the ones stated in Theorem 2.1. We begin by stating the following lemma, which is proved in Appendix 3.6.3.

Lemma 3.1. *Let P_X be any distribution supported on $\mathcal{X}$ and let $\mathcal{H}$ be any star-shaped class of functions (i.e., $\mathcal{H} = \mathrm{star}(\mathcal{H})$) mapping $\mathcal{X}$ to $\mathbb{R}$. Then, for any $\gamma > 0$ we have*

$$\mathfrak{R}_n^{off}(P_X, \mathcal{H}, \gamma) \leq \mathfrak{R}_n^{loc}(P_X, \mathcal{H}, \gamma).$$

An immediate consequence of the above lemma is the following corollary, which shows that the classical local Rademacher complexity bounds hold when the Bernstein condition is replaced via the *estimator-dependent* offset condition.

Corollary 3.1. *Consider the setting of Theorem 3.1. For any $\delta_1, \delta_2 \in (0, 1)$ with probability at least $1 - \delta_1 - \delta_2$, we have*

$$\mathcal{E}(\widehat{f}, \mathcal{G}) \leq c_1 C_b' \mathfrak{R}^{loc}(P_X, \mathrm{star}(\mathcal{F} - g^\star), (C_b')^{-1}\gamma) + c_2 \frac{\gamma^{-1}(C_b')^2 \log(1/\delta_1)}{n} + \varepsilon(\delta_2),$$

where $c_1, c_2 > 0$ are some universal constants and $C_b' = C_b + \gamma b$.

It remains to discuss the relationship between the offset and Bernstein conditions. A typical example where the Bernstein condition holds for any distribution P is when $\mathcal{F} = \mathcal{G}$ is a convex class, and the loss function is strongly convex. In such regimes, any empirical risk minimizer over $\mathcal{F}$ satisfies the offset condition. Thus, when applied to empirical risk minimization estimator, the offset condition can be seen as a dual condition to the Bernstein condition, where the roles played by empirical and population quantities are interchanged. We formalize this observation in the lemma below.

Lemma 3.2. *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a loss function and let $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ be the set of all distributions P supported on $\mathcal{X} \times \mathcal{Y}$. Let*

$f^\star = f^\star(\mathcal{F}, P, \ell)$ be any population risk minimizer over $\mathcal{F}$. Let $\widehat{f}^{(ERM)}$ be an estimator that returns any empirical risk minimizer in class $\mathcal{F}$. If for any $P \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ the tuple $(P, \ell, \mathcal{F}, f^\star)$ satisfies the Bernstein condition with parameter γ, then the estimator $\widehat{f}^{(ERM)}$ satisfies the deterministic offset condition with respect to $(\mathcal{F}, \ell, 0, \gamma)$.

Proof. Given an i.i.d. sample $S_n = (X_i, Y_i)_{i=1}^n$ from some distribution $P \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, let P_n denote a distribution on $\mathcal{X} \times \mathcal{Y}$ assigning equal mass to each (X_i, Y_i). Since $P_n \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, by the assumption of this lemma $(P_n, \ell, \mathcal{F}, \widehat{f}^{(ERM)}(S_n))$ satisfies the Bernstein condition with parameter γ. This is equivalent to saying that $\widehat{f}^{(\mathrm{ERM})}$ satisfies the deterministic offset condition with respect to $(\mathcal{F}, \ell, 0, \gamma)$. □

3.5 Example Applications

In this section, we discuss some example applications to problems where the Bernstein condition does not hold, yet there exist estimators that satisfy the offset condition. As a result, sharp deviation-optimal excess risk rates can be obtained for such estimators via the theory developed in this paper.

In what follows, we will use the notation $a \lesssim b$ to denote the existence of some universal constant c such that $a \leq cb$. For any function class $\mathcal{H}$ mapping $\mathcal{X}$ to $\mathbb{R}$ and any sample $S_n^X = (X_i)_{i=1}^n$, where $X_i \in \mathcal{X}$, define

$$\mathfrak{R}^{\mathrm{off}}(S_n^X, \mathcal{H}, \gamma) = \mathbf{E}_\sigma \left[\sup_{h\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i) - \gamma h(X_i)^2 \right\} \middle| S_n^X \right],$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ denotes a sequence of i.i.d. Rademacher random variables. Observe, in particular, that for any distribution P_X supported on $\mathcal{X}$, we have

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma) \leq \mathbf{E}_{S_n^X} \left[\mathfrak{R}^{\mathrm{off}}(S_n^X, \mathcal{H}, \gamma) \right]. \tag{3.3}$$

Thus, upper bounds on $\mathfrak{R}_n^{\mathrm{off}}(S_n^X, \mathcal{H}, \gamma)$ imply corresponding upper bounds on the offset Rademacher complexity. Let us now state a bound on $\mathfrak{R}_n^{\mathrm{off}}(S_n^X, \mathcal{H}, \gamma)$ for sparse linear classes, which will be used to yield sharp bounds for the examples considered in this chapter.

Lemma 3.3. *For any $w \in \mathbb{R}^d$ let $\|w\|_0$ denote the number of non-zero coordinates of w. Denote a class of k-sparse linear predictors by*

$$\mathcal{H}_{lin}^{d,k} = \{\langle w, \cdot\rangle : w \in \mathbb{R}^d, \|w\|_0 \leq k\}.$$

Let $S_n^\Phi = (\Phi_i)_{i=1}^n$, where $\Phi_i \in \mathbb{R}^d$ are arbitrary. Then, for any $\gamma > 0$ we have

$$\mathfrak{R}^{off}(S_n^\Phi, \mathcal{H}_{lin}^{d,k}, \gamma) \lesssim \frac{1}{\gamma} \log\left(\frac{ed}{k}\right) \frac{k}{n}.$$

The above lemma is proved in Section 3.6.4 via a direct argument involving comparison inequalities for Rademacher and Gaussian chaos. As an immediate consequence, let us state the following corollary that will simplify the exposition of the example applications to follow.

Corollary 3.2. *Let $\mathcal{G} = \{g_1, \dots, g_m\}$ denote a finite class of arbitrary functions mapping $\mathcal{X}$ to $\mathbb{R}$. For any positive integer $k \in \{1, \dots, m\}$ define the function class containing k-sparse linear combinations of elements of $\mathcal{G}$ by*

$$\mathcal{G}_{lin}^k = \left\{ g_w(\cdot) = \sum_{i=1}^m w_i g_i(\cdot) : w \in \mathbb{R}^d \text{ and } \|w\|_0 \leq k \right\}$$

Let $k_1, k_2 \in \{1, \dots, m\}$, $\mathcal{F} = \mathcal{G}_{lin}^{k_1}$, and fix any $g^\star \in \mathcal{G}_{lin}^{k_2}$. Then, for any distribution P_X supported on $\mathcal{X}$ and for any $\gamma > 0$ we have

$$\mathfrak{R}_n^{off}(P_X, \mathrm{star}(\mathcal{F} - g^\star), \gamma) \lesssim \frac{1}{\gamma} \log\left(\frac{em}{(k_1 + k_2)}\right) \frac{(k_1 + k_2)}{n}.$$

Proof. Let $k = k_1 + k_2$ and note that $\mathrm{star}(\mathcal{F} - g^\star) \subseteq \mathcal{G}_{\mathrm{lin}}^k$. Hence, the bound (3.3) yields

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathrm{star}(\mathcal{F} - g^\star), \gamma) \leq \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{G}_{\mathrm{lin}}^k, \gamma) \leq \mathbf{E}_{S_n^X}\left[\mathfrak{R}^{\mathrm{off}}(S_n^X, \mathcal{G}_{\mathrm{lin}}^k, \gamma)\right]. \tag{3.4}$$

For any sample S_n^X and any $i = 1, \dots, n$ define $\Phi_i^X \in \mathbb{R}^m$ by $(\Phi_i^X)_j = g_j(X_i)$. Then, for any $w \in \mathbb{R}^d$ and $g_w = \sum_{i=1}^m w_i g_i$ we have $g_w(X_i) = \sum_{j=1}^m w_j g_j(X_i) = \langle w, \Phi_i^X \rangle$. Hence, letting $S_n^\Phi(S_n^X) = (\Phi_i^X)_{i=1}^n$ and applying Lemma 3.3 yields

$$\mathfrak{R}^{\mathrm{off}}(S_n^X, \mathcal{G}_{\mathrm{lin}}^k, \gamma) = \mathfrak{R}^{\mathrm{off}}(S_n^\Phi(S_n^X), \mathcal{F}_{\mathrm{lin}}^{m,k}, \gamma) \lesssim \frac{1}{\gamma} \log\left(\frac{em}{k}\right) \frac{k}{n}.$$

Plugging in the above inequality into (3.4) completes the proof. □

We now turn to the example applications.

3.5.1 Example Applications to Model Selection Aggregation

In a model selection aggregation problem, we are given a finite dictionary $\mathcal{G} = \{g_1, \dots, g_m\}$ of functions mapping $\mathcal{X}$ to $[-b, b]$. Given a sample $S_n = (X_i, Y_i)_{i=1}^n$, a statistical estimator $\widehat{f}$ aims to construct a new function such that the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$ is small with high probability. For the model selection aggregation problem, the Bernstein condition (cf. Section 2.5.1) does not hold due to the non-convexity of the reference class $\mathcal{G}$. In turn, the classical theory of localization does not directly apply for this problem.

In what follows, we consider loss functions $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ that are C_b-Lipschitz and γ-strongly convex in the first coordinate. More precisely, we assume that for any $y, y_1, y_2 \in [-b, b]$ we have $|\ell(y_1, y) - \ell(y_2, y)| \leq C_b |y_1 - y_2|$ and for any $\lambda \in [0, 1]$ we have $\ell(\lambda y_1 + (1-\lambda)y_2, y) \leq \lambda \ell(y_1, y) + (1-\lambda)\ell(y_2, y) - \frac{\gamma}{2}\lambda(1-\lambda)(y_1 - y_2)^2$.

An identical setup to the one described above was recently treated by Lecué and Rigollet [115], Wintenberger [220]. Optimal model selection aggregation rates $\gamma^{-1} C_b^2 \log(m/\delta)/n$ were obtained therein for the q-aggregation and online Bernstein aggregation procedures. Below, we show how the offset Rademacher complexity analysis yields the same rates for two other estimators: Audibert's star algorithm and the midpoint estimator.

Audibert's Star Algorithm. The star algorithm due to [6] is defined by

$$\widehat{f}^{(\mathrm{star})} = \mathrm{argmin}_{f \in \mathcal{G}, \lambda \in [0,1]} R_n(\lambda \widehat{f}^{(\mathrm{ERM})} + (1-\lambda) f), \text{ where } \widehat{f}^{(\mathrm{ERM})} = \mathrm{argmin}_{f \in \mathcal{G}} R_n(f).$$

Generalizing an argument of Liang, Rakhlin, and Sridharan [120, Lemma 1], the recent work Vijaykumar [213, Proposition 5] shows that $\widehat{f}^{(\mathrm{star})}$ satisfies the $(\mathcal{G}, \ell, 0, c\gamma)$-deterministic offset condition, where $c > 0$ is some universal constant.

In the view of Corollary 3.2, the range of the star estimator $\widehat{f}^{(\mathrm{star})}$ is equal to $\{\lambda g + (1-\lambda) g' : g, g' \in \mathcal{G}, \lambda \in [0,1]\} \subseteq \mathcal{G}^2_{\mathrm{lin}}$. Thus, combining Theorem 3.1 (see also Remark 3.1) and Corollary 3.2 yields, for any $\delta \in (0,1)$ with probability at least $1 - \delta$

$$\mathcal{E}(f^{(\mathrm{star})}, \mathcal{G}) \lesssim \gamma^{-1} C_b^2 \frac{\log(m/\delta)}{n}.$$

Midpoint Estimator. Let $c_1 > 0$ be some sufficiently large universal constant (as elaborated in the proof of Lemma 3.4). For any $\delta \in (0,1)$, the midpoint estimator is defined by

$$\widehat{f}^{(\mathrm{mid})}_{\delta} = \mathrm{argmin}_{f \in \mathcal{G}_{\delta, c_1}(S_n)} R_n\left(\frac{\widehat{f}^{(\mathrm{ERM})} + f}{2}\right),$$

where $\widehat{f}^{(\mathrm{ERM})} = \widehat{f}^{(\mathrm{ERM})}(S_n)$ is any function in $\mathcal{G}$ that minimizes the empirical risk $R_n(\cdot)$ (induced by the sample S_n) and the set $\mathcal{G}_{\delta, c_1}(S_n)$ is a random (data-dependent) set of *almost empirical risk minimizers* defined by

$$\mathcal{G}_{\delta, c_1}(S_n) = \{g \in \mathcal{G} : R_n(g) \leq R_n(f^{(\mathrm{ERM})}) + c_1 C_b d_{\delta, n}(\widehat{f}^{(\mathrm{ERM})}, g)\}$$

with the empirical distance function $d_{\delta,n}$ given by, for any functions g, g':

$$d_{\delta,n}(g, g') = \sqrt{\frac{n^{-1} \sum_{i=1}^n (g(X_i) - g'(X_i))^2 \cdot \log(2m/\delta)}{n}} + \frac{b \log(2m/\delta)}{n}.$$

In the context of model selection aggregation, the idea of applying empirical risk minimization over some set preselected set of almost minimizers goes back to Lecué and Mendelson [112]. For the recent use of midpoint procedures in statistical literature, see, for example, [145, 34, 154].

Since $\widehat{f}^{(\mathrm{mid})}$ outputs 2-sparse convex combinations of elements of the dictionary $\mathcal{G}$, similarly to the above analysis of Audibert's star algorithm, it is enough to establish that $\widehat{f}^{(\mathrm{mid})}$ satisfies the offset condition. For the midpoint estimator, this fact is already implicit in the proofs of Puchkin and Zhivotovskiy [169] in the context of active learning. While, admittedly, the direct analysis of the midpoint estimator is no more difficult than the below lemma, for exposition purposes, let us demonstrate that $\widehat{f}^{(\mathrm{min})}$ does indeed satisfy the offset condition.

Lemma 3.4. *Fix any $\delta \in (0,1)$ and any distribution P supported on $\mathcal{X} \times [-b, b]$. In the setup described above, the estimator $\widehat{f}_{\delta}^{(mid)}$ satisfies the $(\mathcal{G}, \ell, \varepsilon, (64)^{-1}\gamma)$-offset condition for the distribution P, with $\varepsilon(\delta) \lesssim C_b^2 \gamma^{-1} \log(2m/\delta)/n$.*

The proof is deferred to Appendix 3.6.5. An immediate consequence of the above lemma, via an application of Theorem 3.1 (with $\delta_1 = \delta_2 = \delta/2$) and Corollary 3.2 is that for any $\delta \in (0,1)$ with probability at least $1 - \delta$ the following holds:

$$\mathcal{E}(\widehat{f}_{\delta}^{(\mathrm{mid})}, \mathcal{G}) \lesssim \gamma^{-1} C_b^2 \frac{\log(4m/\delta)}{n}.$$

3.5.2 Example Applications to Iterative Regularization

Establishing that the offset condition is satisfied by a family of regularization schemes, characterized by mirror descent optimization algorithms, is the subject of Chapter 4.

3.6 Proofs

This section contains the proofs of the main results.

3.6.1 Proof of Theorem 3.1

Recall that P denotes the underlying distribution of (X, Y) and let P_n denote its empirical counterpart supported on the sample S_n so that

$$P\ell = \mathbf{E}_{(X,Y)\sim P}[\ell(X,Y)] \text{ and } P_n\ell = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i, Y_i) \text{ for any function } \ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R};$$

$$Ph = \mathbf{E}_{X\sim P_X}[h(X)] \text{ and } P_n h = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \text{ for any function } h : \mathcal{X} \to \mathbb{R}.$$

With the above notation we have $R(f) = P\ell_f$ and $R_n(f) = P_n\ell_f$. Denote the event

$$E_{\delta_2} = \{P_n\ell_{\widehat{f}} - P_n\ell_{g^\star} \leq -\gamma P_n(\widehat{f} - g^\star)^2 + \varepsilon(\delta_2)\}$$

Since $\widehat{f}$ satisfies the $(\mathcal{G}, \ell, \varepsilon, \gamma)$-offset condition we have $\mathbf{P}(E_{\delta_2}) \geq 1 - \delta_2$; on E_{δ_2} we have

$$\begin{aligned} P\ell_{\widehat{f}} - P\ell_{g^\star} &= (P - P_n)(\ell_{\widehat{f}} - \ell_{g^\star}) + P_n(\ell_{\widehat{f}} - \ell_{g^\star}) \\ &\leq (P - P_n)(\ell_{\widehat{f}} - \ell_{g^\star}) - \gamma P_n(\widehat{f} - g^\star)^2 + \varepsilon(\delta_2) \\ &\leq \underbrace{\sup_{f \in \mathcal{F}} \left\{(P - P_n)(\ell_f - \ell_{g^\star}) - \gamma P_n(f - g^\star)^2\right\}}_{:= Z} + \varepsilon(\delta_2). \end{aligned}$$

The rest of the proof is structured as follows:

1. We first symmetrize a suitably rearranged Laplace transform of the empirical offset process Z. Since for $\lambda \geq 0$ the map $x \mapsto e^{\lambda x}$ is convex and non-decreasing, this step of the proof follows via standard arguments.

2. Next, we apply Talagrand's contraction lemma to the symmetrized offset empirical process. This step turns our process into a multiplier-type process of Proposition 3.1.

3. We conclude the proof via an application of Proposition 3.1, which yields a Bernstein-type upper bound on the moment generating function of the random variable $Z - \mathfrak{R}_n^{\text{off}}(\text{star}(\mathcal{H}), \gamma')$, for a suitably defined constant $\gamma' > 0$. The desired tail bound then follows via Markov's inequality.

Remark 3.2. Our proof strategy is inspired by the work of Lecué and Rigollet [115], where symmetrization and contraction arguments are also performed on the Laplace transform of the empirical process of interest. The contraction step is needed there to make the corresponding complexity measure linear in the model parameters so that the supremum over a convex hull is attained a vertex. In contrast, we need to apply the contraction step to put us in the setting of Proposition 3.1.

Symmetrization step. We begin by rewriting the random variable Z as follows:

$$\begin{aligned} Z &= \sup_{f \in \mathcal{F}} \left\{(P - P_n)\left(\ell_f - \ell_{g^\star}\right) - \gamma P_n(f - g^\star)^2\right\} \\ &= \sup_{f \in \mathcal{F}} \left\{(P - P_n)\left(\ell_f - \ell_{g^\star} + \frac{3\gamma}{4}(f - g^\star)^2\right) - \frac{\gamma}{4}P_n(f - g^\star)^2 - \frac{3\gamma}{4}P(f - g^\star)^2\right\} \end{aligned} \tag{3.5}$$

where in the last equation above we have added and subtracted $(3\gamma/4)P(f-g^\star)^2$. For any function $f \in \mathcal{F}$ introduce a shorthand notation

$$\phi_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \text{ such that } \phi_f(X,Y) = \ell_f(X,Y) - \ell_{g^\star}(X,Y) + \frac{3\gamma}{4}(f(X)-g^\star(X))^2.$$

Let $S_n' = (X_i', Y_i')_{i=1}^n$ denote an independent copy of $S_n = (X_i, Y_i)_{i=1}^n$ and denote $\mathbf{E}'$ as a shorthand notation for expectation computed with respect to S_n' only, conditionally on all other random variables. Let P_n' denote a counterpart to P_n with the sample S_n replaced by S_n'. Carrying on from equation (3.5) we can rewrite Z as follows:

$$\begin{aligned} Z &= \sup_{f\in\mathcal{F}} \left\{ (P-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{3\gamma}{4}P(f-g^\star)^2 \right\} \\ &= \sup_{f\in\mathcal{F}} \left\{ (P-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{\gamma}{4}P(f-g^\star)^2 - \frac{2\gamma}{4}P(f-g^\star)^2 \right\} \\ &= \sup_{f\in\mathcal{F}} \left\{ (\mathbf{E}'P_n'-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{\gamma}{4}\mathbf{E}'P_n'(f-g^\star)^2 - \frac{2\gamma}{4}P(f-g^\star)^2 \right\} \end{aligned} \tag{3.6}$$

Observe that in the above equation we have left the term $(2\gamma/4)P(f-g^\star)$ unchanged. This is needed to put us in the setting of Proposition 3.1, as we shall see below.

Let us now introduce a sequence of n independent Rademacher (symmetric and $\{\pm 1\}$ valued) random variables σ_i and let $\mathbf{E}_\sigma$ denote expectation with $\sigma_1, \ldots, \sigma_n$ only, conditionally on all other random variables. Let P_n^σ denote the symmetrized empirical measure so that for any function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and any function $h : \mathcal{X} \to \mathbb{R}$ we have

$$P_n^\sigma \ell = \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(X_i, Y_i) \quad \text{and} \quad P_n^\sigma h = \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i).$$

For $\lambda > 0$ the map $x \mapsto e^{\lambda x}$ is convex and non-decreasing; hence, for any $\lambda > 0$, using the identity (3.6), we can proceed to symmetrize the Laplace transform of Z as follows:

$$\begin{aligned} \mathbf{E}\exp(\lambda Z) &\le \mathbf{E}\mathbf{E}' \exp\left(\lambda \sup_{f\in\mathcal{F}} \left\{ (P_n'-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 \right.\right. \\ &\qquad \left.\left. - \frac{\gamma}{4}P_n'(f-g^\star)^2 - \frac{2\gamma}{4}P(f-g^\star)^2 \right\} \right) \\ &\le \mathbf{E}\mathbf{E}_\sigma \exp\left(2\lambda \sup_{f\in\mathcal{F}} \left\{ P_n^\sigma \phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{\gamma}{4}P(f-g^\star)^2 \right\} \right). \end{aligned} \tag{3.7}$$

Notice that the above moment generating function is almost of the form that can be bounded via Proposition 3.1. It remains to replace the term $P_n^\sigma \phi_f$ with a term $\rho P_n^\sigma(f-g^\star)$, for some constant ρ. This is the aim of the contraction step of this proof, which follows below.

Contraction step. Recall that by the assumptions of this theorem, there exists some constant C_b such that for any $f, f' \in \mathcal{F}, x \in \mathcal{X}, y \in \mathcal{Y}$ we have

$$|\ell_f(x,y) - \ell_{f'}(x,y)| \le C_b \, |f(x) - f'(x)| \, .$$

In particular, for any $f, f' \in \mathcal{F}$ and any $x \in \mathcal{X}, y \in \mathcal{Y}$ we have

$$\begin{aligned}
|\phi_f(x,y) - \phi_{f'}(x,y)| &= \left| \ell_f(x,y) + \frac{3\gamma}{4}(f(x) - g^\star(x))^2 - \ell_{f'}(x,y) - \frac{3\gamma}{4}(f'(x) - g^\star(x)) \right| \\
&\le C_b \, |f(x) - f'(x)| + \frac{3\gamma}{4} \, |(f(x) - f'(x))(f(x) + f'(x) - 2g^\star(x))| \\
&\le (C_b + 3\gamma b) \, |f(x) - f'(x)| \\
&= (C_b + 3\gamma b) \, |(f(x) - g^\star(x)) - (f'(x) - g^\star(x))| \, .
\end{aligned}$$

Hence, applying Talagrand's contraction inequality [118, Theorem 4.12] (conditionally on the sample S_n) with the set T_{S_n} and contraction mappings $\phi^{(i)}_{S_n}$:

$$\begin{aligned}
T_{S_n} &= \{((f - g^\star)(X_1), \dots, (f - g^\star)(X_n))^\mathsf{T} : f \in \mathcal{H}\}, \\
\phi^{(i)}_{S_n}(t_i) &= (2C_b + 6\gamma b)^{-1} \cdot 2 \left(\ell(t_i + g^\star(X_i), Y_i) - \ell_{g^\star}(X_i, Y_i) - \frac{3\gamma}{4} t_i^2 \right)
\end{aligned}$$

we may proceed upper bounding (3.7) as follows (cf. [115, Eq. (3.11)]):

$$\begin{aligned}
&\mathbf{E} \exp(\lambda Z) \\
&\le \mathbf{E}\mathbf{E}_\sigma \exp\left(\lambda \sup_{f \in \mathcal{F}} \left\{ P_n^\sigma 2\phi_f - \frac{\gamma}{2} P_n (f - g^\star)^2 - \frac{\gamma}{2} P (f - g^\star)^2 \right\} \right) \\
&\le \mathbf{E}\mathbf{E}_\sigma \exp\left(\lambda \sup_{f \in \mathcal{F}} \left\{ (2C_b + 6\gamma b) P_n^\sigma (f - g^\star) - \frac{\gamma}{2} P_n (f - g^\star)^2 - \frac{\gamma}{2} P (f - g^\star)^2 \right\} \right) \\
&= \mathbf{E}\mathbf{E}_\sigma \exp\left(\lambda \sup_{h \in \mathcal{H}} \left\{ (2C_b + 6\gamma b) P_n^\sigma h - \frac{\gamma}{2} P_n h^2 - \frac{\gamma}{2} P h^2 \right\} \right) \\
&\le \mathbf{E}\mathbf{E}_\sigma \exp\Bigg(\frac{\lambda}{n} \cdot \underbrace{n \sup_{h \in \mathrm{star}(\mathcal{H})} \left\{ (2C_b + 6\gamma b) P_n^\sigma h - \frac{\gamma}{2} P_n h^2 - \frac{\gamma}{2} P h^2 \right\}}_{:= U} \Bigg),
\end{aligned}$$

where in the penultimate line we introduced $\mathcal{H} = \{f - g^\star : f \in \mathcal{F}\}$, and in the last step the inequality comes from replacing $\mathcal{H}$ by $\mathrm{star}(\mathcal{H}) = \{\lambda h : h \in \mathcal{H}, \lambda \in [0,1]\}$.

We will now show that the random variable U is a supremum of an offset multiplier process satisfying the conditions of Proposition 3.1. Let $\zeta_i = (2C_b + 6\gamma b)\sigma_i$ and denote the distribution of ζ by P_ζ. Then, for any $h \in \mathcal{H}$ and for (X, ζ) distributed according

to the product distribution $P_X \otimes P_\zeta$, we have $\mathbf{E}[\zeta h(X)] = 0$. Therefore,

$$\begin{aligned} U &= n \cdot \sup_{h \in \mathrm{star}(\mathcal{H})} \left\{ (2C_b + 6\gamma b) P_n^\sigma h - \frac{\gamma}{2} P_n h^2 - \frac{\gamma}{2} P h^2 \right\} \\ &= \sup_{h \in \mathrm{star}(\mathcal{H})} \left\{ \sum_{i=1}^n \zeta_i h(X_i) - \mathbf{E}_{(X,\zeta) \sim P_X \otimes P_\zeta}[\zeta h(X)] - \frac{\gamma}{2} h(X_i)^2 - \frac{\gamma}{2} \mathbf{E}_{X \sim P_X} h(X)^2 \right\}. \end{aligned}$$

Hence, the moment generating function of the random variable U can be bounded via Proposition 3.1, taking $P_{(X,\zeta)} = P_X \otimes P_\zeta$.

Concluding the proof. Let $c_3 > 0$ be some universal constant such that

$$\eta = 8((2C_b + 6\gamma b)^2(\gamma/2)^{-1} + (\gamma/2)4b^2) \le c_3(\gamma^{-1}C_b^2 + bC_b + \gamma b^2).$$

Relabelling λ/n by λ and applying Proposition 3.1 to the random variable U, the following holds for any $\lambda \in (0, 1/\eta)$:

$$\log \mathbf{E} \exp\left(\lambda((nZ) - \mathbf{E}\mathbf{E}_\sigma U)\right) \le \log \mathbf{E}\mathbf{E}_\sigma \exp\left(\lambda(U - \mathbf{E}\mathbf{E}_\sigma U)\right) \le \frac{\lambda^2 \eta \mathbf{E}\mathbf{E}_\sigma U}{2(1 - \eta\lambda)}. \tag{3.8}$$

The desired tail bound now follows via standard arguments that we sketch below. By [31, Section 2.4], the upper bound (3.8) shows that the random variable $nZ - \mathbf{E}\mathbf{E}_\sigma U$ is sub-gamma on the right-tail with variance proxy $\eta \mathbf{E}\mathbf{E}_\sigma U$ and scale parameter η. Hence, via Markov's inequality, for any $\delta_1 \in (0, 1]$ we have

$$\mathbf{P}\left(nZ - \mathbf{E}\mathbf{E}_\sigma[U] \ge \sqrt{2\eta \mathbf{E}\mathbf{E}_\sigma[U] \log(\delta^{-1})} + \eta \log(\delta_1^{-1}) \right) \le \delta_1.$$

Subtracting $\mathbf{E}\mathbf{E}_\sigma[U]$ from both sides of the inequality defining the event inside $\mathbf{P}(\cdot)$ and optimizing the quadratic function in $\sqrt{\mathbf{E}\mathbf{E}_\sigma[U]}$, we deduce that

$$\begin{aligned} \delta_1 &\ge \mathbf{P}\left(nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \ge \sqrt{2\eta \mathbf{E}\mathbf{E}_\sigma[U] \log(\delta_1^{-1})} - \mathbf{E}\mathbf{E}_\sigma[U'] + \eta \log(\delta_1^{-1}) \right) \\ &\ge \mathbf{P}\left(nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \ge \sup_{x \in \mathbb{R}} \left\{ \sqrt{2\eta x \log(\delta^{-1})} - x^2 \right\} + \eta \log(\delta_1^{-1}) \right) \\ &= \mathbf{P}\left(nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \ge (3/2)\eta \log(\delta_1^{-1}) \right). \end{aligned}$$

Thus, denoting the event

$$E_{\delta_1} = \{ nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \le (3/2)\eta \log(\delta_1^{-1}) \}$$

we have $\mathbf{P}(E_{\delta_1}) \ge 1 - \delta_1$. Finally, observe that

$$\begin{aligned} \mathbf{E}_{S_n} \mathbf{E}_\sigma U &= n(2C_b + 6\gamma b) \mathfrak{R}_n^{\mathrm{off}}\left(P_X, \mathrm{star}(\mathcal{H}), \frac{\gamma}{2} \cdot (2C_b + 6\gamma b)^{-1} \right) \\ &\le 74 \cdot n(C_b + \gamma b) \mathfrak{R}_n^{\mathrm{off}}\left(P_X, \mathrm{star}(\mathcal{H}), \gamma \cdot (C_b + \gamma b)^{-1} \right). \end{aligned}$$

The desired result follows by the union bound on the events E_{δ_1} and E_{δ_2}. □

3.6.2 Proof of Proposition 3.1

Let us first discuss the key insight into our proof. Without loss of generality, assume that the supremum in the definition of the random variable U (cf. (3.2)) is always attained by some function, and denote this (random) function by $\tilde{h} = \tilde{h}(S_n)$. The following lemma shows that the empirical and population L_2 norms of $\tilde{h}$ are upper bounded by $c^{-1}U$. Thus, intuitively the supremum over $\mathrm{star}(\mathcal{H})$ in the multiplier process is computed over a "self-localized" (in a random/data-dependent way) subset of $\mathrm{star}(\mathcal{H})$. In contrast, we remark that the classical theory of localization via fixed-point equations proceeds by localizing the function class $\mathrm{star}(\mathcal{H})$ by constraining it to an *explicitly* chosen subset of functions with small L_2 population or empirical norms.

Lemma 3.5. *Consider the setting of Proposition 3.1 and let $\tilde{h} = \tilde{h}(S_n)$ denote a random function that attains the supremum of the offset multiplier process U (cf. (3.2)) given the sample $S_n = (X_i, \zeta_i)_{i=1}^n$. That is, $\tilde{h}$ satisfies*

$$\sum_{i=1}^{n} \Big(\zeta_i \tilde{h}(X_i) - \mathbf{E}[\zeta \tilde{h}(X) | S_n] - \gamma \tilde{h}(X_i)^2 - \gamma \mathbf{E}[\tilde{h}(X)^2 | S_n] \Big) = U(S_n).$$

Then, the following deterministic inequality holds for any realization of S_n:

$$\sum_{i=1}^{n} \Big(\mathbf{E}[\tilde{h}(X)^2 | S_n] + \tilde{h}(X_i)^2 \Big) \leq \frac{1}{\gamma} U(S_n).$$

Proof. Fix any realization $S_n = (X_i, \zeta_i)_{i=1}^n$ and in the rest of this proof we work conditionally on S_n. For any $h \in \mathrm{star}(\mathcal{H})$, define $A(h)$ and $B(h)$ as follows:

$$A(h) = \sum_{i=1}^{n} \left(\zeta_i h(X_i) - \mathbf{E}[\zeta h(X) | S_n] \right), \quad B(h) = \gamma \sum_{i=1} \Big(\mathbf{E}[h(X)^2 | S_n] + h(X_i)^2 \Big).$$

Thus, since $\tilde{h} = \tilde{h}(S_n)$ denotes a maximizer of the offset multiplier process, we have

$$A(\tilde{h}) - B(\tilde{h}) = \sup_{h \in \mathrm{star}(\mathcal{H})} (A(h) - B(h)) = U(S_n). \tag{3.9}$$

For any $\lambda \in [0, 1)$, let $\lambda h : x \mapsto \lambda h(x)$. Observe that for any h and λ, the term $A(\lambda h)$ scales *linearly* as a function of λ (i.e., $A(\lambda h) = \lambda A(h)$), while the term $B(\lambda h)$ scales *quadratically* (i.e., $B(\lambda h) = \lambda^2 B(h)$) as a function of λ. Fix any $\lambda \in [0, 1)$ and note that by the definition of star-hulls, the function $\lambda \tilde{h}$ is in the set $\mathrm{star}(\mathcal{H})$. Therefore, the identity (3.9) implies that

$$\lambda A(\tilde{h}) - \lambda^2 (B(\tilde{h})) = A(\lambda \tilde{h}) - B(\lambda \tilde{h}) \leq \sup_{h \in \mathrm{star}(\mathcal{H})} (A(h) - B(h)) = U(S_n). \tag{3.10}$$

Rearranging the identity (3.9) we also have $A(\tilde{h}) = U(S_n) + B(\tilde{h})$, which plugged into the left hand side of (3.10) yields

$$\lambda(1-\lambda)B(\tilde{h}) \leq (1-\lambda)U(S_n).$$

Dividing both sides by $(1-\lambda) > 0$ shows that $\lambda B(\tilde{h}) \leq U(S_n)$. Since the last equation holds for any $\lambda \in [0,1)$ it follows that $B(\tilde{h}) \leq U(S_n)$ which completes the proof of this lemma. □

With the above lemma in place, we are ready to prove Proposition 3.1. In the below proof, we follow the standard approach for obtaining Bernstein-type concentration bounds for the supremum of empirical processes (see [31, Section 12.2]). In particular, such bounds often build on the entropy method, which in our case appears through an application of the exponential Efron-Stein inequality. For a survey of tail bounds on the supremum of empirical processes, see the bibliographic remarks in [31, Section 12]. We now introduce some additional notation.

1. Let $S_n^{(i)}$ be equal to the sample S_n with the i-th element (X_i, ζ_i) replaced by an independent copy $(X_i', \zeta_i') \sim P_{(X,\zeta)}$.

2. For $i = 1, \ldots, n$, let $U_i' = U(S_n^{(i)}))$. Thus U_i' is the supremum of the offset multiplier process computed on the sample $S_n^{(i)}$, which differs from S_n by the i-th sample only.

3. Let $\mathbf{E}'[\cdot] = \mathbf{E}[\cdot|S_n]$ denote the expectation computed with respect to the random variables (X_i', ζ_i') only. In particular, we have $\mathbf{E}'[U] = U$.

The exponential Efron-Stein inequality [31, Theorem 6.16] asserts that for $\theta > 0$ and any $\lambda \in (0, 1/\theta)$ we have

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \leq \frac{\lambda\theta}{1-\lambda\theta} \log \mathbf{E} e^{\lambda V^+/\theta}, \quad \text{where} \quad V^+ = \sum_{i=1}^n \mathbf{E}'[(U - U_i')_+^2]. \tag{3.11}$$

To complete the proof of Proposition 3.1, it remains to upper bound the random variable V^+. This will be achieved via a combination of Lemma 3.5 and boundedness assumptions on the function class $\mathcal{H}$ and the multipliers ζ. Indeed, let $\tilde{h} = \tilde{h}(S_n)$ be a function that attains the supremum in the definition of U (cf. Lemma 3.5) Then, evaluating the multiplier process defined on the sample $S_n^{(i)}$ with the function $\tilde{h}$ yields a lower bound on U_i. Therefore, for $i = 1, \ldots, n$ we have

$$U - U_i' \leq \zeta_i \tilde{h}(X_i) - \gamma \tilde{h}(X_i)^2 - \zeta_i' \tilde{h}(X_i') + \gamma \tilde{h}(X_i')^2$$

and hence,

$$(U - U_i')_+^2 \le \left(\zeta_i \tilde{h}(X_i) - \gamma \tilde{h}(X_i)^2 - \zeta_i' \tilde{h}(X_i') + \gamma \tilde{h}(X_i')^2\right)^2.$$

Noting that for any $a, b, c, d \in \mathbb{R}$ we have $(a+b+c+d)^2 \le 4a^2 + 4b^2 + 4c^2 + 4d^2$ (for example, by the Cauchy-Schwarz inequality) it follows that

$$\begin{aligned}
\mathbf{E}'[(U - U_i')_+^2] &\le 4\mathbf{E}'[\zeta_i^2 \tilde{h}(X_i)^2 + \gamma^2 \tilde{h}(X_i)^4 + \zeta_i'^2 \tilde{h}(X_i')^2 + \gamma^2 \tilde{h}(X_i')^4] \\
&\le 4\mathbf{E}'[(\sigma^2 + \gamma^2\kappa^2)(\tilde{h}(X_i)^2 + \tilde{h}(X_i')^2)] \\
&\le 4(\sigma^2 + \gamma^2\kappa^2)(\tilde{h}(X_i)^2 + \mathbf{E}[\tilde{h}(X)^2 | S_n]),
\end{aligned}$$

where the second line follows by the boundedness assumptions and the last line follows by noting that $\tilde{h}(X_i)$ depends on S_n only and renaming X_i' to X. Hence, we can now obtain an upper bound on V^+ defined in (3.11) via Lemma 3.5 as follows:

$$0 \le V^+ \le 4(\sigma^2 + \gamma^2\kappa^2) \sum_{i=1}^n \left(\tilde{h}(X_i)^2 + \mathbf{E}[\tilde{h}(X)^2 | S_n]\right) \le 4(\sigma^2\gamma^{-1} + \gamma\kappa^2)U$$

Plugging the above upper bound on V^+ into the exponential Efron-Stein inequality (3.11) with the choice $\theta = 4(\sigma^2\gamma^{-1} + \gamma\kappa^2)$ yields, for any $\lambda \in (0, 1/\theta)$:

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \le \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbf{E} e^{\lambda U} = \frac{\lambda\theta}{1 - \lambda\theta} \left(\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} + \lambda \mathbf{E}U\right).$$

Rearranging the above inequality, we obtain

$$\frac{1 - 2\lambda\theta}{1 - \lambda\theta} \log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \le \frac{\lambda^2\theta \mathbf{E}U}{1 - \lambda\theta}.$$

For any $\lambda \in (0, 1/(2\theta))$ we have $(1 - 2\lambda\theta)/(1 - \lambda\theta) > 0$, thus for $\lambda \in (0, 1/(2\theta))$ we have

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \le \frac{\lambda^2\theta \mathbf{E}[U]}{1 - 2\lambda\theta} = \frac{\lambda^2(\eta \mathbf{E}U)}{2(1 - \eta\lambda)},$$

where $\eta = 2\theta$. This finishes our proof. □

3.6.3 Proof of Lemma 3.1

Fix any $\varepsilon > 0$ and let $\lambda = (1 + \varepsilon)^{-1} \in (0, 1)$. Let $\lambda\mathcal{H} = \{\lambda h : h \in \mathcal{H}\}$ and observe that by the star-shapedness assumption we have $\lambda\mathcal{H} \subseteq \mathcal{H}$. It follows that

$$\mathfrak{R}_n^{\text{off}}(P_X, \mathcal{H}, \gamma) = \lambda^{-1}\mathfrak{R}_n^{\text{off}}(P_X, \lambda\mathcal{H}, \lambda^{-1}\gamma) \le \lambda^{-1}\mathfrak{R}_n^{\text{off}}(P_X, \mathcal{H}, \lambda^{-1}\gamma). \tag{3.12}$$

We now proceed via a peeling argument. For any $r_1 \ge 0, r_2 > 0$ denote $\mathcal{H}(r_1, r_2) = \{h \in \mathcal{H} : \mathbf{E}_{X \sim P_X}[h(X)^2] \in [r_1, r_2]\}$. Denote $\mathfrak{R}_n^{\text{loc}} = \mathfrak{R}_n^{\text{loc}}(P_X, \mathcal{H}, \gamma)$. Let $\mathcal{H}_0 = \mathcal{H}(0, \gamma^{-1}\mathfrak{R}_n^{\text{loc}})$

and for $k = 1, 2, \ldots$, let $\mathcal{H}_k = \mathcal{H}(\lambda^{1-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}, \lambda^{-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}) \cup \{h_0\}$, where h_0 denotes the identically zero function. Since $\mathcal{H} = \cup_{k\geq 0}\mathcal{H}_k$, by (3.12) we have

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma) \leq \lambda^{-1} \sum_{k\geq 0} \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, \lambda^{-1}\gamma). \tag{3.13}$$

Observe that by the definition of $\mathfrak{R}_n^{\mathrm{loc}}$ (cf. Definition 2.5) we have

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_0, \lambda^{-1}\gamma) \leq \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_0, 0) \leq \mathfrak{R}_n^{\mathrm{loc}}.$$

At the same time, for any $k \geq 1$ we have $h_0 \in \mathcal{H}$ and hence $\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, \lambda^{-1}\gamma) \geq 0$. Also, by [19, Lemmas 3.2 and 3.4] we have $\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}(0, \lambda^{-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}), 0) \leq \lambda^{-k}\mathfrak{R}_n^{\mathrm{loc}}$ and consequently

$$\begin{aligned} 0 &\leq \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, \lambda^{-1}\gamma) \leq \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, 0) - \lambda^{-1}\gamma \cdot \lambda^{1-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}} \\ &= \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, 0) - \lambda^{-k}\mathfrak{R}_n^{\mathrm{loc}} \leq \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}(0, \lambda^{-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}), 0) - \lambda^{-k}\mathfrak{R}_n^{\mathrm{loc}} \leq 0. \end{aligned}$$

Hence, using the two display equations above, the inequality (3.13) simplifies to

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma) \leq \lambda^{-1}\mathfrak{R}_n^{\mathrm{loc}} = (1+\varepsilon)\mathfrak{R}_n^{\mathrm{loc}}.$$

Since the choice of $\varepsilon > 0$ is arbitrary, our proof is complete. □

3.6.4 Proof of Lemma 3.3

Let $\Phi \in \mathbb{R}^{n\times d}$ denote a matrix such that $\Phi_{i,j} = (\Phi_i)_j$ for any $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$. To simplify the notation let $\mathcal{F} = \mathcal{F}_{\mathrm{lin}}^{d,k}$. For any $S \subseteq \{1, 2, \ldots, d\}$, let $\Phi_S \in \mathbb{R}^{n\times|S|}$ denote the matrix obtained by keeping only the columns of Φ indexed by the set S and let

$$\mathcal{S}^{d,k} = \{S \subseteq \{1, \ldots, d\} : |S| \leq k\}.$$

Observe that for any $\lambda > 0$ by Jensen's inequality, the fact that $x \mapsto e^{\lambda x}$ is increasing, and replacing maximum by a sum, we have

$$\begin{aligned}
&n\mathfrak{R}^{\text{off}}(S_n^{\Phi}, \mathcal{F}, \gamma) \\
&= \mathbf{E}_\sigma \sup_{\langle w,\cdot\rangle \in \mathcal{F}} \left\{ \sum_{i=1}^n \sigma_i \langle w, \Phi_i \rangle - \gamma \langle w, \Phi_i \rangle^2 \right\} \\
&= \mathbf{E}_\sigma \sup_{\langle w,\cdot\rangle \in \mathcal{F}} \left\{ \langle \Phi w, \sigma \rangle - \gamma w^\top (\Phi^\top \Phi) w \right\} \\
&= \mathbf{E}_\sigma \max_{S \in \mathcal{S}^{d,k}} \sup_{w \in \mathbb{R}^{|S|}} \left\{ \langle \Phi_S w, \sigma \rangle - \gamma w^\top (\Phi_S^\top \Phi_S) w \right\} \\
&\leq \frac{1}{\lambda} \log \mathbf{E}_\sigma \exp \left(\lambda \max_{S \in \mathcal{S}^{d,k}} \sup_{w \in \mathbb{R}^{|S|}} \left\{ \langle \Phi_S w, \sigma \rangle - \gamma w^\top (\Phi_S^\top \Phi_S) w \right\} \right) \\
&\leq \frac{1}{\lambda} \log \sum_{S \in \mathcal{S}^{d,k}} \mathbf{E}_\sigma \exp \left(\lambda \sup_{w \in \mathbb{R}^{|S|}} \left\{ \langle \Phi_S w, \sigma \rangle - \gamma w^\top (\Phi_S^\top \Phi_S) w \right\} \right) \\
&\leq \frac{1}{\lambda} \log \left(\left| \mathcal{S}^{d,k} \right| \max_{S \in \mathcal{S}^{d,k}} \mathbf{E}_\sigma \exp \left(\lambda \sup_{w \in \mathbb{R}^{|S|}} \left\{ \langle \Phi_S w, \sigma \rangle - \gamma w^\top (\Phi_S^\top \Phi_S) w \right\} \right) \right). \quad (3.14)
\end{aligned}$$

We now proceed to upper bound the expectation inside the logarithm. For any matrix A, denote its Moore-Penrose inverse by $A^\dagger$. Fix any $S \in \mathcal{S}^{d,k}$. For any vector $\sigma \in \mathbb{R}^n$, the vector $\Phi_S^\top \sigma$ belongs to the orthogonal complement of the null space of $\Phi_S^\top \Phi_S$. Hence, following [177, Section 12, page 108], the following identity holds:

$$\begin{aligned}
\sup_{w \in \mathbb{R}^{|S|}} \left\{ \langle \Phi_S w, \sigma \rangle - \gamma w^\top (\Phi_S^\top \Phi_S) w \right\} &= \sup_{w \in \mathbb{R}^{|S|}} \left\{ \left\langle w, \Phi_S^\top \sigma \right\rangle - \gamma w^\top (\Phi_S^\top \Phi_S) w \right\} \\
&= (4\gamma)^{-1} \sigma^\top \Phi_S (\Phi_S^\top \Phi_S)^\dagger \Phi_S^\top \sigma.
\end{aligned}$$

To simplify the notation, denote by $H = \Phi_S (\Phi_S^\top \Phi_S)^\dagger \Phi_S^\top$ the hat matrix, keeping the dependence on an arbitrary fixed $S \in \mathcal{S}^{d,k}$ implicit. By the above equation, it follows that

$$\mathbf{E}_\sigma \exp \left(\lambda \sup_{w \in \mathbb{R}^{|S|}} \left\{ \langle \Phi_S w, \sigma \rangle - \gamma w^\top (\Phi_S^\top \Phi_S) w \right\} \right) = \mathbf{E}_\sigma \exp \left(\frac{\lambda}{4\gamma} \sum_{i,j=1}^n \sigma_i \sigma_j H_{i,j} \right).$$

We will now control the moment generating function of the above Rademacher chaos by decoupling and comparison with Gaussian chaos. Let $\sigma' = (\sigma'_1, \ldots, \sigma'_n)^\top$ be an independent copy of σ. Let $g = (g_1, \ldots, g_n)^\top \in \mathbb{R}^n$ be a vector of independent standard Normal random variables and let g' be an independent copy of g. Then, for some

universal constant $c_1 > 0$ we have

$$\begin{aligned}
&\mathbf{E}_\sigma \exp\left(\frac{\lambda}{4\gamma}\sum_{i,j=1}^n \sigma_i\sigma_j H_{i,j}\right) \\
&\leq \mathbf{E}_{\sigma,\sigma'} \exp\left(\frac{\lambda}{\gamma}\sum_{i,j=1}^n \sigma_i\sigma'_j H_{i,j}\right) && \text{[212, (Decoupling) Theorem 6.1.1]} \\
&\leq \mathbf{E}_{g,g'} \exp\left(\frac{c_1\lambda}{\gamma}\sum_{i,j=1}^n g_i g'_j H_{i,j}\right) && \text{[212, (Comparison) Lemma 6.2.3].}
\end{aligned}$$

Let $\|\cdot\|_{\text{op}}$ denote the operator norm and let $\|\cdot\|_F$ denote the Frobenius norm. Then, by the Gaussian chaos moment generating function bound [212, Lemma 6.2.2], there exist some universal constants $c_2, c_3 > 0$ such that for any $\lambda \in (0, \gamma c_2/\|H\|_{\text{op}}]$ we have

$$\mathbf{E}_{g,g'} \exp\left(\frac{c_1\lambda}{\gamma}\sum_{i,j=1}^n g_i g'_j H_{i,j}\right) \leq \exp\left(\frac{c_3\lambda^2}{\gamma^2}\|H\|_F^2\right).$$

We will now plug in the above bound into (3.14). Notice that the hat matrix H has at most $|S|$ non-zero eigenvalues, all of which are equal to 1; hence,$\|H\|_{op} = 1$ and $\|H\|_F^2 \leq |S|$. It follows that for any $\lambda \in (0, \gamma c_2]$ we have

$$\mathbf{E}_\sigma \sup_{w\in\mathbb{R}^d, \|w\|_0\leq k}\left\{\langle \Phi w, \sigma\rangle - \gamma w^\top(\Phi^\top\Phi)w\right\} \leq \frac{1}{\lambda}\log|\mathcal{S}^{d,k}| + \frac{c_3\lambda k}{\gamma^2}. \tag{3.15}$$

Recalling the standard bound

$$|\mathcal{S}^{d,k}| = \sum_{i=1}^k \binom{d}{i} \leq \left(\frac{ed}{k}\right)^k$$

and plugging in $\lambda = \gamma c_2$ in (3.15) yields the desired result

$$n\mathfrak{R}^{\text{off}}(S_n^\Phi, \mathcal{F}, \gamma) \leq \frac{1}{\gamma}\left(c_2^{-1}k\log\frac{ed}{k} + c_2c_3k\right) \lesssim \frac{1}{\gamma}\log\left(\frac{ed}{k}\right)k.$$

□

3.6.5 Proof of Lemma 3.4

For any $g, g' \in \mathcal{G}$ define the event

$$E(g,g') = \left\{R(g) - R(g') \leq R_n(g) - R_n(g') + c_1C_b d_{\delta,n}(g,g')\right\}.$$

By the empirical Bernstein inequality [136, Thereom 11] applied to the random variables $(2bC_b)^{-1}(\ell_g(X_i,Y_i) - \ell_{g'}(X_i,Y_i))$ we have $\mathbf{P}(E(g,g')) \geq 1 - \delta/m^2$. Hence, defining the event $E = \cup_{g,g'\in\mathcal{G}}E(g,g')$, by the union bound $\mathbf{P}(E) \geq 1 - \delta$.

We will now show that on the event E, the estimator $\widehat{f}^{(\mathrm{mid})}$ satisfies the offset condition. First observe that on the event $E(\widehat{f}^{(\mathrm{ERM})}, g^\star) \subseteq E$, the population risk minimizer $g^\star$ belongs to the set $\mathcal{G}_{\delta,c_1}(S_n)$ of the empirical almost minimizers. Define the diameter

$$D_n^{\max} = \max_{g,g' \in \mathcal{G}_{\delta,c_1}(S_n)} \|g - g'\|_n^2, \quad \text{where} \quad \|g - g'\|_n^2 = \frac{1}{n}\sum_{i=1}^n (g(X_i) - g'(X_i))^2.$$

We may assume without loss of generality that $D_n^{\max} > 0$ since otherwise the offset condition is trivially satisfied. Since $g^\star \in \mathcal{G}_{\delta,c_1}(S_n)$, it follows that $\|\widehat{f}^{(\mathrm{mid})} - g^\star\|_n^2 \leq D_n^{\max}$. Also, since $D_n^{\max} > 0$, there exists some function $g' \in \mathcal{G}_{\delta,c_1}(S_n)$ such that $\|\widehat{f}^{(\mathrm{ERM})} - g'\| \geq D_n^{\max}/4$. Hence, on the event E it holds that

$$\begin{aligned}
&R_n(\widehat{f}^{(\mathrm{mid})}) - R_n(g^\star) \\
&\leq R_n\left(\frac{\widehat{f}^{(\mathrm{ERM})} + g'}{2}\right) - R_n(g^\star) \\
&\leq \frac{1}{2}(R_n(\widehat{f}^{(\mathrm{ERM})}) - R_n(g^\star)) + \frac{1}{2}(R_n(g') - R_n(g^\star)) - \frac{\gamma}{32} D_n^{\max}, \\
&\leq \left(\frac{1}{2} c_1 C_b \sqrt{\frac{D_n^{\max} \log(2m/\delta)}{n}} - \frac{\gamma}{64} D_n^{\max}\right) + \frac{1}{2} c_1 b C_b \frac{\log(2m/\delta)}{n} - \frac{\gamma}{64} D_n^{\max}, \\
&\leq \left(4 c_1^2 C_b^2 \gamma^{-1} + \frac{1}{2} c_1 b C_b\right) \frac{\log(2m/\delta)}{n} - \frac{\gamma}{64} \|f^{(\mathrm{mid})} - g^\star\|_n^2,
\end{aligned}$$

where the third line follows by the strong convexity of the loss function; the fourth line follows by the fact that $g' \in \mathcal{G}_{\delta,c_1}(S_n)$ and $R_n(\widehat{f}^{(\mathrm{ERM})}) - R_n(g^\star) \leq 0$; the fifth line follows by optimizing the quadratic function in $\sqrt{D_n^{\max}}$ in the brackets and replacing $D_n^{\max}$ by $\|\widehat{f}^{(\mathrm{mid})} - g^\star\|_n^2$. By Remark 3.1, we have $bC_b \leq \gamma^{-1} C_b^2$ and thus our proof is complete. □

3.7 Limitations and Open Directions

One limitation of the framework presented in this chapter is naturally the imposed boundedness assumptions, which play a crucial role in establishing our main concentration result (Proposition 3.1). Hence, we cannot analyze unbounded, and in particular, heavy-tailed problems that have recently received a lot of attention; see the survey by Lugosi and Mendelson [124]. For progress in this direction, we refer to the works by Mendelson [142, 144] and Oliveira [163], where one-sided concentration arguments and moment-equivalence assumptions play a central role. It would be interesting to see if the offset Rademacher complexity framework could be extended along similar lines.

However, even staying within the bounded framework, there exist improper statistical estimators that do not fully fit within the theory developed in this chapter. Let us discuss a specific example, namely, the Q-aggregation algorithm [58, 115], which is one of the existing deviation-optimal algorithms for model selection aggregation. For simplicity, let $\ell(y_1, y_2) = (y_1 - y_2)^2$ denote the quadratic loss, and let R_n be the associated empirical risk functional. Let $\mathcal{G} = \{g_1, \dots, g_m\}$ be a finite dictionary of reference functions and let $g^\star$ be any element of the dictionary minimizing the population risk. Let $\Delta_m = \{\theta \in \mathbb{R}^m : \theta_i \geq 0, \sum_{i=1}^m \theta_i = 1\}$ be the m-dimensional probability simplex and consider the class of functions $\mathcal{F} = \{f_\theta = \sum_{i=1}^m \theta_i g_i : \theta \in \Delta_m\}$. Fix some $\nu \in (0, 1)$ and define the Q-aggregation estimator as follows:

$$\widehat{f}_\nu^{(\text{Q-agg})} = \widehat{f}_{\widehat{\theta}}, \text{ where } \widehat{\theta} = \operatorname{argmin}_{\theta \in \mathbb{R}^m} R_n(f_\theta) + \nu V_n(f_\theta), \tag{3.16}$$

where

$$V_n(f_\theta) = \sum_{i=1}^m \theta_i \|g_i - f_\theta\|_n^2.$$

The definition of the Q-aggregation objective is motivated by the following identity, which holds for any $\theta \in \Delta_m$:

$$R_n(f_\theta) = \|f_\theta - Y\|_n^2 = \sum_{i=1}^m \theta_i \|g_i - Y\|_n^2 - V_n(f_\theta).$$

Thus, the parameter $\nu \in [0, 1]$ in the Q-aggregation objective (3.16) allows to interpolate between the empirical risk minimization estimator over the simplex (for $\nu = 0$), and, at the other extreme, an estimator that minimizes the linearized loss (for $\nu = 1$). Plugging in $\nu = \frac{1}{3}$, and denoting $f_{1/3}^{(\text{Q-agg})} = f_{\widehat{\theta}}$, it is possible to show that (calculations omitted) the following deterministic inequality holds:

$$R_n(f_{\widehat{\theta}}) - R_n(g^\star) \leq -\frac{1}{3}\|f_{\widehat{\theta}} - g^\star\| - \frac{1}{3}\sum_{i=1}^m \widehat{\theta}_i \|f_i - g^\star\|_n^2.$$

The first term on the right-hand side is the term appearing in the definition of the offset condition (cf Definition 3.1). However, the second term plays an additional role in regularizing the resulting "offset-type" empirical process. Indeed, provided that Proposition 3.1 can be extended to the offset terms of the type appearing in the above display equation, the following complexity measure (up to modifying absolute constants) would upper bound the excess risk of the Q-aggregation estimator:

$$\begin{aligned}
&\mathbf{E}_{\sigma,X}\left[\frac{1}{n}\sup_{\theta\in\Delta_m}\left\{\sum_{i=1}^n \sigma_i(f_\theta - g^\star)(X_i) - \frac{1}{3}\|f_\theta - g^\star\|_n^2 - \frac{1}{3}\sum_{i=1}^n \theta_i \|f_i - g^\star\|_n^2\right\}\right] \\
&\leq \mathbf{E}_{\sigma,X}\left[\frac{1}{n}\sup_{\theta\in\Delta_m}\left\{\sum_{i=1}^n \sigma_i(f_\theta - g^\star)(X_i) - \frac{1}{3}\sum_{i=1}^n \theta_i \|f_i - g^\star\|_n^2\right\}\right].
\end{aligned}$$

The key idea of Q-aggregation is that the above supremum, computed over the simplex Δ_m, is a linear function of θ. Hence, the above supremum is attained at one of the m vertices of the simplex Δ_m. The optimal upper bound of order $\log(m)/n$ on the above complexity measure can then be obtained via standard arguments, namely, applying log and exp composition inside the expectation, taking the log outside of the expectation by Jensen's inequality, and approximating the maximum by a finite sum involving m terms inside the expectation.

Thus, motivated by the Q-aggregation example discussed above, one avenue for extending the results of this chapter is to try to generalize the concentration of shifted multiplier-type processes (Proposition 3.1) to more general offset terms than the one suggested in the offset condition definition (Definition 3.1). Indeed, the only key property exploited in the proof of Proposition 3.1 is that the offset term grows faster than the terms involving Rademacher variables. Extending Proposition 3.1 to more general offset terms would allow bounding the excess risk for a more general class of statistical estimators. In addition, it would provide a way to design statistical estimators that explicitly regularize the resulting offset-type complexity measure, exploiting the particular problem structure.

3.8 Bibliographic Remarks

The results of this chapter are based on work performed in collaboration with Varun Kanade and Patrick Rebeschini, currently in preparation for submission.

The idea of symmetrizing empirical processes to obtain excess risk bounds in Statistical Learning dates back to the early works by Vapnik and Chervonenkis [205, 208, 206]. For some results involving global Rademacher complexities, see, for example, [18, 96, 151]. However, it is well-known that computing global Rademacher averages cannot capture excess risk convergence rates faster than $1/\sqrt{n}$, where n is the sample size (see, e.g., [16, Theorem 2.3]). In contrast, local Rademacher complexities, developed by many authors (e.g., [107, 104, 18, 130, 19, 105]), and culminating in the seminal works by Bartlett, Bousquet, and Mendelson [19], Koltchinskii [105], yield sharp excess risk rates (at least in terms of the sample size) for the empirical risk minimization algorithm in many problem settings. See also the survey by Boucheron, Bousquet, and Lugosi [30] and the textbooks by Koltchinskii [106] and Wainwright [216, Chapters 13 and 14]. The localization approach via offset Rademacher complexities, which the present chapter extends to a high probability framework, is due to Rakhlin and Sridharan [170], Liang, Rakhlin, and Sridharan [120].

The above-cited works on local Rademacher complexities primarily consider bounded problem settings. Boundedness is needed for two technical reasons. First, it allows us to apply Talagrand's contraction lemma [117, Chapter 4] for the symmetrized expected empirical process. It is discussed by Mendelson [142] that the contraction step prevents obtaining bounds that scale correctly with the noise level of the problem; we shall return to this point in Chapter 5 of this thesis. The second reason for boundedness is due to the application of Talagrand's concentration inequality for empirical processes [191, 192]. As a short digression from the topic of excess risk bounds, we remark that Talagrand's inequality for empirical processes is a topic of independent interest, with various refinements and extensions available in the literature, e.g., [117, 131, 32, 99, 140, 116]. In the context of Statistical Learning, however, there is an issue related to Talagrand's inequality: it is too strong in the sense that it provides a two-sided concentration result while bounding the excess risk only requires one-sided concentration, which can be obtained without boundedness assumptions. Regarding the last point, we point the interested reader to the works of Mendelson [142] and Oliveira [163], where one-sided concentration arguments are established for unbounded problems, namely, under moment-equivalence-type assumptions on the underlying data-generating mechanism. As a result, the above line of work provides powerful tools for treating many unbounded and potentially heavy-tailed problems of interest that fall outside the present chapter's scope. However, let us remark that moment-equivalence assumptions do not allow for immediate distribution-free treatment of bounded settings at the level of generality offered via the classical local complexities; see the paper by Saumard [184] as well as the discussions in Chapters 5 and 6 of this thesis.

To better contextualize the results obtained in this chapter, let us briefly review some of the other general frameworks for obtaining excess risk bounds. One of the simplest ways to obtain sharp bounds on the *expected* excess risk without imposing strong distributional assumptions is via *average stability* (or *leave-one-out*) arguments [178, 59, 82]. Among other approaches are in-expectation guarantees obtainable via stochastic approximation arguments (e.g., [176, 217, 156, 64]), or by transporting regret bounds from the framework of prediction of individual sequences [47] to the stochastic setting via an online-to-batch conversion (e.g., [48, 7]). However, in this chapter, we focused on obtaining bounds that hold with high probability. As already discussed in the introduction, this is of particular interest in improper learning settings since, in such regimes, in-expectation bounds do not imply any meaningful deviation

bounds on the excess risk. Regarding the last point, see also the discussions in Chapter 6 of this thesis.

Turning to excess risk bounds that hold with high probability, along with the already discussed local Rademacher complexity approach, another powerful framework pioneered by McAllester [137] is the so-called PAC-Bayesian approach; see also the monograph by Catoni [43]. This approach is particularly well-suited to exploit specific problem structures via a properly chosen prior distribution. For some sharp results obtained over the past decade particularly tailored to linear problem structures, see [8, 9, 45, 152, 227]. Recently, some high-probability guarantees were also established for the class of *uniformly stable* algorithms (which is a more restrictive notion than average stability mentioned above; see [33]). Specifically, "fast rate" excess risk bounds that hold with high probability were recently obtained by Klochkov and Zhivotovskiy [102]. Also, for "fast rate" high-probability excess risk bounds obtainable via online-to-batch conversions, see the work by Kakade and Tewari [95] and the references therein. In terms of probabilistic tools, the former work builds on the notion of (weakly) self-bounding functions [29, 135], while the latter relies on the tail bound for martingales due to Freedman [68]. However, both works cited above impose strong assumptions on the loss function not present in this chapter, excluding classical settings of interest such as bounded regression with the squared loss.

Let us conclude this section by highlighting one difference between the offset and Bernstein conditions. In some settings, the latter condition is used as a *distributional assumption* (in the sense that only some of the distributions with the given support satisfy the condition, as opposed to all distributions). For example, in the classification setting with zero-one loss, the Bernstein condition corresponds to bounded noise assumptions (see the discussions in [30]), under which empirical risk minimization estimator can achieve fast rates of convergence of the excess risk despite the lack of curvature in the zero-one loss function. For sharp treatment of the classification setting under the bounded noise assumptions via ideas related to offset Rademacher averages, see [228]. At the same time, let us remark that the offset condition can be exploited to design prediction procedures that achieve fast classification rates without Bernstein-type bounded noise assumptions, provided an option to abstain from prediction at a smaller cost than misclassification [34, 160, 169].

4 Iterative Regularization via Early-Stopped Mirror Descent

This chapter is based on results obtained in collaboration with Varun Kanade and Patrick Rebeschini published in the paper [211].

In this chapter, we study iterative regularization schemes characterized by the family of mirror descent algorithms. Such schemes generate a sequence of models by applying an optimization algorithm to the *unregularized* empirical risk. The regularization effect is induced by early termination of the optimization procedure in an attempt to stabilize the obtained solution. An attractive feature of iterative regularization schemes is that each model – a point on the optimization path – is computationally relatively cheap to obtain, in the sense to be explained in the introduction. The main contribution of this chapter is to connect the analysis of iterative regularization schemes to the statistical notion of offset Rademacher complexity.

The material presented in this chapter presupposes familiarity with the mathematical framework of Statistical Learning (see Section 2.1), local Rademacher complexity measures (see Section 2.5 and Chapter 3), and the mirror descent algorithm (see Section 2.4). For a summary of notation, see Section 2.6.

4.1 Introduction

Among the most studied statistical estimators is the *empirical risk minimization* (ERM) algorithm, which, given a reference class of predictors $\mathcal{G}$, outputs a function $\widehat{g}^{(\mathrm{ERM})}$ defined as

$$\widehat{g}^{(\mathrm{ERM})} \in \arg\min_{g \in \mathcal{G}} R_n(g). \tag{4.1}$$

Recall that R_n denotes the empirical risk functional defined by

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(X_i), Y_i),$$

where $S_n = (X_i, Y_i)_{i=1}^n$ denotes an i.i.d. sample drawn from some unknown distribution P such that $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, the class $\mathcal{G}$ contains functions mapping $\mathcal{X}$ to $\mathcal{Y}$, and $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is the loss function.

Recall that the success of a prediction procedure $\widehat{f} = \widehat{f}(S_n)$, given a reference class of functions $\mathcal{G}$, is measured by a random variable called the *excess risk*:

$$\mathcal{E}(\widehat{f}, \mathcal{G}) = R(\widehat{f}) - \inf_{g \in \mathcal{G}} R(g),$$

where $R(g) = \mathbf{E}_{(X,Y)\sim P}[\ell(g(X), Y)]$ denotes the *population risk*. It is well-known that, in general, the ERM estimator $\widehat{g}^{(\mathrm{ERM})}$ suffers from overfitting – a phenomenon where peculiarities of the observed data are fit too strongly, hindering the generalization ability of the output predictor and resulting in suboptimal levels of its excess risk.

One of the most widely studied ways to mitigate overfitting effects is replacing the $\widehat{g}^{(\mathrm{ERM})}$ estimator (4.1) with its regularized counterpart

$$\widehat{g}_{\lambda}^{(\mathrm{ERM})} \in \arg\min_{g\in\mathcal{G}} \{R_n(g) + \lambda\Psi(g)\}, \tag{4.2}$$

where the function Ψ penalizes prediction functions $g \in \mathcal{G}$ based on some notion of their complexity, while the parameter λ controls the trade-off between data fit and regularization. For example, if the reference class of functions $\mathcal{G}$ is the class of all m-dimensional linear functions $\{g_\alpha = \langle w, \cdot\rangle : \alpha \in \mathbb{R}^m\}$, popular choices of penalty functions are $\Psi(g_\alpha) = \|\alpha\|_1$ and $\Psi(g_\alpha) = \|\alpha\|_2^2$. The above two choices of Ψ result in procedures known by the names (in Statistics literature) of the *lasso* [52, 194, 51] and *ridge regression* [86], respectively.

Traditionally, in Learning Theory, statistical and computational properties of ERM estimators have been considered separately. From a statistical point of view, localized Rademacher complexity measures have become a default tool in statistical learning theory and empirical processes theory for controlling the excess risk of algorithms based on the empirical risk minimization principle [19, 105]. A rich and general theory regarding these complexity measures has been developed and used to provide excess risk bounds in both classification and regression settings, yielding minimax-optimal results in several cases. Such complexity measures depend on combinatorial or geometric parameters of interest, such as the VC-dimension or eigenvalue decay of the kernel matrix and, in particular, they may serve as a guiding principle to choose a suitable *explicit regularizer* for inducing a set of candidate models $(\widehat{g}_\lambda)_{\lambda\in\Lambda}$, where $\lambda \in \Lambda$ is a hyper-parameter that controls the amount of regularization. In practice, some $\lambda^\star \in \Lambda$ is then chosen via some model selection procedure such as cross-validation, aiming to select a model with the smallest risk. From a computational point of view, computing the estimators $(\widehat{g}_\lambda)_{\lambda\in\Lambda}$ can be done by solving the corresponding optimization problems defined in Equation (4.2), one for each $\lambda \in \Lambda$. An appealing aspect of this approach is that the design and analysis of efficient optimization algorithms for approximating solutions to the objective (4.2), exploiting the problem geometry that arises from the structure of the model as well as the data generating distribution P, can be done independently of the statistical analysis of the problem.

Recent years have also witnessed an increased interest in directly studying the statistical properties of models trained by gradient-based methods, particularly in relation to the notions of *implicit regularization* and *early stopping*. For a family of functions $\mathcal{F} = \{f_\alpha : \alpha \in \mathbb{R}^m\}$ parametrized by a vector α, such methods are fully characterized by the initialization point α_0 and an update rule, which given α_t and the gradient of the empirical risk at α_t, generates the next iterate α_{t+1}, yielding a set of candidate estimators $(\widehat{f}_{\alpha_t})_{t\geq 0}$. Early stopping has an effect akin to *explicit* regularization discussed above, and the *stopping time* $t^\star$ can be chosen in practice via cross-validation, just as in the case of choosing the explicit regularization parameter $\lambda^\star$ corresponding to the best model among $(\widehat{g}_\lambda)_{\lambda\in\Lambda}$. In modern large-scale machine learning applications, early stopping may often be the preferred way to perform model selection, since obtaining a new model is as cheap as performing a step of gradient descent, as opposed to solving a new optimization problem with a different regularization parameter.

It is by now well understood that changing the update rule that generates the sequence $(\widehat{g}_{\alpha_t})_{t\geq 0}$, e.g., by changing the optimization algorithm or parametrization of the model class, can directly affect both the statistical properties of the iterates $\widehat{g}_{\alpha_t}$, as well as computational properties, such as an upper-bound on the optimal stopping time $t^\star$ (see Section 4.1.1 for an extended discussion). However, most of the prior literature has focused on the investigation of vanilla gradient descent updates: $\alpha_{t+1} = \alpha_t - \eta\nabla_{\alpha_t} R_n(\widehat{g}_{\alpha_t})$; a review of related work is deferred to Section 4.6. The existing theory does not easily generalize to other update rules corresponding to different problem geometries. A general theory that connects the notion of early stopping for a more general class of update rules with the well-established theory of localized complexities is still missing. More broadly, a general "language" to reason about the statistical properties of trajectories traced by optimization algorithms applied to the unregularized empirical risk is still lacking.

In this chapter, we study a *family* of update rules given by the mirror descent algorithm [157, 21] (see Section 2.4 for background). Mirror descent, which includes vanilla gradient descent as a special case, is increasingly becoming the tool of choice in optimization and machine learning, applied well beyond the traditional setting of convex optimization. Our choice to study iterative regularization within mirror descent framework is motivated by its ability to exploit non-Euclidean geometries via properly designed mirror maps, the fact that the algorithm admits a general potential-based convergence analysis in terms of Bregman divergences, and its ability to represent a large class of algorithms in a unified and well-developed framework.

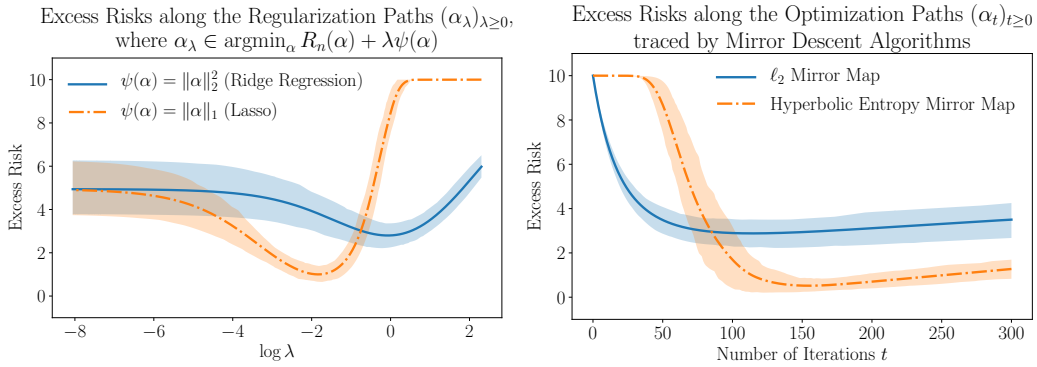
4.1.1 Optimization Algorithms as Regularizers

In implicit regularization literature, the fact that optimization algorithms used to minimize the empirical risk can be seen to introduce an implicit bias, which in turn results in a regularizing effect, has been attracting increasing attention (see, e.g., [75, 76, 190]). This serves as an important source of motivation to consider a general family of optimization algorithms (i.e., the mirror descent family) in the present chapter, as opposed to focusing on vanilla gradient descent updates, known to be closely linked to Tikhonov/ridge type of regularization (see, e.g., [223, 172, 2]).

To better motivate our work, let us now discuss some numeric simulations. We aim to make the following two points:

1. considering iterative regularization mirror descent schemes, different choices of mirror maps are more appropriate for different problems;

2. to explain optimal performance, early stopping needs to be considered, similarly to the consideration of regularization strength in explicitly penalized regularization schemes.



(a) Explicit regularization.

(b) Implicit regularization.

Figure 4.1: A comparison between explicit and implicit (or iterative) regularization schemes. The data generating distribution P is defined by $X \sim N(0, I_d)$ and $Y|X = x \sim \langle \alpha', x\rangle + N(0, 5^2)$ for some parameter $\alpha' \in \mathbb{R}^{100}$, only 10 coordinates of which are non-zero. The sample size is taken to be $n = 200$. In the plots above, the solid lines denote means over 100 runs whereas the shaded regions correspond to the 10-th and the 90-th percentiles.

In Figure 4.1, we demonstrate that different choices of optimization algorithms applied to the unregularized empirical risk R_n yield different statistical performance along the optimization path $(\widehat{g}_{\alpha_t})_{t\geq 0}$, in a similar way that a choice of an explicit regularizer affects the statistical performance along the corresponding regularization path. Due to the sparsity of α' (the optimal model used to generate noisy data in the

simulations shown in Figure 4.1), explicit regularization via ℓ_1 penalization results in a class of models $(\alpha_\lambda)_{\lambda \geq 0}$ that at its minimum achieves significantly lower risk than the class of models generated via ℓ_2 penalization (cf. Figure 4.1a). Figure 4.1b demonstrates a similar phenomenon from an implicit regularization point of view. Due to the sparsity of α', the choice of a hyperbolic entropy mirror map (see Section 2.4.4 for the definition) yields an optimization path that at its minimum achieves excess risk nearly an order of magnitude lower than the path generated by the vanilla gradient descent updates.

Finally, observe that the minimum risk in the iterative regularization schemes (cf. Figure 4.1b) is attained at some finite stopping time $t^\star < \infty$; that is, the optimal performance is achieved *before* the mirror descent iterates converge to an empirical risk minimizer. In a close similarity, optimal performance for explicitly penalized schemes is achieved via a properly tuned regularization strength parameter λ. As a result, our work presented in this chapter focuses on the analysis of early-stopped iterates, which should be contrasted with a more recent line of work where implicit regularization properties are investigated at convergence (see, e.g., [75, 76]). Another important consideration is that early stopping provides a way to save computational resources. As we shall see, the optimal stopping time is connected to the statistical properties of the problem. Hence, the amount of computation required to achieve statistical optimality can be upper-bounded in terms of the statistical complexity of the problem.

4.1.2 Improperness of Early-Stopped Mirror Descent Iterates

The improperness of statistical estimators is a central theme in this thesis. Let us now discuss in what sense the early-stopped mirror descent iterates can be considered improper. As already discussed in the background part of this thesis (see Section 2.2), we define improperness in the context of having a fixed reference class of functions $\mathcal{G}$. Then, an estimator is said to be improper if it is allowed to output predictors that do not belong to the reference class $\mathcal{G}$.

In iterative regularization, improperness arises naturally due to the unconstrained nature of the algorithm. Consider the Figure 4.2. For any $R \geq 0$ let $\mathcal{G}_R = \{\langle \alpha, \cdot \rangle : \|\alpha\|_2 \leq R\}$ be the function class of linear predictors identified with an ℓ_2 ball of radius R. Let $\alpha_{\mathcal{G}_R} = \operatorname{argmin}_{\alpha \in \mathcal{G}_R} R(\alpha)$ denote the population risk minimizer in $\mathcal{G}_R$. In the simulations performed above, we fix $\alpha' = (1.5, 0.5)^\mathsf{T}$, $n = 100$, and consider a

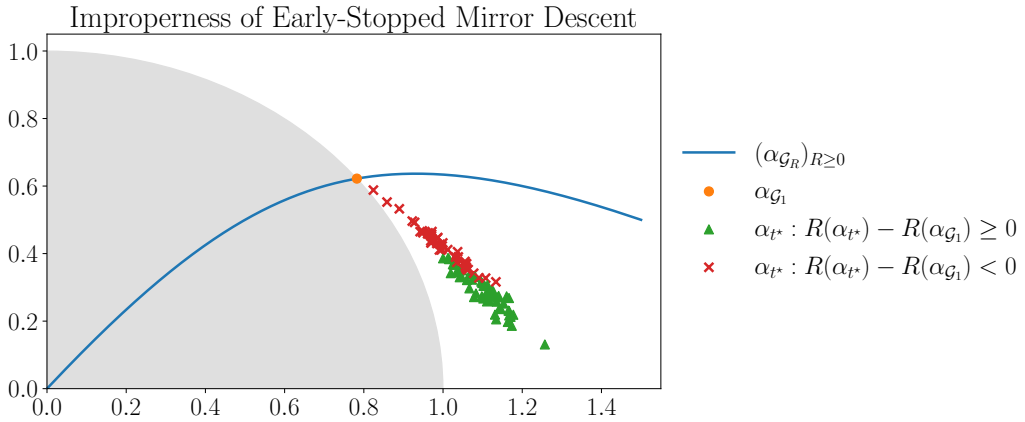


Figure 4.2: A demonstration of improperness of early-stopped mirror descent. The shaded ball denotes a reference class of linear predictors $\mathcal{G}_1$ identified with vectors whose Euclidean norm is bounded by one. The red crosses and green triangles denote early-stopped iterates $\alpha_{t^\star}$ (over different runs for re-sampled data) with the stopping rule $t^\star = \min\{\, t \geq 0 : R_n(\alpha_t) \leq R_n(\alpha_{\mathcal{G}_1})\}$. The plotted iterates fall outside of the shaded region (for the above stopping rule), thus demonstrating the improperness of the considered estimator.

distribution P defined by

$$X \sim N(0, \Sigma) \text{ and } Y|X = x \sim \langle \alpha', x \rangle + N(0, 0.5^2), \text{ where } \Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}.$$

We run the mirror descent algorithm with the mirror map $\psi(\alpha) = \alpha^{\mathsf{T}}\Sigma\alpha/2$, the initialization $\alpha_0 = 0$ and the step-size $\eta = 10^{-3}$. The stopping time is defined by $t^\star = \min\{\, t \geq 0 : R_n(\alpha_t) \leq R_n(\alpha_{\mathcal{G}_1})\}$ so that our early-stopped estimator is identified with the parameter $\alpha_{t^\star}$. The stopping rule considered in the main results section always results in larger stopping times, so that improperness demonstrated for the above stopping rule implies improperness of the estimator to be considered later (cf. Section 4.4). We plot the values of $\alpha_{t^\star}$ (denoted by crosses and triangles) over 100 runs. Figure 4.2 demonstrates the improperness of the considered early-stopped estimator, since the triangles and crosses fall outside of the shaded ball region.

Recall from the discussions in Section 2.5 and Chapter 3 that the so-called Bernstein condition allows us to analyze statistical estimators via the local Rademacher complexity theory, a powerful machinery that provides sharp excess risk bounds in many situations. Bernstein condition requires that for some constant $C > 0$, the inequality $R(\alpha_{t^\star}) - R(\alpha_{\mathcal{G}_1}) \geq C\|f_{\alpha_{t^\star}} - f_{\alpha_{\mathcal{G}_1}}\|^2_{L_2(P)}$ holds. However, in Figure 4.2, the early-stopped iterates marked via red crosses indicate points for which Bernstein condition cannot hold for any non-negative constant C. Of course, this issue would disappear if we replaced the reference class of functions $\mathcal{G}_1$ with $\mathcal{G}_R$ for some large enough $R > 0$ such that the entire blue line in Figure 4.2 is contained in the Euclidean

ball of radius R (i.e., this ball would contain a globally optimal parameter). However, we shall remark that in the spirit of the results contained in this thesis, we aim to provide a general theory of excess risk bounds, where such constraints on the possible reference class of functions should not be placed, and improper settings should be handled.

4.2 Problem Formulation

In this section, we formulate our problem setting. First, we describe our assumptions on the possible representations of prediction functions. Next, we define the mirror descent update schemes. Finally, we set out a goal that we want to achieve in connection with the previously obtained results in the literature.

Representation of Functions. We only consider parametric classes of linear functions in the sense described below. We identify the parameter system by the set $\mathbb{R}^m$ for some natural number m and denote the parameters by $\alpha \in \mathbb{R}^m$. Each vector α identifies a linear function f_α. The linearity of functions is not necessarily defined with respect to the standard basis. Thus, we assume that there exists a data-dependent matrix $Z = Z(S_n) \in \mathbb{R}^{n\times m}$ such that for any input point $X_i \in \mathcal{X}$ in the observed sample S_n and any $\alpha \in \mathbb{R}^m$, it holds that $f_\alpha(X_i) = (Z\alpha)_i$. Let us now provide two example settings commonly studied in iterative regularization literature that admit the above conditions.

Example 4.1 (Linear Regression). The simplest example setting is that of linear regression. Given a data sample $S_n = (X_i, Y_i)_{i=1}^n$, where $X_i \in \mathbb{R}^d$, let $m = d$ and let the i-th row of $Z \in \mathbb{R}^{n\times d}$ be given by the vector X_i^T. For any $\alpha \in \mathbb{R}^d$, let $f_\alpha(\cdot) = \langle \alpha, \cdot \rangle$ be the linear function identified by α. This is the setting of simulations performed in Figures 4.1 and 4.2, for different choices of mirror maps.

Example 4.2 (Kernel Regression). We now discuss a setting of kernel regression frequently considered in iterative regularization literature (e.g., [223, 20, 172, 219]).

Let $S_n = (X_i, Y_i)_{i=1}^n$ denote an observed data sample, where $X_i \in \mathcal{X}$ for some abstract space $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a Mercer kernel which induces a Hilbert space of functions $\mathcal{H}$ equipped with norm $\|\cdot\|_\mathcal{H}$. Then, conditionally on the observed sample S_n, denote by $K \in R^{n\times n}$ a matrix such that $K_{ij} = k(x_i, x_j)$. To each $\alpha \in \mathbb{R}^n$, we may associate a function $f_\alpha \in \mathcal{H}$ defined as $f_\alpha = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$. Thus, for any $i = 1, \ldots, n$, we have $f_\alpha(X_i) = (K\alpha)_i$ and hence we may set $m = n$ and $Z = K$.

We refer the interested reader to the book by Scholkopf and Smola [186] for more background on reproducing kernel Hilbert spaces.

Family of Iterative Regularization Schemes. We consider the family of mirror descent algorithms (see Section 2.4 for background), characterized by an initialization point $\alpha_0 \in \mathbb{R}^m$, the choice of a mirror map $\psi : \mathbb{R}^m \to \mathbb{R}$ (cf. Definition 2.3), and the choice of a step-size sequence $(\eta_t)_{t \geq 0}$. The iterates are defined by the following update rule, for $t = 0, 1, 2, \ldots$:

$$\nabla \psi(\alpha_{t+1}) = \nabla \psi(\alpha_t) - \eta_t \nabla R_n(\alpha_t),$$

where we denote by $R_n(\alpha)$ the empirical risk of the function f_α, i.e., $R_n(\alpha)$ is a shorthand for $R_n(f_\alpha)$. Thus, in light of the function class representation discussed above, for some matrix $Z \in \mathbb{R}^{n \times m}$ that depends on the observed sample S_n we have

$$R_n(\alpha_t) = \frac{1}{n} \sum_{t=1}^{n} \ell(f_\alpha(X_i), Y_i) = \frac{1}{n} \sum_{t=1}^{n} \ell((Z\alpha)_i, Y_i).$$

Our Aims. Iterative regularization is widely studied, primarily in the context of gradient descent updates (we review the literature in Section 4.6). The closest works to our are the papers by Raskutti, Wainwright, and Yu [172], Wei, Yang, and Wainwright [219], where early-stopped gradient descent updates are studied in the setting of reproducing kernel Hilbert spaces (cf. Example 4.2). These are the only prior works that make connections between the analysis of iterative regularization schemes and localized Rademacher complexities (or its Gaussian analogues), a statistical notion of complexity that yields sharp excess risk bounds in a variety of settings.

In this chapter, we aim to extend the works [172, 219] in two ways.

1. First, we aim to provide local Rademacher complexity excess risk bounds without assuming that the data is generated within a well-specified statistical model, an assumption present in the above works concerning the random design setting.

2. Second, we aim to provide a general analysis that holds for a family of mirror descent updates, as opposed to vanilla gradient decent updates.

The following section summarizes how we achieve the above two points in the present chapter by connecting the analysis of early-stopped mirror descent iterates to the notion of offset Rademacher complexity introduced in Section 2.5.2 and further developed in Chapter 3.

4.3 Summary of Contributions

We develop a general theory for learning linear models (including kernel machines) that shows how the optimization trajectory of *unconstrained* mirror descent applied to minimize the unregularized empirical risk is *inherently* connected to excess risk guarantees via offset Rademacher complexity. Unlike in most prior work on early stopping, the notion of statistical complexity appears naturally from intrinsic properties of mirror descent applied to the unregularized empirical risk, without invoking lower-level arguments related to concentration to the *fictitious* population version of the algorithm. Furthermore, our theory leads to an explicit characterization of stopping times from the point of view of both optimization and statistics, which directly yields excess risk bounds and allows us to re-derive previously established results, and some new results, in a more straightforward fashion.

As discussed in the introduction, early-stopped unconstrained iterative algorithms are improper. Hence, following the discussion in Chapter 3, the analysis of such estimators does not easily fit within the mathematical framework of classical localization techniques, partially explaining the scarcity of results connecting localized complexity measures with such algorithms. Offset Rademacher complexities, on the other hand, open up another avenue for establishing such connections via the design of update rules tailored to satisfy the offset condition (cf. Definition 3.1). We show that the mirror descent updates applied to the empirical loss R_n simultaneously *implicitly* minimize the function $R_n(\alpha) - R_n(\alpha') + \|f_\alpha - f_{\alpha'}\|_n^2$ for *all* reference points α' up to a certain stopping time (which depends on α'), while also *staying inside a certain Bregman "ball"* centered at α' up to the corresponding stopping time. Consequently, a suitably early-stopped mirror descent iterate can be shown to satisfy the offset condition while remaining in a certain bounded set centered around the chosen reference function. While mirror descent was developed within the framework of convex optimization, it has also found applications in a wide range of problems, including bandits [1], online learning [84], the k-server problem [39] and metrical task systems [40]. In this respect, our work can be seen as an exposition of another example where mirror descent naturally solves a problem outside its originally intended scope.

Demonstration of the Main Idea. To show the key idea of the proof technique that we present in this chapter, for simplicity, let us temporarily fix the loss function $\ell(y_1, y_2) = (y_1 - y_2)^2$ to be the quadratic loss. In addition, for the sake of the exposition, we consider the continuous-time (instead of discrete-time updates, which are more

complicated to analyze) mirror descent flow given by:

$$\frac{d}{dt}\alpha_t = (\nabla^2\psi(x_t))^{-1}\nabla R_n(\alpha_t),$$

where recall that ψ is a mirror map that characterizes the mirror descent flow.

In what follows, let α' be some arbitrary reference point. The simplified continuous-time setting with the quadratic loss will allow us to exhibit the key insight behind our main results in a clear and straightforward manner. We begin by proving the following lemma.

Lemma 4.1. *Let $R_n(\alpha) = \frac{1}{n}\|Z\alpha - y\|_2^2$ be the empirical risk induced by the quadratic loss. Then, for any $\alpha, \alpha' \in \mathbb{R}^m$, the following holds:*

$$\langle -\nabla R_n(\alpha), \alpha' - \alpha\rangle = R_n(\alpha) - R_n(\alpha') + \|f_\alpha - f_{\alpha'}\|_n^2.$$

Proof. We have

$$\begin{aligned}
&\langle -\nabla R_n(\alpha), \alpha' - \alpha\rangle \\
&= \frac{2}{n}\left\langle -Z^\mathsf{T}(Z\alpha - y), \alpha' - \alpha\right\rangle \\
&= \frac{2}{n}\left\langle -(Z\alpha - Z\alpha' + Z\alpha' - y), Z(\alpha' - \alpha)\right\rangle \\
&= \frac{2}{n}\|Z\alpha - Z\alpha'\|_2^2 - \frac{1}{n}\cdot 2\left\langle Z\alpha' - y, Z(\alpha' - \alpha)\right\rangle \\
&= \frac{2}{n}\|Z\alpha - Z\alpha'\|_2^2 - \frac{1}{n}\cdot\left(\|Z\alpha' - y\|_2^2 + \|Z(\alpha - \alpha')\|_2^2 - \|Z\alpha - y\|_2^2\right) \\
&= \frac{2}{n}\|Z\alpha - Z\alpha'\|_2^2 - \frac{1}{n}\cdot\left(nR_n(\alpha') + \|Z(\alpha - \alpha')\|_2^2 - nR_n(\alpha)\right) \\
&= R_n(\alpha) - R_n(\alpha') + \|f_\alpha - f_{\alpha'}\|_n^2,
\end{aligned}$$

where the fourth line follows by applying the equality $2\langle a, b\rangle = \|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2$, which holds for any vectors $a, b \in \mathbb{R}^m$. □

To appreciate the significance of the above lemma in our context, we will now revisit the potential-based proof of mirror descent presented in Equation (2.7) in Section 2.4.2. This time, instead of using the convexity of R_n which gives $\langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t\rangle \geq R_n(\alpha_t) - R_n(\alpha')$, we directly plug in the identity given in Lemma 4.1 into Equation (2.7). In turn, we obtain the following *equality*:

$$-\frac{d}{dt}D_\psi(\alpha', \alpha_t) = R_n(\alpha_t) - R_n(\alpha') + \|f_{\alpha_t} - f_{\alpha'}\|_n^2.$$

The above equation shows that while $R_n(\alpha_t) - R_n(\alpha') + \|f_{\alpha_t} - f_{\alpha'}\|_n^2 > 0$, the iterates of mirror descent stay within the Bregman ball $\{\alpha \in \mathbb{R}^m : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0)\}$.

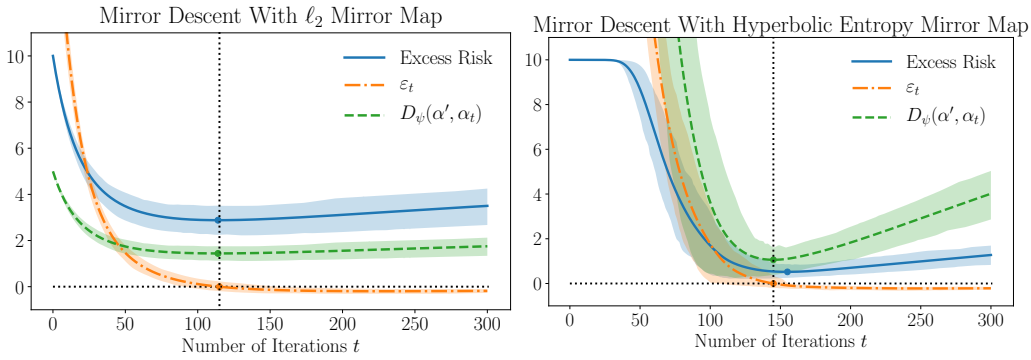


Figure 4.3: Consider the setting of Figure 4.1 and let $\varepsilon_t = R_n(\alpha_t) - R_n(\alpha') + \|f_{\alpha_t} - f_{\alpha'}\|_n^2$. The above plots illustrate the following two points. First, there exists a stopping time time $t^\star$ such that $\varepsilon_{t^\star} \approx 0$ (denoted by the vertical dotted line). Hence, the early stopped estimator $\alpha_{t^\star}$ satisfies the offset condition. Second, while $\varepsilon_t \geq 0$, the Bregman divergence $D_\psi(\alpha', \alpha_t)$ denoted by the green line is non-increasing. It follows that the estimator $f_{\alpha_{t^\star}}$ is constrained to lie in the set $\{f_\alpha : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0)\}$, the offset complexity of which can be used to upper-bound the excess risk of interest. Crucially, this type of analysis does not directly rely on the particular form taken by the mirror descent update rules, which bypasses the limitations present in prior works exploiting closed form solutions of gradient descent iterates. In the plot above, the solid lines denote means over 100 runs, the dots denote the minimum of each solid line, whereas the shaded regions correspond to the 10-th and the 90-th percentiles.

At the same time, the non-negativity of the Bregman divergence $D_\psi(\alpha', \alpha_t)$ shows that the right hand side of the above display equation cannot stay larger than any positive $\varepsilon > 0$ indefinitely, since otherwise the non-negativity of Bregman divergence would be contradicted. Thus, the early-stopped mirror descent iterates (for some stopping time) satisfy the offset condition (cf. Definition 3.1). For a visual demonstration of the above proof sketch see Figure 4.3. We provide full details of this argument (for strongly convex losses) in the proof of Theorem 4.1. A discrete-time result under additional smoothness assumption on the loss is presented in Theorem 4.2.

To summarize, our main contributions are the following:

1. Our work extends the scope of offset Rademacher complexities to a family of early-stopped mirror descent methods. Additionally, we extend the scope of mirror descent to be used as a computationally efficient statistical device in an i.i.d. batch statistical learning setting.

2. Our main results, in a short and transparent way, yield bounds on the excess risk of the iterates of (both continuous-time and discrete-time) mirror descent using offset Rademacher complexities. In contrast to prior work, our arguments require no direct use of low-level mathematical techniques such as symmetrization, peeling, or concentration to the population version of the algorithm.

4.4 Main Results

This section contains the main results, which establish that suitably early-stopped mirror descent iterates satisfy the offset condition. It follows that the excess risk of early-stopped mirror descent statistical estimator can be controlled via the theory of offset Rademacher complexities, introduced in Section 2.5.2 and extended in Chapter 3 of this thesis.

Before stating and proving the main theorems, let us formulate the assumptions needed in our analysis. The first assumption is the one concerning the representation of functions, discussed in Section 4.2.

Assumption 4.1 (Function Class Representation)**.** Let $S_n = (X_i, Y_i)_{i=1}^n$ denote the observed data sample. We assume that there exists a (possibly data dependent) matrix $Z = Z(S_n) \in \mathbb{R}^{n \times m}$, such that for any $\alpha \in \mathbb{R}^m$, the corresponding function f_α satisfies $f(X_i) = (Z\alpha)_i$ for any $i = 1, \ldots, n$.

Next, we shall restrict our attention to strongly convex and differentiable functions, in the sense described below.

Assumption 4.2 (Strong Convexity and Differentiability)**.** Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be the loss function. For any $y' \in \mathcal{Y}$ let $\ell_{y'}$ be the function $y \mapsto \ell(y, y')$. We assume that the following two conditions are satisfied.

1. For any $y' \in \mathcal{Y}$, the function $\ell_{y'}$ is differentiable.
2. For any $y' \in \mathcal{Y}$, the function $\ell_{y'}$ is γ-strongly convex, in the sense that for any $y_1, y_2 \in \mathcal{Y}$ the following inequality holds:
$$\ell_{y'}(y_1) \geq \ell_{y'}(y_2) + \ell'_{y'}(y_2)(y_1 - y_2) + \frac{\gamma}{2}(y_1 - y_2)^2.$$

A simple example of a loss function satisfying the above condition is the quadratic loss $\ell(y, y') = (y - y')^2$, which is 2-strongly convex. Observe that the above condition is much weaker than assuming that the empirical risk function $\alpha \to R_n(\alpha)$ is strongly convex.

In the argument sketched in the previous section, we have restricted ourselves to the case of the quadratic loss that allowed us to prove the *equality* $\langle -\nabla R_n(\alpha), \alpha' - \alpha \rangle = R_n(\alpha) - R_n(\alpha') + \|f_\alpha - f_{\alpha'}\|_n^2$ (cf. Lemma 4.1). However, an observant reader may have already noticed that the argument presented in the previous section only relied on having a *lower bound* on $\langle -\nabla R_n(\alpha), \alpha' - \alpha \rangle$. A lower bound on the latter object follows directly from the strong convexity assumption, as we show in the following lemma.

Lemma 4.2. *Suppose that the loss function ℓ satisfies the Assumption 4.2. Further, suppose that there exists some matrix $Z \in \mathbb{R}^{n\times m}$ such that for any $\alpha \in \mathbb{R}^m$, the function f_α identified with the parameter α satisfies $Z\alpha = (f_\alpha(X_1), \dots, f_\alpha(X_n))^\top$. Then, the following inequality holds for any $\alpha' \in \mathbb{R}^d$:*

$$\langle -\nabla R_n(\alpha), \alpha' - \alpha \rangle \geq R_n(\alpha) - R_n(\alpha') + \frac{\gamma}{2}\|f_\alpha - f_{\alpha'}\|_n^2.$$

Proof. Recall that $(X_i, Y_i)_{i=1}^n$ denotes the observed data sample and that for any $y' \in \mathbb{R}$ we denote by $\ell_{y'}$ the function $y \mapsto \ell(y, y')$. Hence, by the strong convexity assumption on $\ell_{y'}$ for all y' (cf. Assumption 4.2), the following holds for any $i = 1, \dots, n$:

$$\ell_{Y_i}(f_{\alpha'}(X_i)) \geq \ell_{Y_i}(f_\alpha(X_i)) + \ell'_{Y_i}(f_\alpha(X_i))(f_{\alpha'}(X_i) - f_\alpha(X_i)) + \frac{\gamma}{2}(f_\alpha(X_i) - f_{\alpha'}(X_i))^2.$$

Summing the above equation for $i = 1, \dots, n$ and dividing both sides by n yields

$$R_n(\alpha') \geq R_n(\alpha) + \frac{1}{n}\sum_{i=1}^n \ell'_{Y_i}(f_\alpha(X_i))(f_{\alpha'}(X_i) - f_\alpha(X_i)) + \frac{\gamma}{2}\|f_\alpha - f_{\alpha'}\|_n^2. \tag{4.3}$$

Finally, using the fact that $f_\alpha(X_i) = (Z\alpha)_i = z_i^\top \alpha$, where $z_i \in \mathbb{R}^m$ is the i-th row of the matrix Z, we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \ell'_{Y_i}(f_\alpha(X_i))(f_{\alpha'}(X_i) - f_\alpha(X_i)) &= \frac{1}{n}\sum_{i=1}^n \ell'_{Y_i}(z_i^\top\alpha)(z_i^\top\alpha' - z_i^\top\alpha) \\
&= \frac{1}{n}\sum_{i=1}^n \left(z_i \ell'_{Y_i}(z_i^\top\alpha)\right)^\top (\alpha' - \alpha) \\
&= \frac{1}{n}\sum_{i=1}^n \left(\nabla_\alpha \ell_{Y_i}(f_\alpha(X_i))\right)^\top (\alpha' - \alpha) \\
&= \langle \nabla R_n(\alpha), \alpha' - \alpha \rangle .
\end{aligned}$$

Plugging the above identity into the inequality (4.3) and rearranging completes the proof. □

We will now state and prove our main results. The rest of this section is split into two parts. First, we prove a continuous-time result (Section 4.4.1) under the assumptions stated above. Next, in Section 4.4.2, we prove a discrete-time result under the additional assumptions on the smoothness of the loss and strong convexity of the mirror map.

4.4.1 Continuous-Time Version of the Main Result

This section establishes that the continuous-time mirror descent flow, along its optimization path, visits a point that satisfies the offset condition, with respect to an arbitrary reference point. More specifically, the first part of the below theorem shows that the iterates of mirror descent stay within a certain Bregman ball up to the prescribed stopping time $t^\star$. The second part of the theorem establishes that for any fixed reference point α', the early-stopped estimator $f_{\alpha_{t^\star}}$ satisfies the deterministic offset condition (cf. Definition 3.1) with $\gamma = 1$ and an arbitrary accuracy parameter $\varepsilon > 0$. Observe that the upper bound on the stopping time $t^\star$ is of order $D_\psi(\alpha', \alpha_0)/\varepsilon$, so that achieving higher statistical accuracy requires more computational power. In the applications considered in the paper on which the present chapter is based [211], ε is chosen of the same size as the resulting offset complexity, since choosing smaller ε increases computational cost without further improving the resulting excess risk bound. Thus, the statistical notion of offset Rademacher complexity also controls the amount of necessary computational resources from the point of view of the upper bounds presented in this chapter (note that we do not present any lower bound). Finally, we note that the dependence of $t^\star$ on the unknown radius $D_\psi(\alpha', \alpha_0)$ is unavoidable purely from an optimization point of view.

Theorem 4.1. *Suppose that function representation assumption (Assumption 4.1) and the γ-strong convexity assumption (Assumption 4.2) hold. Let $\alpha_0 \in \mathbb{R}^m$ be the initialization point and $\psi : \mathbb{R}^m \to \mathbb{R}$ be a mirror map (cf. Definition 2.3). Consider the continuous-time mirror descent dynamics given by*

$$\frac{d}{dt}\alpha_t = (\nabla^2\psi(\alpha_t))^{-1}\nabla R_n(\alpha_t).$$

Then, for any chosen reference point α' and any $\varepsilon > 0$, there exists a stopping time $t^\star = t^\star(S_n, \psi, \alpha_0, \alpha') \leq 2D_\psi(\alpha', \alpha_0)/\varepsilon$ such that:

1. *For all $0 \leq t \leq t^\star$, $f_{\alpha_t} \in \mathcal{G}(\psi, \alpha_0, \alpha') = \{f_\alpha \in \mathbb{R}^m : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0)\}$.*
2. *At the stopping time $t^\star$, we have $R_n(\alpha_{t^\star}) - R_n(\alpha') + \frac{\gamma}{2}\|f_{\alpha_{t^\star}} - f_{\alpha'}\|_n^2 \leq \varepsilon$.*

Proof. Using Lemma 4.2 instead of Lemma 4.1, we may repeat the proof sketched in Section 4.3. To simplify the notation let $\delta_t = R_n(\alpha_t) - R_n(\alpha')$ and $r_t = \frac{\gamma}{2}\|f_{\alpha_t} - f_{\alpha'}\|_n^2$. Recall that in Section 2.4.2, we showed that the continuous-time mirror descent iterates satisfy the following identity

$$-\frac{d}{dt}D_\psi(\alpha', \alpha_t) = \langle -\nabla R_n(\alpha), \alpha' - \alpha\rangle .$$

Combining the above equation with Lemma 4.2 we obtain the following bound on the continuous-time change of Bregman divergence

$$-\frac{d}{dt}D_\psi(\alpha', \alpha_t) \geq r_t + \delta_t.$$

Let $T = 2D_\psi(\alpha', \alpha_0)/\varepsilon$. Integrating both sides of the above inequality we obtain

$$\begin{aligned} D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_T) &= \int_0^T -\frac{d}{dt}D_\psi(\alpha', \alpha_t)dt \geq \int_0^T (r_t + \delta_t)dt \\ \implies \inf_{0\leq t\leq T}\{r_t + \delta_t\} &\leq \frac{1}{T}\int_0^T (r_t + \delta_t)dt \leq \frac{D_\psi(\alpha', \alpha_0)}{T} \leq \frac{\varepsilon}{2}. \end{aligned}$$

It follows that the following infimum is well defined:

$$t^\star = \inf\{0 \leq t \leq T \mid r_t + \delta_t \leq \varepsilon\}.$$

Hence, $r_{t^\star} + \delta_{t^\star} \leq \varepsilon$ and for all $0 \leq t \leq t^\star$ we have

$$D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_t) \geq \int_0^t (r_t + \delta_t)dt \geq t\varepsilon \geq 0.$$

The above inequality implies that $D_\psi(\alpha', \alpha_t) \leq D_\psi(\alpha', \alpha_0)$, which concludes our proof. □

4.4.2 Discrete-Time Version of the Main Result

In the following theorem, we prove a discrete-time counterpart to the continuous-time theorem proved in the previous section. We will show a variant of a discrete-time result under smoothness of the empirical loss function R_n and under strong convexity of the mirror map; such assumptions are natural from the optimization point of view (see, e.g., the monograph by Bubeck [38]).

Let $\|\cdot\|$ denote any norm. We say that R_n is β-smooth with respect to $\|\cdot\|$ if $R_n(\alpha') \leq R_n(\alpha) + \langle \nabla R_n(\alpha), \alpha' - \alpha\rangle + \frac{\beta}{2}\|\alpha - \alpha'\|^2$ for any α, α' in the domain of R_n. We also say that the mirror map ψ is ρ-strongly convex with respect to $\|\cdot\|$ if for any α, α' we have $D_\psi(\alpha', \alpha) \geq \frac{\rho}{2}\|\alpha' - \alpha\|^2$.

With the definition of smoothness and strong convexity with respect to general norms in place, we are now ready to state the discrete time theorem.

Theorem 4.2. *Suppose that function representation assumption (Assumption 4.1) and the γ-strong convexity assumption (Assumption 4.2) hold. Additionally, suppose that the empirical risk function R_n is β-smooth and the mirror map ψ is ρ-strongly*

convex with respect to some norm $\|\cdot\|$. *Let* $\alpha_0 \in \mathbb{R}^m$ *be the initialization point, let* $0 < \eta \leq \frac{\rho}{\beta}$ *be the step size. Consider the discrete-time mirror descent updates given by*

$$\nabla\psi(\alpha_{t+1}) = \nabla\psi(\alpha_t) - \eta\nabla R_n(\alpha_t).$$

Then, for any chosen reference point α' *and any* $\varepsilon > 0$, *there exists a stopping time* $t^\star = t^\star(S_n, \psi, \alpha_0, \alpha', \eta) \leq (D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha'))/(\eta\varepsilon)$ *such that:*

1. *For all* $0 \leq t \leq t^\star$, $f_{\alpha_t} \in \mathcal{G}(\psi, \alpha_0, \alpha', \eta) = \{f_\alpha : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha')\}$.

2. *At the stopping time* $t^\star$, *we have* $R_n(\alpha_{t^\star}) - R_n(\alpha') + \frac{\gamma}{2}\|f_{\alpha_{t^\star}} - f_{\alpha'}\|_n^2 \leq \varepsilon$.

Before providing the proof we briefly comment on the above theorem. First, the step-size condition $\eta \leq \rho/\beta$ and the number of iterations $O(1/\varepsilon)$ needed to reach a desired level of accuracy are identical to the guarantees proved in purely convex optimization settings (cf. Theorem 4.4 in [38]). On the other hand, comparing Theorems 4.1 and 4.2, in the discrete setting we pay a price of $\eta R_n(\alpha')$ in the radius of the Bregman ball where our early-stopped estimator lies. This is consistent with prior work in the early stopping literature, where such an expansion of the radius dependent on the noise level propagates into the resulting bounds (cf. definition of C in Theorem 1 in [219]). Our work, on the other hand, allows for a more fine-grained control of statistical-computational trade-offs via a selection of a small enough step-size η.

We now introduce an auxiliary lemma supporting the proof of the above theorem. The following lemma proves a discrete-time counterpart to the inequality $-\frac{d}{dt}D_\psi(\alpha', \alpha_t) \geq r_t + \delta_t$, where we remind the reader that $\delta_t = R_n(\alpha_t) - R_n(\alpha')$ and $r_t = \frac{\gamma}{2}\|f_{\alpha_t} - f_{\alpha'}\|_n^2$.

Lemma 4.3. *Consider the setting of Theorem 4.2. The following inequality holds for all* $t \geq 0$*:*

$$D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t).$$

Proof. Recall the identity proved in Lemma 2.2 in Section 2.4.3, which holds for any x, y, z in the domain of the mirror map ψ:

$$D_\psi(z, x) - D_\psi(z, y) = \langle\nabla\psi(y) - \nabla\psi(x), z - x\rangle - D_\psi(x, y).$$

Combining the above identity with the definition of discrete-time mirror descent updates, we obtain

$$\begin{aligned}
&D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \\
&= \langle \nabla\psi(\alpha_t) - \nabla\psi(\alpha_{t+1}), \alpha_t - \alpha' \rangle - D_\psi(\alpha_t, \alpha_{t+1}) \\
&= \langle \eta \nabla R_n(\alpha_t), \alpha_t - \alpha' \rangle - (\psi(\alpha_t) - \psi(\alpha_{t+1}) - \langle \nabla\psi(\alpha_{t+1}), \alpha_t - \alpha_{t+1} \rangle) \\
&= \langle \eta \nabla R_n(\alpha_t), \alpha_t - \alpha' \rangle - (-D_\psi(\alpha_{t+1}, \alpha_t) + \langle \nabla\psi(\alpha_t) - \nabla\psi(\alpha_{t+1}), \alpha_t - \alpha_{t+1} \rangle) \\
&= \langle \eta \nabla R_n(\alpha_t), \alpha_t - \alpha' \rangle - (-D_\psi(\alpha_{t+1}, \alpha_t) + \langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1} \rangle) \\
&= \langle \eta \nabla R_n(\alpha_t), \alpha_t - \alpha' \rangle + D_\psi(\alpha_{t+1}, \alpha_t) + \langle -\eta \nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1} \rangle \,. \qquad (4.4)
\end{aligned}$$

By the ρ-strong convexity of the mirror map ψ, the second term in Equation (4.4) can be lower-bounded as $D_\psi(\alpha_{t+1}, \alpha_t) \geq \frac{\rho}{2} \|\alpha_{t+1} - \alpha_t\|^2$. The last term in Equation (4.4) can be lower-bounded by using the β-smoothness condition of the empirical risk function R_n, which yields $\langle -\nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1} \rangle = \langle \nabla R_n(\alpha_t), \alpha_{t+1} - \alpha_t \rangle \geq R_n(\alpha_{t+1}) - R_n(\alpha_t) - \frac{\beta}{2}\|\alpha_{t+1} - \alpha_t\|^2$. We can hence continue from Equation (4.4) as follows:

$$\begin{aligned}
&D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \\
&= \langle \eta \nabla R_n(\alpha_t), \alpha_t - \alpha' \rangle + D_\psi(\alpha_{t+1}, \alpha_t) + \langle -\eta \nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1} \rangle \\
&\geq \langle \eta \nabla R_n(\alpha_t), \alpha_t - \alpha' \rangle + \frac{\rho}{2} \|\alpha_{t+1} - \alpha t\|^2 + \eta \left(R_n(\alpha_{t+1}) - R_n(\alpha_t) - \frac{\beta}{2} \|\alpha_{t+1} - \alpha_t\|^2 . \right) \\
&= \langle \eta \nabla R_n(\alpha_t), \alpha_t - \alpha' \rangle + \left(\frac{\rho - \eta\beta}{2} \right) \|\alpha_{t+1} - \alpha_t\|^2 + \eta(\delta_{t+1} - \delta_t).
\end{aligned}$$

Since $\eta \leq \rho/\beta$, the second term is lower-bounded by 0. Also, by Lemma 4.2, the first term becomes $\eta\langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t \rangle \geq \eta(\delta_t + r_t)$. Combining these two observations with the last equation above we obtain

$$D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t),$$

which completes our proof. □

With Lemma 4.3 at hand, we can prove Theorem 4.2 following along the same steps used to prove Theorem 4.1, albeit with the continuous-time equation $-\frac{d}{dt} D_\psi(\alpha', \alpha_t) = \delta_t + r_t$ replaced with its discrete-time counterpart $D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t)$. In the discrete-time equation, δ_t is replaced with δ_{t+1}, which results in the expansion of the radius of the Bregman ball in which the mirror descent iterates lie before the prescribed stopping time (see the discussion following the statement of Theorem 4.2 above).

Proof of Theorem 4.2. By Lemma 4.3 we have $D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t)$. Let $T = \left\lceil \frac{D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha')}{\eta\varepsilon} \right\rceil$. Summing both sides of the above equation for $t = 0, \ldots, T$ we obtain

$$D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_T) \geq \eta r_0 + \sum_{t=1}^{T} \eta(r_t + \delta_t) + \eta\delta_{T+1}$$
$$\implies \min_{t=1,\ldots,T} \{\delta_t + r_t\} \leq \frac{\sum_{t=1}^{T} r_t + \delta_t}{T} \leq \frac{D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha')}{\eta T} \leq \varepsilon,$$

where in the last line we have used the definition of T and facts that $D_\psi(\alpha', \alpha_T) \geq 0$, $r_0 \geq 0$, and $\delta_{T+1} \geq -R_n(\alpha')$.

It follows that the following minimum is well defined: $t^\star = \min\{t = 0, \ldots, T \mid r_t + \delta_t \leq \varepsilon\}$. Hence, $r_{t^\star} + \delta_{t^\star} \leq \varepsilon$, which completes the second part of the theorem. To complete the first part of the theorem, note that for any $1 \leq t \leq t^\star$ by telescoping the equation $D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t)$ from 0 to $t - 1$ we obtain

$$D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_t) \geq \eta r_0 + \sum_{t=1}^{t-1} \eta(r_t + \delta_t) + \eta\delta_t$$
$$\implies D_\psi(\alpha', \alpha_t) \leq D_\psi(\alpha', \alpha_0) - \sum_{t=1}^{t-1} \eta(r_t + \delta_t) - \eta\delta_t \leq D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha'),$$

where in the last line we have used the facts that $\delta_t + r_t > \varepsilon > 0$ and $-\delta_t \leq R_n(\alpha')$. □

4.5 Limitations and Open Directions

Among the most natural directions for extending the results presented in this chapter are extensions of our results to stochastic or accelerated versions of mirror descent; local Rademacher complexity analysis of iterative regularization schemes in such regimes have yet to be established. Let us, however, explain a different direction in connection to sparse linear prediction.

A part of the motivation for considering mirror descent updates in this chapter was to provide a simpler analysis for earlier work performed in collaboration with Varun Kanade and Patrick Rebeschini [210]. The problem considered therein was composed of a design matrix $Z \in \mathbb{R}^{n \times d}$, coupled with a well-specified model of observations, i.e., the existence of a vector α' such that the observations $y \in \mathbb{R}^n$ follow the distribution $y = Z\alpha' + \xi$, where ξ is a vector with i.i.d. zero-mean σ^2-subGaussian components. Assume that α' is a sparse vector and that the design matrix Z satisfies some regularity properties (particularly, the restricted isometry property). Given the design matrix Z and the observations y, the aim is to find a vector $\widehat{\alpha}$ that is close to α' in the ℓ_2 norm.

Considering the above *sparse estimation problem*, it was shown in [210] that the following iterative regularization scheme for computing $(\alpha_t)_{t\geq 0}$ achieves minimax optimal statistical estimation rates[1] (if stopped at an appropriate time):

$$\begin{aligned} u_0 = v_0 = \sqrt{\gamma/2}\cdot \mathbb{1}, \quad \alpha_t = u_t \odot u_t - v_t \odot v_t, \\ u_{t+1} = u_t \odot (\mathbb{1} - 2\eta \nabla R_n(\alpha_t)), \quad v_{t+1} = v_t \odot (\mathbb{1} + 2\eta \nabla R_n(\alpha_t)), \end{aligned} \tag{4.5}$$

where $\odot$ denotes the Hadamard product (i.e., component-wise multiplication of vectors). One limitation of the analysis presented in [210] is that the proofs there follow an ad-hoc strategy that is rather long and does not easily extend to other iterative regularization schemes. In contrast, the analysis presented in this chapter simultaneously treats a family of iterative regularization schemes while also following conventional proof strategies originating in Statistics and Optimization literature.

Noting that $1 + x \approx e^x$ for small x, we can approximate the above updates (with the step-size η rescaled by a constant factor by the unconstrained exponential gradient (EG$\pm$) algorithm of Kivinen and Warmuth [98], whose updates are given by

$$\begin{aligned} \alpha_0^+ = \alpha_0^- = (\gamma/2)\mathbb{1}, \quad \alpha_t = \alpha_t^+ - \alpha_t^-, \\ \alpha_{t+1}^+ = \alpha_t^+ \odot \exp(-\eta \nabla R_n(\alpha_t)), \quad \alpha_{t+1}^- = \alpha_{t+1}^- \odot \exp(\eta \nabla R_n(\alpha_t)), \end{aligned} \tag{4.6}$$

where the exponentiation operation is applied componentwise. It was shown in [69] that the above updates correspond to running unconstrained mirror descent initialized at the origin with the hyperbolic entropy mirror map defined in Section 2.4.4; thus, the above updates can be analyzed within the framework of the present chapter. Indeed, it is shown in [211] – the paper on which the present chapter is based – that the iterative regularization scheme given by (4.6) can achieve, up to multiplicative logarithmic factors, a minimax optimal rate for the *in-sample prediction error* $\frac{1}{n}\|Z\alpha - Z\alpha'\|_2^2$, under ℓ_1 norm constraints on the parameter α'.

We can now state one limitation of the framework presented in this chapter. While the theory developed in this chapter can yield nearly optimal rates for the in-sample prediction error under ℓ_1 norm constraints on the optimal prediction vector, it is not clear if the results obtained for the updates (4.5) in [210] *under sparsity constraints on the optimal parameter* and the restricted isometry assumption on the design can be recovered using the technique developed in this chapter. A particular difficulty in recovering the results of [210] within the framework presented in this chapter is that it is not enough to constrain the mirror descent iterates to lie in a certain Bregman

[1] Of course, other procedures, such as the lasso, also obtain the minimax optimal rate under the same assumptions; see the discussions in [210] for more details.

ball induced by the hyperbolic entropy mirror map to conclude that the path traced by the mirror descent algorithm will only contain sparse vectors up to the prescribed stopping time. The latter fact is established in [210] via an analysis tailored to the particular form of the updates (4.5) and the specific properties of matrices that satisfy the restricted isometry property. In other words, it is not clear if the theory developed in this chapter can be applied to recover the results of [210] in a clean way, without resorting to ad-hoc proof strategies.

4.6 Bibliographic Remarks

This chapter is based on joint work performed in collaboration with Varun Kanade and Patrick Rebeschini [211]. The presentation in this thesis departs from the paper [211] in two aspects. First, we reformulate main results for strongly-convex loss functions, instead of the quadratic loss considered in [211]. Second, we omit the example applications available in [211].

As discussed in Section 4.5, a part of the motivation for studying non-Euclidean iterative-regularization schemes stemmed from the prior work of the same authors on early-stopping applied to a sparse recovery problem [210], the latter work not covered in this thesis beyond what is mentioned in the preceding section. Applications of early stopping regularization to sparse recovery problems were considered even earlier, with a different algorithm involving soft truncations, but nevertheless closely related to the ideas of mirror descent; see the work by Osher, Ruan, Xiong, Yao, and Yin [164] and the references therein for further details.

The idea of iterative regularization has a long history. Early ideas can be traced back to the stochastic approximation arguments of Robbins and Monro [176]. Even more closely related are the ideas put forth by Louis Landweber [110], yielding one of the regularization schemes in the theory of inverse problems; see the book by [65] for further details and a more extensive background from the inverse problems point of view. In the Statistics literature, the first work to analyze early-stopped gradient descent in connection to minimax optimality appears to be due to Bühlmann and Yu [41], formulated in the context of L_2-boosting algorithms. Regarding early stopping regularization for boosting algorithms, see also the works [93, 225, 25, 17]. However, from the practical perspective, early stopping regularization was used long before, for example, in neural network training [168].

Closer to the setting investigated in the present chapter, statistical and computational properties of unconstrained *gradient descent* updates have been a subject of intense study over the past two decades. Most of the existing results focus on

regression in reproducing kernel Hilbert spaces (RKHS). For the quadratic loss, see the works [41, 223, 20, 172, 28], while for general loss functions, see [123, 219]. It shall be noted that some of the works investigating early stopping focus on attaining bounds in the $\|\cdot\|_n$ or in the $\|\cdot\|_{L_2(P)}$ norms. The former notion of quality measures the adequacy of the learned function assuming the design is non-random, while the latter is different from the excess risk considered in this chapter, and bounds obtained in $L_2(P)$ norms do not, in general, imply bounds on the excess risk (see [189, Section 1] for an example). In addition, the analysis in the above-cited works [41, 223, 20, 172] loss function is closely tied to the ℓ_2 geometry of the gradient descent updates, which allows one to express predictions of early-stopped in a relatively simple linear operator equation, where the linear operator acts on the observed labels. Spectral properties of these linear operators (depending on the stopping time) are then analyzed as a function of the number of iterations, which can be solved for a stopping time via some form of bias-variance decomposition. Our work, in contrast, enables simultaneously studying a family of iterative regularization schemes without relying on closed form solutions of the early-stopped iterates, which are unavailable within the general context considered in this chapter.

We note that the works [28, 123, 219] do not rely on a particular form of the iterates and are closer to the technique of the present chapter. In particular, they implicitly exploit potential-based convergence proofs from the optimization literature to constrain the iterates to some set prior to early stopping. In our work, the main difference is that we obtain extra quadratic negative terms that guarantee that the early-stopped estimator (using our stopping rule) satisfies the offset condition, from which the random design excess risk bounds can be deduced.

One of the primary contributions of our work is the connection between mirror descent iterates and localized complexity measures. To the best of our knowledge, there are only two prior works making connections of a similar nature, albeit only in the setting of Euclidean gradient descent updates, that is, with the choice of the mirror map $\psi(\alpha) = \|\alpha\|_2^2/2$ [172, 219]. Such connections are observed in an algebraic fashion in the former work, while localized complexities appear more naturally in [219], via the analysis of the range of estimators defined by gradient descent iterates up the prescribed stopping time. In this respect, the work [219] is the closest to ours.

Beyond the Euclidean setup, interest in understanding the generalization properties of neural networks has sparked research into *implicit* regularization properties of various factorized models. In the context of neural networks, the authors of [75, 119, 5, 221, 70] show that iterates of gradient descent applied to factorized matrix models are implicitly

biased towards some sparsity-inducing structure such as low-rankness or low nuclear norm. Such results, however, hold under certain limit statements, such as vanishing initialization or step-size, the number of iterations going to infinity, or no noise in the problem. In the setting of linear regression, matrix factorization models reduce to vector Hadamard product factorizations, where early-stopped gradient descent was shown to yield minimax optimal rates for sparse recovery with the analysis vitally relying on the restricted isometry property [226, 210].

Implicit regularization properties of mirror descent have recently attracted a considerable amount of attention; however, most results in this area either focus on optimization guarantees that do not provide any direct link to statistical guarantees on out-of-sample prediction [76, 10], or establish a connection to statistics via some forms of explicit regularization [190]. The work [190] shows connections between the iterates on the entire path and the solutions on the regularization path for a suitable regularized risk minimization problem. Yet other papers have used early stopping to solvers applied directly to appropriately constrained problems and regularization-promoting structures encoded directly into the loss function [134].

Recent work has also focused on providing statistical guarantees for iterates generated via gradient descent updates in stochastic [179, 122, 159, 3], accelerated [53, 166], and distributed settings [121, 173, 174]. One open direction concerning our work is extending the offset Rademacher complexity analysis of early-stopped estimators in the above settings.

5 Suboptimality of Constrained Least Squares

This chapter is based on joint work with Nikita Zhivotovskiy [209]. We investigate the performance of the empirical risk minimization estimator under arguably the most favorable setting where the loss function is quadratic, and the reference class of functions is convex; thus, in particular, the Bernstein condition is satisfied. The main finding is that even in such a regime, improper predictors can offer significant dimension-dependent improvements in the incurred excess risk.

The results of this chapter assume familiarity with the mathematical framework of Statistical Learning (Section 2.1) and the concept of improper learning (Section 2.2). Familiarity with local Rademacher complexity measures (Section 2.5) is helpful but not essential. A summary of notation is available in Section 2.6.

5.1 Introduction

In Chapters 3 and 4, we studied problem settings concerning general classes of reference functions and general loss functions. In this chapter, we will restrict our attention to a more specialized problem of learning linear functions with the quadratic loss, arguably among the most well-studied problem setups in Learning Theory and Statistics. We now introduce notation/problem setting specific to this thesis chapter:

1. for a positive constant $b > 0$, we fix the reference class of functions $\mathcal{G}_b = \{\langle w, \cdot \rangle : w \in \mathcal{W}_b\}$, where $\mathcal{W}_b = \{w \in \mathbb{R}^d : \|w\| \leq b\}$ and $\|\cdot\|$ denotes the Euclidean norm;

2. we fix the loss function to be the quadratic loss $\ell(y, y') = (y - y')^2$;

3. for constants $r, m > 0$, we let the data generating distribution $P = P_{r,m}$ be a distribution such that $(X, Y) \sim P$ is contained, with probability one, in the set $\{x \in \mathbb{R}^d : \|x\| \leq r\} \times [-m, m]$;

4. for any $w \in \mathbb{R}^d$, we make the following notational simplifications: $R(w) \coloneqq R(\langle w, \cdot \rangle)$ and $R_n(w) \coloneqq R_n(\langle w, \cdot \rangle)$.

We study the statistical performance of any empirical risk minimizer over $\mathcal{G}_b$, defined by

$$\widehat{w}_b^{\mathrm{ERM}} \in \operatorname*{arg\,min}_{w \in \mathcal{W}_b} R_n(w).$$

Thence, we aim to understand to what extent, as characterized by its excess risk, the empirical risk minimization estimator $\widehat{w}_b^{\mathrm{ERM}}$ predicts as well as the best linear predictor in a bounded Euclidean ball with respect to the squared loss.

As a motivating example, consider the well-specified model $Y = \langle w^*, X\rangle + \xi$. Here $w^* \in \mathbb{R}^d$ and ξ is zero mean random variable independent of X. Assuming additionally that $n \geq 2d$, ξ is Gaussian and that X is zero mean multivariate Gaussian with invertible covariance matrix Σ, a classical result of Breiman and Freedman [36, Theorem 1.1] shows that the *excess risk* of unconstrained least squares estimator $\widehat{w}_\infty^{\mathrm{ERM}}$ (also known as the ordinary least squares estimator) satisfies

$$\mathbf{E}R(\widehat{w}^{\mathrm{ERM}}) - R(w^*) \lesssim \frac{dR(w^*)}{n}, \tag{5.1}$$

where the expectation is taken with respect to the observed data sample $S_n = (X_i, Y_i)_{i=1}^n$, the notation $\lesssim$ suppresses absolute multiplicative constants, and the optimal risk $R(w^*)$ is equal to the variance of the noise random variable ξ. Remarkably, the bound (5.1) depends neither on the exact form of the covariance matrix Σ nor on the magnitude of w^*. Recent work of Mourtada [152, Theorem 1] shows that if the model is well-specified then for any distribution of the covariates X such that the sample covariance matrix is almost surely invertible and any $n \geq d$, the excess risk of unconstrained least squares is *exactly* equal to the minimax risk.

While the above result attests to the existence of regimes where least squares is a statistically optimal estimator in a minimax sense, there is a growing interest in the statistics and machine learning communities in understanding the robustness of statistical estimators to various forms of model *misspecification*. For instance, the regression function $\mathbf{E}(Y|X = \cdot)$ might be non-linear, or the distribution of the noise random variable ξ might depend on the corresponding covariate X. Many authors have matched the d/n rate (5.1) for ERM-based algorithms under less restrictive assumptions than that of a well-specified model with Gaussian covariates; see Section 5.7 for an overview of the existing literature.

Despite the many existing results on the performance of the constrained least squares estimator, a complete understanding of the magnitude of its excess risk in the *distribution-free* setting is lacking. Specifically, by distribution-free, we mean a setting where only boundedness assumptions are imposed, but no other assumptions on $(X, Y) \sim P$ are assumed (such as a well-specified model, well-behaved covariance matrix, zero-mean noise, etc.). Regarding the boundedness assumptions, we remark that for the study of ERM estimator in the setting outlined above, *some* assumptions need to be made since otherwise, any algorithm that returns a linear predictor

(including the least squares estimator) can incur arbitrarily large excess risk (see the lower bounds in [189, 114, 45]).

We now turn to the next section, where we discuss the precise sense in which the current understanding of the statistical performance of the estimator $\widehat{w}_b^{\mathrm{ERM}}$ is insufficient, and we formulate the problem to be addressed in the present chapter.

5.2 Problem Formulation

Recall that in the context of this chapter, an estimator is called proper if it always returns a linear function $\langle \widehat{w}, \cdot \rangle$ with $\widehat{w}$ contained in the bounded Euclidean ball $\mathcal{W}_b$ of radius b. Otherwise, an estimator is called improper. Fix any proper estimator $\widehat{w}_b$ and any constants $r, m > 0$. The work of Shamir [189] shows that there exists a distribution $P = P_{r,m}(\widehat{w}_b)$ satisfying $\|X\| \leq r$ almost surely and $\|Y\|_{L_\infty(P)} \leq m$, such that the following lower bound holds

$$\mathbf{E}R(\widehat{w}_b) - \inf_{\omega \in \mathcal{W}_b} R(\omega) \gtrsim \min\left\{m^2, \min\left\{\frac{dm^2}{n}, \frac{rbm}{\sqrt{n}}\right\} + \frac{r^2b^2}{n}\right\}, \tag{5.2}$$

For large enough sample sizes ($n \gg d$), the above lower bound reduces to the following simplified form:

$$\mathbf{E}R(\widehat{w}_b) - \inf_{w \in \mathcal{W}_b} R(w) \gtrsim \frac{dm^2}{n} + \frac{r^2b^2}{n}. \tag{5.3}$$

To avoid unnecessary technicalities related to the slow rate terms in the bound (5.2), in the rest of this chapter, we will restrict our attention to the lower bound (5.3). Note that the first term in the above lower bound corresponds to the rate in the upper bound that holds for the well-specified models (5.1): the excess risk of the best predictor in the class $\mathcal{W}_b$ is upper bounded by that of a zero function, whose risk is in turn bounded by m^2. On the other hand, the second term in (5.3) shows that in the absence of simplifying distributional assumptions, the statistical performance of linear predictors can deteriorate arbitrarily with respect to the boundedness constants r, b, even in one-dimensional settings; in contrast, the upper bound (5.1) does not depend on b and r.

A baseline for our work is a conjecture proposed by Shamir [189] postulating the statistical optimality of the constrained least squares estimator $\widehat{w}_b^{\mathrm{ERM}}$ in a sense that it matches the lower bound (5.3). For some of the recent discussions and attempts to resolve this conjecture see, for example, the works [108, 12, 74, 218]. The existing results, however, only partially address this conjecture, restricting to the regimes where $br \sim m$ (the notation $a \sim b$ means $a \lesssim b \lesssim a$). Specifically, the best known

guarantees that can be obtained, for instance, via localized Rademacher complexity arguments, yield the following upper bound

$$\mathbf{E}R(\widehat{w}_b^{\mathrm{ERM}}) - \inf_{w \in \mathcal{W}_b} R(w) \lesssim \frac{dm^2}{n} + d \cdot \frac{r^2 b^2}{n}. \tag{5.4}$$

We note that an overlooked aspect of the work [189] is that the lower-bound (5.3) proved for *proper* algorithms is matched there via the *improper* Vovk-Azoury-Warmuth (VAW) forecaster [215, 11] . Among the proper algorithms, least squares is arguably the most natural and most extensively studied candidate that could potentially match the lower bound (5.3) (as conjetured by Shamir). Thus, a natural reformulation of Shamir's conjecture arises:

> Provided that the covariate vectors and the response variable are bounded almost surely, is the constrained least squares estimator $\widehat{w}_b^{\mathrm{ERM}}$ optimal among all (potentially non-linear) estimators in a sense that it always matches the lower bound (5.3)?

We address this question by showing that there exist bounded distributions inducing a multiplicative $\sqrt{d}$ gap between the excess risk achievable by the constrained least squares estimator $\widehat{w}_b^{\mathrm{ERM}}$ and that achievable via non-linear predictors. This will be established in the regime where $m \sim 1$ and $br \sim \sqrt{d}$; as we shall see, this is the only natural regime where proper prediction procedures can compete with improper predictors since improper predictors are not sensitive to the scaling of r and b (see Section 5.4.3).

5.3 Summary of Contributions

When only boundedness of the data generating distribution is assumed, we establish that the least squares estimator constrained to a bounded Euclidean ball does not attain the classical d/n excess risk rate, where d is the dimension of the covariates and n is the number of samples. In particular, we construct a bounded distribution such that the constrained least squares estimator incurs an excess risk of order $d^{3/2}/n$ hence refuting the conjecture of Shamir [189]. In contrast, via known results in the literature, we observe that non-linear predictors can achieve the optimal rate d/n assuming that the response variable is bounded, but without any assumptions on the distribution of the covariates.

It is important to highlight that this statistical gap holds despite performing ERM over a *convex* and *bounded* function class with respect to the *squared loss*, a setting considered to be favorable in the literature (see, e.g., [106, Chapter 5]). In particular, the Bernstein condition (cf. Section 2.5.1) is always satisfied in our setup, which is known to imply fast rates for least squares in the bounded setup whenever the underlying class is not too complex. Our work identifies a contrasting scenario: we find that the least squares algorithm is suboptimal for a convex problem, and as such, the failure of the least squares procedure cannot be attributed to complex/non-convex structure of the underlying class (e.g., as is the case for the model selection aggregation problem, where proper estimators fail due to the non-convexity of the underlying class).

As already mentioned in the previous section, the upper bound (5.4) can be readily obtained via the classical localized Rademacher complexity arguments or the more recently introduced offset Rademacher complexity (see Section 2.5 for background). Crucially, the lower bound obtained by Shamir (5.3) and the sharpest known upper bound (5.4) differ by a factor of d in the worst case. It appears that the suboptimal dependence on the boundedness constants br and m arises due to an application of the Talagrand's contraction inequality [118]. In particular, when $\|Y\|_{L_\infty(P_r)} \leq m$, the quadratic loss is $2(br+m)$-Lipschitz on $\mathcal{W}_b$, and the constant $(br+m)^2$ propagates into the resulting upper bounds. Issues related to the application of contraction lemma are discussed in greater detail by Mendelson [142]; an approach for avoiding the use of contraction lemma, via localized analysis of the so-called quadratic and multiplier processes, is also discussed therein.

To avoid applying the contraction lemma in our analysis – following the works of Lecué and Mendelson [113], Mendelson [142] – we adopt the proof strategy of splitting the empirical process of the excess loss class into two parts: the multiplier and quadratic processes, which are then localized separately. In this thesis, we depart from the presentation used in the paper [209] and prove our main upper bound using the localization arguments via the offset Rademacher complexity approach (cf. Section 2.5.2 and Chapter 3). To control the offset quadratic process in our setting of bounded distributions (and in the absence of moment-equivalence assumptions, under which the quadratic processes are typically controlled), we adapt to our setting the proof of matrix Bernstein inequality based on Lieb's trace inequality (cf. [196] and [212, Section 5.4]). The multiplier process is controlled via a direct computation.

Despite improving upon the previously sharpest known upper bound (5.4), our established upper bound on the excess risk of constrained least squares estimator

(Theorem 5.1) does not match Shamir's lower bound (5.3). In particular, for some data generating distributions, the localized multiplier process can be substantially larger than Shamir's lower bound (5.3). To prove our main lower bound, we construct a distribution tailored to make the localized multiplier process ill-behaved. Specifically, we construct a distribution that simultaneously violates moment equivalence assumptions on the noise and the statistical leverage distribution of the covariate vectors since otherwise, the multiplier process can be shown to match Shamir's lower bound (5.3) (these details are omitted in the present chapter but can be found in the paper [209, Section 2.4] on which this chapter is based). The proof of our main lower bound relies on a combination of some delicate exact computations and multiple applications of the matrix Chernoff and Bernstein inequalities [196].

Finally, in Section 5.4.3, we comment on a result due to Forster and Warmuth [66] – an estimator that achieves expected excess risk bound of order dm^2/n without any dependence on the magnitude of the constants b and r. The Forster-Warmuth estimator computes its predictions by first computing a prediction of the ordinary least squares estimator, and then shrinking the size of the prediction based on the leverage of the newly observed point. Due to the shrinkage mechanism, the resulting function is non-linear, and in particular, the Forster-Warmuth estimator is improper. The fact that non-linear estimators can achieve the optimal d/n rate with no assumptions on the design matrix opens new avenues for research concerning statistical prediction under heavy-tailed distributions. This will be further explored in Chapter 6 of this thesis.

5.4 Main Results

This section is organized as follows:

1. Section 5.4.1 contains a sharp upper bound on the excess risk of constrained least squares estimator;

2. the main result is presented in Section 5.4.2, where we establish the suboptimality of the constrained least squares estimator;

3. Section 5.4.3 discusses the potential improvements offered by improper estimators.

Before stating our results, we introduce the following additional notation. For a constant $b > 0$, let w_b^* denote any solution minimizing the population risk $R(\cdot)$ over

$\mathcal{W}_b$ and let $\xi = \xi_b$ denote the associated *noise variable*:

$$w_b^* \in \operatorname*{arg\,min}_{w \in \mathcal{W}_b} R(w) \quad \text{and} \quad \xi = \xi_b(X, Y) = Y - \langle w_b^*, X \rangle.$$

Further, for any $\lambda \geq 0$, denote the *regularized sample second moment matrix* by

$$\widehat{\Sigma}_\lambda = \frac{1}{n}\left(\lambda I_d + \sum_{k=1}^{n-1} X_k X_k^\mathsf{T} + X X^\mathsf{T}\right),$$

where we write X instead of X_n in order to simplify the notation in our main results, and correspondingly, we shall write Y and ξ instead of Y_n and ξ_n.

The proofs are deferred to Section 5.5.

5.4.1 Upper Bound

Our first theorem is an upper bound on the excess risk of any constrained least squares estimator $\widehat{w}_b^{\mathrm{ERM}}$. The proof is deferred to Section 5.5.1. We remark that the below upper bound is non-asymptotic, in contrast to the well-known distribution-free upper bound proved by Audibert and Catoni [9, Theorem 2.1]. Also, we remark that in the following theorem we only assume the integrability of the squared response variable instead of boundedness (the boundedness assumption on Y would enter through a further control on the obtained upper bound).

Theorem 5.1. *For any $n, d, b, r > 0$ and any distribution $P_{r,\infty}$ satisfying $\mathbf{E}Y^2 < \infty$ it holds that*

$$\mathbf{E}R(\widehat{w}_b^{\mathrm{ERM}}) - R(w_b^*) \lesssim \inf_{\lambda > 0}\left(\frac{\mathbf{E}\xi^2 X^\mathsf{T}\widehat{\Sigma}_{\lambda r^2}^{-1} X}{n} + \frac{\lambda r^2 b^2}{n}\right) + \frac{r^2 b^2 \log(1+d)}{n}.$$

We now comment on the structure of the above bound using the interpretation of localized multiplier and quadratic processes due to Mendelson [142] (although the proof of the above theorem shown in this thesis uses the offset Rademacher complexity arguments, instead of the localization arguments based on fixed points). Assume for the sake of presentation that $\widehat{\Sigma}_0$ is invertible. Then, with the choice $\lambda = 0$, we may rewrite the above upper bound as follows:

$$\mathbf{E}R(\widehat{w}_b^{\mathrm{ERM}}) - R(w_b^*) \lesssim \underbrace{\frac{\mathbf{E}\xi^2 X^\mathsf{T}\widehat{\Sigma}_0^{-1} X}{n}}_{\text{Interaction between the noise and covariates}} + \underbrace{\frac{r^2 b^2 \log(1+d)}{n}}_{\text{Low-noise complexity}}. \quad (5.5)$$

The first term, which arises from the supremum of the localized multiplier process, shows the correlation between the noise ξ and the *statistical leverage score* $X^\mathsf{T}\widehat{\Sigma}_0^{-1}X$

(let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the matrix with the i-th row equal to X_i; we may write $X^\mathsf{T}\widehat{\Sigma}_0^{-1}X = H_{nn}$, where $H \in \mathbb{R}^{n \times n}$ is the "hat matrix" defined as $H = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$). If $d \leq n$ and the noise random variable ξ is independent of X, then the first term in (5.5) corresponds essentially to the minimax optimal rate for unconstrained least squares regression [152, Theorem 2] and is hence unimprovable in general. The second term in (5.5), which arises from the supremum of the localized quadratic process, intuitively captures the problem complexity in low-noise regimes, that is, when ξ is relatively small. See the paper on which this chapter is based [209, Proposition 2.3] for a demonstration of a *noiseless* problem such that for *some* constrained least squares estimator, the second term in (5.5) appears in a lower bound on its excess risk. Hence, the second term above is also, in general, unimprovable.

Various assumptions considered in the literature, under which least squares estimators are analyzed, impose a kind of independence between the leverage scores and the noise random variables. Such assumptions, in turn, allow obtaining bounds on the correlation between the leverage scores and between the noise that scale as the classical rate d/n. See [209, Section 2.4] for an extended discussion.

5.4.2 Lower Bound

In this section, we present our main result: a construction of a bounded distribution under which the constrained least squares estimator exceeds Shamir's lower bound (5.3) by a factor proportional to $\sqrt{d}$. In what follows, we shall refer to the correlation between the leverages and the noise – the first term in the upper bound presented in Theorem 5.1 – as the *multiplier term*, since as discussed in the previous section, it arises from the localized multiplier process.

If the noise variables and the leverage scores satisfy certain regularity assumptions that allow us to split them apart in the multiplier term, then an excess risk upper bound that matches Shamir's lower bound (5.3) can be obtained [209, Section 2.4]. Among such assumptions are the L_4–L_2 moment equivalence assumptions that essentially allow upper bounding the multiplier term in the upper bound of Theorem 5.1 via an application of Cauchy-Schwarz inequality. Since Bernoulli random variables with a small parameter p satisfy L_4–L_2 moment equivalence with an ill-behaved constant $1/p$, we aim to construct a distribution such that the noise random variables and the leverage scores both approximately follow Bernoulli distributions with a small parameter that depends on the dimension d. Besides, the noise variables and the leverage scores need to be *highly correlated*; otherwise, the multiplier term would be too small to yield the desired lower bound.

We remark that our construction could be considered somewhat extreme only with respect to the constants appearing in the L_4–L_2 moment equivalence on the leverage scores and the noise. At the same time, our distribution presented below is bounded with a favorable choice of constants, and in addition, the Bernstein condition is satisfied. Hence, by the classical upper bound obtainable via local Rademacher complexity arguments (5.4), the constrained least squares estimator satisfies a non-trivial fast rate excess risk guarantee, making the construction of our main lower bound more challenging.

We now present our construction. For simplicity, we assume that $\sqrt{d}$ is an integer in what follows. Let $\mathbf{1} \in \mathbb{R}^d$ denote an all-ones vector. For a support set $S \subseteq \{1, \dots, d\}$, let $\mathbf{1}_S$ denote a vector such that $(\mathbf{1}_S)_i = 1$ if $i \in S$ and 0 otherwise. Let $\mathcal{S}_{\sqrt{d}} = \{S \subseteq \{1, \dots, d\} : |S| = \sqrt{d}\}$. We define the joint distribution on (X, Y) by first defining the marginal distribution of Y as a Bernoulli random variable with parameter $1 - d^{-1/2}$; thus, Y takes value 1 with probability $1 - d^{-1/2}$ and value 0 otherwise. Next, we define the conditional distribution of X given $Y = 1$ via the Dirac measure supported on $d^{-1}\mathbf{1}$, denoted by $\delta_{\{d^{-1}\mathbf{1}\}}$. Finally, we define the conditional distribution of X given $Y = 0$ as a uniform distribution over the set $\{d^{-1/4}\mathbf{1}_S : S \in \mathcal{S}_{\sqrt{d}}\}$, and we denote this distribution by $\mathrm{Uniform}(d^{-1/4}\mathbf{1}_S : S \in \mathcal{S}_{\sqrt{d}})$. Thus, our construction may be summarized as follows:

$$\begin{aligned} Y &\sim \mathrm{Bernoulli}(1 - d^{-1/2}), \\ X|Y = 1 &\sim \delta_{\{d^{-1}\mathbf{1}\}}, \\ X|Y = 0 &\sim \mathrm{Uniform}(d^{-1/4}\mathbf{1}_S : S \in \mathcal{S}_{\sqrt{d}}). \end{aligned} \tag{5.6}$$

A simple calculation shows that $w^*_\infty \approx \frac{1}{2}\mathbf{1}$ and hence, for $b \gtrsim \sqrt{d}$ we have $w^*_b = w^*_\infty$. In particular, the squared noise variable ξ_i^2 is smaller than 1 for the "high probability" points $(X_i, Y_i) = (d^{-1}\mathbf{1}, 1)$, while $\xi_i^2 \approx \sqrt{d}$ for the "low probability" points $(X_i, Y_i) = (d^{-1/4}\mathbf{1}_S, 0)$. This establishes that ξ_i^2 behaves as Bernoulli random variables. Similarly, since all the "high probability" points are exactly the same, they essentially have zero leverage. On the other hand, the "low probability" points all have high leverage, thus the leverage scores also approximately follow the Bernoulli distribution. Finally, since ξ_i^2 is large exactly for the high leverage points, the squared noise random variables are correlated with the leverage scores. Intuitively, the multiplier term (i.e., the first term in Theorem 5.1) scales as $d^{3/2}/n$ under the distribution (5.6), while Shamir's lower bound (5.3) scales only as d/n provided that $b \sim \sqrt{d}$.

The main result of this chapter is presented below. The proof is deferred to Section 5.5.2.

Theorem 5.2. *Suppose that the distribution P of (X, Y) is given by (5.6); observe that for this distribution we have $r = 1$ and $m = 1$. Then, for any constrained least squares estimator $\widehat{w}_b^{\mathrm{ERM}}$, the following lower bound holds, provided that d is larger than some absolute constant, $b \sim \sqrt{d}$ and $n \gtrsim d^3 \log d$:*

$$\mathbf{E} R(\widehat{w}_b^{\mathrm{ERM}}) - R(w_b^*) \gtrsim \frac{d^{3/2}}{n} \sim \sqrt{d} \cdot \underbrace{\left(\frac{dm^2}{n} + \frac{r^2 b^2}{n} \right)}_{\text{Shamir's lower bound (5.3)}}.$$

We now comment on the above result. Recall that the aim of the construction (5.6) is to maximize the multiplier term under boundedness constraints on the underlying distribution. In view of Shamir's lower bound (5.3), the parameters m, r, b are chosen in the most relevant way in the sense explained below. First, because of the homogeneity of the excess risk for least squares estimator, we may always set $m = 1$ – scaling the response variables by m affects the excess risk of the ordinary least squares estimator by a factor of m^2. Second, the choice $b \sim \sqrt{d}$ is natural for d-dimensional vectors as it allows each coordinate to be of a constant order, particularly, for the underlying parameter w_b^*. Finally, the scaling $dm^2 \sim r^2 d^2$ equalizes the two terms in Shamir's lower bound (5.3) and according to the results discussed in Section 5.4.3, leaves open the possibility that in such regimes improper estimators offer no statistical improvements. Indeed, the best upper bound for non-linear estimators scales as dm^2/n. The above theorem shows that even in the most favorable regime for proper estimators, the constrained least squares incurs an extra factor of $\sqrt{d}$ in its excess risk, in comparison to the excess risk achievable via improper estimators.

Finally, observe that the sample complexity $n \gtrsim d^3 \log d$ appearing in the above lower bound can be improved at most to $n \gtrsim d^2$. To see that, by optimizing the λ term in the upper bound presented in Theorem 5.1, it is possible to obtain a "slow rate" upper bound of order $brm/\sqrt{n}$ (see [Section 2.4][209]). When $r = m = 1$ and $b = \sqrt{d}$, this upper bound reduces to $\sqrt{d}/\sqrt{n}$. Hence, to obtain a lower bound of order $d^{3/2}/n$, we need $n \gtrsim d^2$. It is not clear if our techniques are sufficient to imply to this weaker assumption on the sample complexity.

5.4.3 Separating Proper and Improper Predictors via Known Results

In this section, we observe that non-linear predictors can surpass Shamir's lower bound (5.3) that holds for proper estimators. In particular, an estimator due to Forster

and Warmuth [66][1], which is based on the Vovk-Azoury-Warmuth (VAW) forecaster [215, 11], satisfies an excess risk bound independent of the constants r and b.

We now define the Forster-Warmuth forecaster. Given a data sample S_n, denote by $\widehat{w}^{(\mathrm{ERM})}$ the ordinary least squares solution computed via the following expression:

$$\widehat{w}^{(\mathrm{ERM})} = \left(\sum_{i=1}^{n} X_i X_i^{\mathsf{T}}\right)^{\dagger} \left(\sum_{i=1}^{n} X_i Y_i\right), \tag{5.7}$$

where the notation $A^{\dagger}$ denotes the Moore-Penrose inverse of a matrix A. Further, keeping the observed data sample S_n fixed, for any $X \in \mathbb{R}^d$ define its leverage by

$$h_X = X^{\mathsf{T}} \left(\sum_{i=1}^{n} X_i X_i^{\mathsf{T}} + X X^{\mathsf{T}}\right)^{\dagger} X.$$

The Forster-Warmuth predictor $\widehat{f}^{(\mathrm{FW})}$ is then defined pointwise as follows:

$$\widehat{f}^{(\mathrm{FW})}(X) = (1 - h_X)^2 \left\langle \widehat{w}^{(\mathrm{ERM})}, X \right\rangle. \tag{5.8}$$

Thus, the above function outputs the predictions of the ordinary least squares estimator computed via (5.7), with an extra shrinkage factor that depends on the leverage of a new point X on which the prediction is to be computed. It is important that the shrinkage term is quadratic – should we only shrink by $(1 - h_X)$, the resulting prediction procedure would correspond to the VAW forecaster, for which excess risk upper bound independent of the scale of r and b is not known. However, the VAW forecaster (with an online-to-batch conversion) also surpasses the Shamir's lower bound (5.3); see [209, Section 3.1] for details.

Intuitively, the Forster-Warmuth predictor (5.8) avoids making large errors for high leverage points. In the view of the upper bound presented in Theorem 5.1, such a modification is desirable, as it makes the multiplier term small. We now state a result due to Forster and Warmuth [66].

Theorem 5.3 (Forster and Warmuth [66])**.** *Let $\widehat{f}^{(\mathrm{FW})}(\cdot)$ denote the non-linear predictor defined in (5.8). Let P be any distribution (with possibly unbounded covariates) satisfying $\|Y\|_{L_\infty} \leq m$. Then, for any $d, n > 0$, the following holds:*

$$\mathbf{E} R(\widehat{f}^{(\mathrm{FW})}(X)) - \inf_{w \in \mathbb{R}^d} R(w) \leq \frac{2dm^2}{n}. \tag{5.9}$$

[1]We are thankful to Manfred Warmuth for pointing us to the Forster-Warmuth algorithm.

Observe that the bound (5.9) is closely related to the upper bound (5.1) that holds for unconstrained least squares in the well-specified setup with Gaussian design: both bounds do not depend on the magnitude of the covariates, specific properties of the covariance structure, and the norm of the optimal linear predictor $w^* = \arg\inf_{w\in\mathbb{R}^d} R(w)$. The difference is that $R(w^*)$ is replaced by m^2. It is reported in [66] that the authors could not prove a bound similar to (5.9) for the least squares estimator. Given the discussions of this section is it not surprising – the least squares estimator is proper and it is sensitive to the constants r and b, unlike the Forster-Warmuth estimator.

5.5 Proofs

This section contains proofs of Theorem 5.1 (Section 5.5.1) and Theorem 5.2 (Section 5.5.2). Some technical lemmas are deferred to Section 5.5.3. Let us now recall some notation. The notation $\|\cdot\|$ denotes the Euclidean norm for vectors and the operator norm for matrices. The $d \times d$ identity matrix is denoted by I_d. The notation $\sigma = (\sigma)_{i=1}^n$ denotes a sequence of i.i.d. Rademacher random variables (symmetric $\{\pm\}$-valued random variables); further, $\mathbf{E}_\sigma$ denotes expectation computed with respect to Rademacher random variables only, conditionally on all other random variables. We will sometimes write w instead of $\langle w, \cdot\rangle$. For example, ℓ_w denotes $\ell_{\langle w,\cdot\rangle}$. Finally, when the above convention would introduce confusion, we will use the notation the notation $f_w = \langle w, \cdot\rangle$.

5.5.1 Proof of Theorem 5.1

This section is devoted to the proof of Theorem 5.1. First, observe that by convexity of $\mathcal{W}_b$ and the strong convexity of the quadratic loss, the empirical risk minimization estimator satisfies the deterministic offset condition (cf. Definition 3.1) stating that the following inequality holds with probability one:

$$R_n(\widehat{w}_b^{\mathrm{ERM}}) - R_n(w_b^*) \le -\frac{1}{n}\sum_{i=1}^n \left\langle \widehat{w}_b^{\mathrm{ERM}} - w_b^*, X_i \right\rangle^2 = -P_n(f_{\widehat{w}_b^{\mathrm{ERM}}} - f_{w_b^*})^2. \tag{5.10}$$

We may thus proceed with the offset Rademacher complexity proof technique (cf. Section 2.5.2 , Chapter 3):

$$\begin{aligned}
&\mathbf{E}R_n(\widehat{w}_b^{\mathrm{ERM}}) - R_n(w_b^*) \\
&= \mathbf{E}\left[(P - P_n)(\ell_{\widehat{w}_b^{\mathrm{ERM}}} - \ell_{w_b^*}) + P_n(\ell_{\widehat{w}_b^{\mathrm{ERM}}} - \ell_{w_b^*})\right] \\
&\le \mathbf{E}\left[(P - P_n)(\ell_{\widehat{w}_b^{\mathrm{ERM}}} - \ell_{w_b^*}) - P_n(f_{\widehat{w}_b^{\mathrm{ERM}}} - f_{w_b^*})^2\right],
\end{aligned}$$

where the last step follows from the offset condition (5.10). Following Mendelson [142], we separately localize (in our case, via the offset approach) the *quadratic* and *multiplier* components using the identity

$$\ell_w(X,Y) - \ell_{w^*}(X,Y) = (f_w(X) - f_{w^*}(X))^2 - 2(f_w(X) - f_{w^*}(X)) \underbrace{(Y - f_{w^*}(X))}_{=\xi(X,Y)}.$$

Hence, we can continue the above derivations as follows (also following along the lines of [120, Proof of Theorem 3]):

$$\begin{aligned}
&\mathbf{E} R_n(\widehat{w}_b^{\mathrm{ERM}}) - R_n(w_b^*) \\
&\le \mathbf{E}\left[\frac{3}{2}(P - P_n)\left(f_{\widehat{w}_b^{\mathrm{ERM}}} - f_{w_b^*}\right)^2 - \frac{1}{4}(P + P_n)\left(f_{\widehat{w}_b^{\mathrm{ERM}}} - f_{w_b^*}\right)^2\right] \\
&\quad + \mathbf{E}\left[(P_n - P)2\xi(f_{\widehat{w}_b^{\mathrm{ERM}}} - f_{w_b^*}) - \frac{1}{4}(P + P_n)\left(f_{\widehat{w}_b^{\mathrm{ERM}}} - f_{w_b^*}\right)^2\right] \\
&\le \frac{3}{2}\mathbf{E}\left[\sup_{w \in \mathcal{W}_b}\left\{(P - P_n)\left(f_w - f_{w_b^*}\right)^2 - \frac{1}{6}(P + P_n)\left(f_w - f_{w_b^*}\right)^2\right\}\right] \\
&\quad + 2\mathbf{E}\left[\sup_{w \in \mathcal{W}_b}\left\{(P_n - P)\xi(f_w - f_{w_b^*}) - \frac{1}{8}(P + P_n)\left(f_w - f_{w_b^*}\right)^2\right\}\right] \\
&\le 3\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w \in \mathcal{W}_b}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i \langle w - w_b^*, X_i\rangle^2 - \frac{1}{6}\frac{1}{n}\sum_{i=1}^n \langle w - w_b^*, X_i\rangle^2\right\}\right] \\
&\quad + 4\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w \in \mathcal{W}_b}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i \langle w - w_b^*, X_i\rangle - \frac{1}{8}\frac{1}{n}\sum_{i=1}^n \langle w - w_b^*, X_i\rangle^2\right\}\right] \\
&\le 3\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w \in \mathcal{W}_{2b}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i \langle w, X_i\rangle^2 - \frac{1}{6} w^\top \widehat{\Sigma} w\right\}\right] \qquad \text{(quadratic term)} \\
&\quad + 4\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w \in \mathcal{W}_{2b}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i \langle w, X_i\rangle - \frac{1}{8} w^\top \widehat{\Sigma} w\right\}\right], \qquad \text{(multiplier term)}
\end{aligned}$$

where the penultimate line follows by a classical symmetrization argument and the last line follows by noting that $\|w - w_b^*\| \le 2b$ for $w \in \mathcal{W}_b$. Also recall that $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$.

In the rest of this proof we show how to control the quadratic and multiplier terms. We begin with the multiplier term since it is more straightforward to upper bound.

Bounding the multiplier term. We introduce a tuning parameter $\lambda > 0$. By subtracting and adding $\frac{\lambda r^2}{n}\lambda r^2 \frac{1}{8} w^\top w$ inside the supremum, we obtain

$$\text{(multiplier term)} \le 4\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w \in \mathcal{W}_{2b}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\xi_i \langle w, X_i\rangle - \frac{1}{8} w^\top \widehat{\Sigma}_{r\lambda} w\right\}\right] + \frac{2\lambda r^2 b^2}{n}.$$

We will now control the term involving the supremum. A direct computation yields

$$\begin{aligned}
&\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w\in\mathcal{W}_{2b}}\left\{\frac{1}{n}\sum_{i=1}^n\sigma_i\xi_i\left\langle w,X_i\right\rangle-\frac{1}{8}w^\top\widehat{\Sigma}_{r\lambda}w\right\}\right]\\
&\leq\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w\in\mathbb{R}^d}\left\{\left\langle w,\frac{1}{n}\sum_{i=1}^n\sigma_i\xi_iX_i\right\rangle-\frac{1}{8}w^\top\widehat{\Sigma}_{r\lambda}w\right\}\right]\\
&=2\mathbf{E}\mathbf{E}_\sigma\left[\left(\frac{1}{n}\sum_{i=1}^n\sigma_i\xi_iX_i\right)^\top\widehat{\Sigma}_{r\lambda}^{-1}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i\xi_iX_i\right)\right]\\
&=2\mathbf{E}\mathbf{E}_\sigma\left[\left(\frac{1}{n}\sum_{i=1}^n\sigma_i\xi_iX_i\right)^\top\widehat{\Sigma}_{r\lambda}^{-1}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i\xi_iX_i\right)\right]\\
&=2\mathbf{E}\left[\frac{1}{n^2}\sum_{i=1}^n\xi_i^2X_i^\top\widehat{\Sigma}_{r\lambda}^{-1}X_i\right]\\
&=\frac{2}{n}\mathbf{E}\left[\xi_n^2X_n^\top\widehat{\Sigma}_{r\lambda}^{-1}X_n\right],
\end{aligned}$$

where the last line follows by the fact that X_i,ξ_i are identically distributed. Recalling that $\lambda>0$ is arbitrary, this concludes the first part of the proof, yielding the first term in the upper bound of Theorem 5.1.

Bounding the quadratic term. In order to control the quadratic term, we follow along the lines of the proof of matrix-Bernstein inequality (particularly, the presentation of Vershynin [212, Section 5.4]). Also, we adapt some ideas from the proof of Rudelson's inequailty for sum of rank one operators due to Oliveira [162].

In what follows, any function $f:\mathbb{R}\to\mathbb{R}$ applied to a $d\times d$ symmetric matrix A with spectral decomposition $A=\sum_{i=1}^n\lambda_iu_iu_i^\top$ acts on its spectrum. That is, we let $f(A)=\sum_{i=1}^n f(\lambda_i)u_i^\top u_i$, where $u_1,\ldots,u_d$ form an orthonormal eigenbasis of A. In addition, the inequality sign $\succcurlyeq$ induces a positive-semidefinite partial order on the set of $d\times d$ symmetric matrices. In particular, we write $A\succcurlyeq B$ if and only if $A-B\succcurlyeq 0$, meaning that $A-B$ is positive-semidefinite (equivalently, the eigenvalues of $A-B$ are non-negative). For more details on the above setup we refer to Vershynin [212, Section 5.4.1].

As a first step, we reformulate the quadratic term in the language of matrix

eigenvalues. Letting $\lambda_{\max}(A)$ denote the largest eigenvalue of the matrix A we have

$$
\begin{aligned}
&\text{(quadratic term)} \\
&= 3\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w\in\mathcal{W}_{2b}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i \langle w, X_i\rangle^2 - \frac{1}{6}w^\top\widehat{\Sigma}w\right\}\right] \\
&= 12b^2\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w\in\mathcal{W}_1}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i \langle w, X_i\rangle^2 - \frac{1}{6}w^\top\widehat{\Sigma}w\right\}\right] \\
&= \frac{12b^2}{n}\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w\in\mathcal{W}_1}\left\{w^\top\left(\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)w\right\}\right] \\
&= \frac{12b^2}{n}\mathbf{E}\mathbf{E}_\sigma\left[\max\left\{0, \lambda_{\max}\left(\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right\}\right]
\end{aligned}
$$

Recall that $\mathbf{E}_\sigma[]$ is a shorthand for $\mathbf{E}_\sigma[\cdot|X_1,\ldots,X_n]$. We will now show how to control the inner expectation in the above equation. Introducing a tuning parameter $\mu > 0$, we have by Jensen's inequality

$$
\begin{aligned}
&\mathbf{E}_\sigma\left[\max\left\{0, \lambda_{\max}\left(\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right\}\right] \\
&\le \frac{1}{\mu}\log\mathbf{E}_\sigma\left[\exp\left(\mu\max\left\{0, \lambda_{\max}\left(\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right\}\right)\right] \\
&\le \frac{1}{\mu}\log\left(1 + \mathbf{E}_\sigma\left[\exp\left(\mu\lambda_{\max}\left(\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right)\right]\right).
\end{aligned}
$$

Next, using the bounding the maximum eigenvalue via trace (which is true when all eigenvalues are non-negative) we have

$$
\begin{aligned}
&\mathbf{E}_\sigma\left[\exp\left(\mu\lambda_{\max}\left(\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right)\right] \\
&= \mathbf{E}_\sigma\left[\lambda_{\max}\left(\exp\left(\mu\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right)\right] \\
&= \mathbf{E}_\sigma\left[\mathrm{Trace}\left(\exp\left(\mu\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right)\right].
\end{aligned}
$$

Using the subadditivity property of matrix cumulant generating functions that can be established by a repeated application of Lieb's inequality, we have by Tropp [196, Lemma 3.5.1]

$$
\begin{aligned}
&\mathbf{E}_\sigma\left[\mathrm{Trace}\left(\exp\left(\mu\sum_{i=1}^n \sigma_i X_i^\top X_i - \frac{1}{6}X_iX_i^\top\right)\right)\right] \\
&\le \mathrm{Trace}\exp\left(\sum_{i=1}^n \log\mathbf{E}_{\sigma_i}\left[\exp\left(\mu\sigma_i X_iX_i^\top - \frac{\mu}{6}X_iX_i^\top\right)\right]\right).
\end{aligned}
$$

Now comes the part where we will be able to exploit the negative terms that arise from the offset condition. Using the fact that $\sigma_i X_i X_i^\top$ and $X_i X_i^\top$ commute, we may decompose the exponential terms into products, and products inside logarithms into sum of logarithms (i.e., if A and B commute, then $e^{A+B} = e^A e^B$ and $\log(AB) = \log(A) + \log(B)$). Then, we have

$$\begin{aligned}
&\operatorname{Trace}\exp\left(\sum_{i=1}^{n}\log \mathbf{E}_{\sigma_i}\left[\exp\left(\mu\sigma_i X_i X_i^\top - \frac{\mu}{6}X_i X_i^\top\right)\right]\right)\\
&= \operatorname{Trace}\exp\left(\sum_{i=1}^{n}\log\left(\mathbf{E}_{\sigma_i}\left[\exp\left(\mu\sigma_i X_i X_i^\top\right)\right]\exp\left(-\frac{\mu}{6}X_i X_i^\top\right)\right)\right)\\
&= \operatorname{Trace}\exp\left(\sum_{i=1}^{n} -\frac{\mu}{6}X_i X_i^\top + \log \mathbf{E}_{\sigma_i}\left[\exp\left(\mu\sigma_i X_i X_i^\top\right)\right]\right)\\
&\leq \operatorname{Trace}\exp\left([\mu^2 r^2/2 - \mu/6]\sum_{i=1}^{n} X_i X_i^\top\right),
\end{aligned}$$

where the last line follows from applying a Hoeffding-type moment generating function bound [196, Lemma 4.6.3], which combined with the bound $\|X_i\|_2 \leq r$ yields:

$$\log \mathbf{E}_{\sigma_i}\left[\exp\left(\mu\sigma_i X_i X_i^\top\right)\right] \preccurlyeq \frac{1}{2}\mu^2 (X_i X_i^\top)^2 = \frac{1}{2}\mu^2 \|X_i\|_2^2 X_i X_i^\top \preccurlyeq \frac{1}{2}r^2\mu^2 X_i X_i^\top.$$

Taking $\mu = 1/r^2$ (so that $\mu^2 r^2/2 - \mu/6 = 0$) and putting everything together, we have

$$\begin{aligned}
&\text{(quadratic term)}\\
&= \frac{12b^2}{n}\mathbf{E}\mathbf{E}_\sigma\left[\sup_{w\in\mathcal{W}_1}\left\{w^\top\left(\sum_{i=1}^{n}\sigma_i X_i^\top X_i - \frac{1}{6}X_i X_i^\top\right)w\right\}\right]\\
&\leq \frac{36b^2r^2}{n}\mathbf{E}\log\left(1 + \operatorname{Trace}\exp(0_{d\times d})\right)\\
&= \frac{36b^2r^2}{n}\log\left(1 + \operatorname{Trace} I_d\right)\\
&= \frac{36b^2r^2}{n}\log\left(1 + d\right).
\end{aligned}$$

The proof is complete. □

5.5.2 Proof of Theorem 5.2

The proof of this result is split into several steps. In this section, we provide three technical lemmas and demonstrate how they imply the result of Theorem 5.2. The first two lemmas are based on exact computations using the Sherman-Morrison formula. The proof of the third lemma, for which we sketch a simple heuristic argument before presenting the formal proof, is based on matrix concentration inequalities. We

always assume that d (and therefore n, since it satisfies $n \gtrsim d^3 \log d$) is large enough. Within the proofs, we use auxiliary variables α, β, x, y, that are sometimes redefined throughout the text.

Before we proceed, let us remark that (X, Y) distributed according to (5.6) satisfies $\|X\| \leq 1$ almost surely and $\|Y\|_{L_\infty} \leq 1$, thus $r = m = 1$. Our first lemma, proved in Appendix 5.5.3.1, provides an excess risk lower bound for any vector $w \in \mathbb{R}^d$, provided that b is large enough.

Lemma 5.1. *Suppose that $b \geq \sqrt{d}/2, d \geq 4$, and (X, Y) is distributed according to (5.6). Then, for any $w \in \mathbb{R}^d$ we have*

$$R(w) - R(w_b^*) \geq \frac{1}{2} d^{-3/2} \|w - w_b^*\|^2 \quad \text{and also} \quad w_b^* = \frac{\sqrt{d} - 1}{2\sqrt{d} - 1} \cdot \mathbf{1}.$$

Further, we define an unconstrained least squares solution as (dropping the superscript ERM in our notation):

$$\widehat{w}_\infty = \left(n\widehat{\Sigma}\right)^{-1} \left(\sum_{i=1}^n X_i Y_i\right). \tag{5.11}$$

In the proof of Theorem 5.2 we work on the event where $\widehat{\Sigma}$ is invertible which will be shown to hold with sufficient probability. This ensures the uniqueness of $\widehat{w}_\infty$ hence, we remark that the result of Theorem 5.2 holds for *any* constrained least squares estimator. Our proof strategy is quite straightforward: using Lemma 5.1 we show that the excess risk of $\widehat{w}_\infty$ is lower bounded by $cd^{3/2}/n$, while for large enough b, $\widehat{w}_\infty$ is also a least squares solution constrained to the ball of radius b. Before stating our next lemma we introduce some additional notation. Let

$$I = \{i \in \{1, \ldots, n\} : X_i \neq \mathbf{1}/d\}$$

denote the (random) subset of data points whose covariates are not equal to $\mathbf{1}/d$. Denote

$$A = \sum_{i \in I} X_i X_i^\mathsf{T} \quad \text{and hence} \quad \widehat{\Sigma} = \frac{1}{n}\left((n - |I|)\, d^{-2} \mathbf{1}\mathbf{1}^\mathsf{T} + A\right). \tag{5.12}$$

Further, let $v, \zeta \in \mathbb{R}^d$ denote the (random) vectors such that

$$v_i = A_{ii}\sqrt{d} \quad \text{and} \quad \zeta = v - |I|\, d^{-1/2}\mathbf{1}. \tag{5.13}$$

In words, the i-th entry of v denotes the number of observations in the set I whose i-th entry is non-zero. Observe that conditionally on the size of the set I, we have $\mathbf{E}(v \mid |I|) = |I|\, d^{-1/2}\mathbf{1}$ and hence, ζ represents the noise present in the counts vector v.

We will repeatedly rely on the following identities, which can be shown via a simple counting argument:

$$A\mathbf{1} = v = |I|\, d^{-1/2}\mathbf{1} + \zeta \quad \text{and} \quad \langle \zeta, \mathbf{1} \rangle = 0. \tag{5.14}$$

The following lemma provides a sharp inequality for the norm of $\widehat{w}_\infty$ as well as an exact expression for the vector $\widehat{w}_\infty$ itself. The proof is deferred to Appendix 5.5.3.2.

Lemma 5.2. *Let $\widehat{w}_\infty$ be defined by (5.11). The following two identities hold whenever the matrix A defined in (5.12) is invertible:*

$$\widehat{w}_\infty = \frac{d^{3/2}\,|I|^{-1}}{(n-|I|)^{-1}d^2 + \mathbf{1}^\mathsf{T} A^{-1}\mathbf{1}}\mathbf{1} - \frac{d^{3/2}\,|I|^{-1}}{(n-|I|)^{-1}d^2 + \mathbf{1}^\mathsf{T} A^{-1}\mathbf{1}} A^{-1}\zeta, \tag{5.15}$$

and

$$\|\widehat{w}_\infty\|^2 \le n^2 d^{-2}\mathbf{1}^\mathsf{T} A^{-2}\mathbf{1}.$$

Note that the first summand in (5.15) as well as the vector w_b^* are both proportional to $\mathbf{1}$. However, it will be shown later that the second summand in (5.15), which is proportional to $A^{-1}\zeta$, is almost orthogonal to $\mathbf{1}$. Combining this observation with the fact that $\langle \mathbf{1}, \zeta \rangle = 0$ will yield the desired lower bound via Lemma 5.1, provided that the magnitude of the second term in (5.15) is large enough.

Combining Lemmas 5.1 and 5.2, the excess risk of the unconstrained least squares solution $\widehat{w}_\infty$ can be expressed in terms of the random quadratic form $\mathbf{1}^\mathsf{T} A^{-1}\mathbf{1}$ and the random vector $A^{-1}\zeta$. Also, the norm of $\widehat{w}_\infty$ can be upper-bounded in terms of $\mathbf{1}^\mathsf{T} A^{-2}\mathbf{1}$. The following result provides sharp bounds on all the random quantities that we need.

Lemma 5.3. *Suppose that d is large enough and $n \gtrsim d^3 \log d$. Then, the following results hold simultaneously, with probability at least $1/2$:*

(a) $|I| \sim nd^{-1/2}$;

(b) $\|\zeta\|^2 \sim n$;

(c) The matrix A defined by (5.12) is invertible;

(d) $\zeta^\mathsf{T} A^{-1}\zeta \gtrsim d^{3/2}$;

(e) $\mathbf{1}^\mathsf{T} A^{-1}\mathbf{1} \lesssim n^{-1}d^2$;

(f) $\mathbf{1}^\mathsf{T} A^{-2}\mathbf{1} \lesssim n^{-2}d^3$.

Before presenting the formal proof (see Appendix 5.5.3.3), we discuss the intuition behind the proof of this lemma. First, observe that (a) follows from the fact that $|I|$ is Binomially distributed with parameters n, $d^{-1/2}$, and so $\mathbf{E}|I| = nd^{-1/2}$. The magnitude of $\|\zeta\|^2$ follows from a direct computation of its expectation and variance. For large enough n, we expect that $A \approx \mathbf{E}A$. Assuming this, we may focus on $\mathbf{E}A = \mathbf{E}\mathbf{E}(A \mid |I|)$, which has the following simple form:

$$\mathbf{E}A = \mathbf{E}\mathbf{E}(A \mid |I|) = \mathbf{E}\,|I| \left((d^{-1} - (d^{3/2}+d)^{-1})I_d + (d^{3/2}+d)^{-1}\mathbf{1}\mathbf{1}^\mathsf{T}\right).$$

Observe that the eigenvector corresponding to the largest eigenvalue of $\mathbf{E}A$ is proportional to $\mathbf{1}$, and the remaining eigenvectors complement this direction and form an orthonormal basis. Moreover, the above expression for $\mathbf{E}A$ implies that $\lambda_1(\mathbf{E}A) \sim \mathbf{E}\,|I|\,d^{-1/2} = nd^{-1}$ and $\lambda_j(\mathbf{E}A) \sim \mathbf{E}\,|I|\,d^{-1} = nd^{-3/2}$ for $j = 2, \ldots, d$. Thus, $\mathbf{E}A$ is invertible and in particular, we have

$$\mathbf{1}^\mathsf{T}(\mathbf{E}A)^{-1}\mathbf{1} = d/\lambda_1(\mathbf{E}A) \lesssim n^{-1}d^2, \quad \text{and} \quad \mathbf{1}^\mathsf{T}(\mathbf{E}A)^{-2}\mathbf{1} = d/(\lambda_1(\mathbf{E}A))^2 \lesssim n^{-2}d^3.$$

Finally, since by (5.14) we have $\langle \zeta, \mathbf{1}\rangle = 0$, the random vector ζ is orthogonal to the first eigenvalue of $\mathbf{E}A$. Therefore, the following inequality holds with probability one:

$$\zeta^\mathsf{T}(\mathbf{E}A)^{-1}\zeta \geq \|\zeta\|^2/\lambda_2(\mathbf{E}A) \gtrsim d^{3/2}.$$

With the above lemmas at hand, we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. We work on the event of Lemma 5.3. First, note that combining Lemmas 5.2 and 5.3 we have

$$\|\widehat{w}_\infty\|^2 \lesssim d.$$

Thus, on the event of Lemma 5.3, the unconstrained ERM solution $\widehat{w}_\infty$ is also a solution over the Euclidean ball of any radius b that satisfies $b \geq c\sqrt{d}$, where c is some absolute constant.

We will now lower bound the expected excess risk of $\widehat{w}_\infty$. Observe that for any vector x and a unit vector u we have $\|x\| \geq |\langle x, u\rangle|$. Consider the unit vector $u = \zeta/\|\zeta\|$. Denote

$$\beta = \frac{d^{3/2}\,|I|^{-1}}{(n-|I|)^{-1}d^2 + \mathbf{1}^\mathsf{T}A^{-1}\mathbf{1}}.$$

Combining Lemmas 5.1 and 5.2 together with $\langle \zeta, \mathbf{1} \rangle = 0$ given by (5.14) we have

$$\begin{aligned} R(\widehat{w}_\infty) - R(w_b^*) &\geq \frac{1}{2} d^{-3/2} \left\| \left(\beta - \frac{\sqrt{d}-1}{2\sqrt{d}-1} \right) \mathbf{1} - \beta A^{-1} \zeta \right\|^2 \\ &\geq \frac{1}{2} d^{-3/2} \left(\left\langle \frac{\zeta}{\|\zeta\|}, \left(\beta - \frac{\sqrt{d}-1}{2\sqrt{d}-1} \right) \mathbf{1} - \beta A^{-1} \zeta \right\rangle \right)^2 \\ &= \frac{1}{2} d^{-3/2} \left(\left\langle \frac{\zeta}{\|\zeta\|}, \beta A^{-1} \zeta \right\rangle \right)^2 . \end{aligned} \tag{5.16}$$

By Lemma (5.3) we have $\beta \gtrsim 1$, with probability at least $\frac{1}{2}$. Hence, the lower bound (5.16) implies on the event of Lemma 5.3 that

$$R(\widehat{w}_\infty) - R(w_b^*) \geq \frac{1}{2} d^{-3/2} \beta^2 \left(\frac{\zeta^\mathsf{T} A^{-1} \zeta}{\|\zeta\|} \right)^2 \gtrsim \frac{d^{3/2}}{n}.$$

Since the event of Lemma 5.3 holds with probability at least $\frac{1}{2}$, it follows that $\mathbf{E} R(\widehat{w}_\infty) - R(w_b^*) \gtrsim \frac{d^{3/2}}{n}$. This concludes the proof of our theorem. □

5.5.3 Proofs of Technical Lemmas

This section contains proofs of technical lemmas used in the proof of Theorem 5.2. Before presenting the proofs, let us recall the *Bernstein condition*, which follows from the convexity of $\mathcal{W}_b$ and the strong convexity of the quadratic loss:

$$\text{for any } w \in \mathcal{W}_b \text{ it holds that } R(w) - R(w_b^*) \geq \mathbf{E} \langle w - w_b^*, X \rangle^2. \tag{5.17}$$

5.5.3.1 Proof of Lemma 5.1

The proof is split into two steps. We first compute $w_\infty^* = \inf_{w \in \mathbb{R}^d} R(w)$ and show that $\|w_\infty^*\| \leq \sqrt{d}/2$ so that $w_\infty^* = w_b^*$ whenever $b \geq \sqrt{d}/2$. Next, we show that the lower bound follows via the Bernstein condition (5.17).

Computing w_b^*. Differentiating $R(w)$ with respect to w and applying the first order optimality conditions, we obtain the following well-known expression for an unconstrained minimizer of the population risk over $\mathbb{R}^d$: $w_\infty^* = \Sigma^{-1} \mathbf{E} XY$, where $\Sigma = \mathbf{E} XX^\mathsf{T}$. A simple calculation shows that

$$\begin{aligned} \Sigma = \alpha \mathbf{1}\mathbf{1}^\mathsf{T} + \beta I_d, \quad \text{with} \quad & \alpha = (1 - d^{-1/2}) d^{-2} + (d^2 + d^{3/2})^{-1} \\ \text{and} \quad & \beta = d^{-3/2} - (d^2 + d^{3/2})^{-1}. \end{aligned} \tag{5.18}$$

By the Sherman-Morrison formula, we have

$$\Sigma^{-1} = (\beta I_d)^{-1} - \frac{(\beta I_d)^{-1} \alpha \mathbf{1}\mathbf{1}^\mathsf{T} (\beta I_d)^{-1}}{1 + \alpha \mathbf{1}^\mathsf{T} (\beta I_d)^{-1} \mathbf{1}} = \beta^{-1} I_d - \frac{\alpha \beta^{-2}}{1 + \alpha \beta^{-1} d} \mathbf{1}\mathbf{1}^\mathsf{T},$$

which plugged into the equation $w_\infty^* = \Sigma^{-1} \mathbf{E} XY$ yields

$$w_b^* = \left(\beta^{-1} - \frac{\alpha \beta^{-2} d}{1 + \alpha \beta^{-1} d} \right) (1 - d^{-1/2}) d^{-1} \cdot \mathbf{1} = \frac{\sqrt{d} - 1}{2\sqrt{d} - 1} \cdot \mathbf{1}.$$

For all $d \geq 1$ we have $0 \leq (\sqrt{d} - 1)/(2\sqrt{d} - 1) \leq 1/2$ and, in particular, $\|w_\infty^*\| \leq \sqrt{d}/2 \leq b$.

Lower bounding the excess risk. Let w denote any parameter vector in $\mathbb{R}^d$. Since we have already shown that w_b^* minimizes $R(w)$ over all of $\mathbb{R}^d$, by the Bernstein condition stated in (5.17) we have

$$\begin{aligned} R(w) - R(w_b^*) \geq \mathbf{E} \langle X, w - w_b^* \rangle^2 &= (w - w_b^*)^\mathsf{T} \Sigma (w - w_b^*) \\ &= (w - w_b^*)(\alpha \mathbf{1}\mathbf{1}^\mathsf{T} + \beta I_d)(w - w_b^*), \end{aligned}$$

with the values of α and β given in (5.18). Since $\mathbf{1}\mathbf{1}^\mathsf{T}$ is positive semi-definite, it hence follows that

$$R(w) - R(w_b^*) \geq \beta \|w - w_b^*\|^2 .$$

Finally, for $d \geq 4$ we have $\beta \geq \frac{1}{2} d^{-3/2}$, which completes our proof. □

5.5.3.2 Proof of Lemma 5.2

Computing $\widehat{w}_\infty$. We set once again $\alpha = (n - |I|)d^{-2}$ and $y = \mathbf{1}^\mathsf{T} A^{-1} \mathbf{1}$. Combining (5.11) and (5.12) with $\sum_{i=1}^n X_i Y_i = (n - |I|)\mathbf{1}/d$ and the Sherman-Morrison formula we have

$$\widehat{w}_\infty = d\alpha \left(A^{-1}\mathbf{1} - \frac{\alpha y A^{-1}\mathbf{1}}{1 + \alpha y} \right) = \left(d\alpha - \frac{d\alpha^2 y}{1 + \alpha y} \right) A^{-1}\mathbf{1} = \frac{d\alpha}{1 + \alpha y} A^{-1}\mathbf{1}. \tag{5.19}$$

By (5.14), we have $A\mathbf{1} = |I|\, d^{-1/2}\mathbf{1} + \zeta$. Multiplying both sides by A^{-1} and rearranging yields

$$A^{-1}\mathbf{1} = |I|^{-1}\, d^{1/2} (\mathbf{1} - A^{-1}\zeta).$$

Plugging the above into (5.19) yields

$$\widehat{w}_\infty = \frac{d^{3/2}\, |I|^{-1}\, \alpha}{1 + \alpha y} (\mathbf{1} - A^{-1}\zeta).$$

Computing $\|\widehat{w}_\infty\|^2$. Using the computations as above, we obtain

$$\|\widehat{w}_\infty\|^2 = \langle \widehat{w}_\infty, \widehat{w}_\infty \rangle = (n - |I|)^2 d^{-2} \cdot \mathbf{1}^\mathsf{T} (n\widehat{\Sigma})^{-2} \mathbf{1}. \tag{5.20}$$

To simplify the notation, let $\alpha = (n-|I|)d^{-2}$, $x = \mathbf{1}^\mathsf{T} A^{-2} \mathbf{1}$ and $y = \mathbf{1}^\mathsf{T} A^{-1} \mathbf{1}$. Applying the Sherman-Morrison formula together with (5.12) we have

$$\begin{aligned}
&\mathbf{1}^\mathsf{T} (n\widehat{\Sigma})^{-2} \mathbf{1} \\
&= \mathbf{1}^\mathsf{T} \left(A^{-1} - \frac{\alpha A^{-1} \mathbf{1}\mathbf{1}^\mathsf{T} A^{-1}}{1 + \alpha y} \right)^2 \mathbf{1} \\
&= \mathbf{1}^\mathsf{T} \left(A^{-2} - \frac{\alpha A^{-2} \mathbf{1}\mathbf{1}^\mathsf{T} A^{-1}}{1 + \alpha y} - \frac{\alpha A^{-1} \mathbf{1}\mathbf{1}^\mathsf{T} A^{-2}}{1 + \alpha y} + \frac{\alpha^2 A^{-1} \mathbf{1}\mathbf{1}^\mathsf{T} A^{-2} \mathbf{1}\mathbf{1}^\mathsf{T} A^{-1}}{(1 + \alpha y)^2} \right) \mathbf{1} \\
&= x - \frac{\alpha x y}{1 + \alpha y} - \frac{\alpha y x}{1 + \alpha y} + \frac{\alpha^2 y x y}{(1 + \alpha y)^2} = \frac{x}{(1 + \alpha y)^2}.
\end{aligned}$$

Plugging the above into (5.20) yields

$$\|\widehat{w}_\infty\|^2 = (n - |I|)^2 d^{-2} \frac{\mathbf{1}^\mathsf{T} A^{-2} \mathbf{1}}{(1 + (n - |I|) d^{-2} \mathbf{1}^\mathsf{T} A^{-1} \mathbf{1})^2} \leq n^2 d^{-2} \mathbf{1}^\mathsf{T} A^{-2} \mathbf{1}.$$

The claim follows. □

5.5.3.3 Proof of Lemma 5.3

The proof is based on applying the union bound on the probability of several events. Adjusting the constants one may always guarantee that the statement of Lemma 5.3 holds with probability at least $\frac{1}{2}$. By writing that the event holds with *sufficient probability* we mean that it holds with probability at least $\frac{99}{100}$.

Controlling $|I|$. The result follows from Chebyshev's inequality since $|I|$ follows the Binomial distribution with parameters $n, d^{-1/2}$.

Bound on $\|\zeta\|^2$. Recalling (5.12) and (5.13), we may rewrite $v_i = \sum_{j=1}^{|I|} v_{i,j}$, where $v_{i,j}$ follows the Bernoulli distribution with parameter $d^{-1/2}$. Moreover, for any fixed i we have that $v_{i,1}, \ldots, v_{i,|I|}$ are independent and for any j it holds that $\sum_{i=1}^{d} v_{i,j} = d^{1/2}$.

Combining these facts we have

$$\begin{aligned}\|\zeta\|^2 &= \sum_{i=1}^{d}\left(\sum_{j=1}^{|I|}\left(v_{i,j}-d^{-1/2}\right)\right)^2\\ &= \sum_{i=1}^{d}\sum_{j=1}^{|I|}\left(v_{i,j}-d^{-1/2}\right)^2+\sum_{i=1}^{d}\sum_{j\neq k}^{|I|}\left(v_{i,j}-d^{-1/2}\right)\left(v_{i,k}-d^{-1/2}\right)\\ &= d^{1/2}|I|-|I|+\sum_{i=1}^{d}\sum_{j\neq k}^{|I|}\left(v_{i,j}-d^{-1/2}\right)\left(v_{i,k}-d^{-1/2}\right).\end{aligned}$$

We proceed with analysis of the zero mean sum $\sum\limits_{i=1}^{d}\sum\limits_{j\neq k}^{|I|}\left(v_{i,j}-d^{-1/2}\right)\left(v_{i,k}-d^{-1/2}\right)$. Observe that for any given j the values $v_{1,j},\dots,v_{d,j}$ are not independent but are sampled with replacement. However, it is possible to avoid this problem using a direct computation. First, for $i_1\neq i_2$ and any j we have

$$\begin{aligned}\mathbf{E}\left(v_{i_1,j}-d^{-1/2}\right)\left(v_{i_2,j}-d^{-1/2}\right) &= \mathbf{E}v_{i_1,j}v_{i_2,j}-d^{-1}=\frac{d^{1/2}}{d}\cdot\frac{d^{1/2}-1}{d-1}-d^{-1}\\ &= -\frac{1}{d^{3/2}+d}.\end{aligned}$$

This implies the following correlation identity

$$\begin{aligned}&\mathbf{E}\left(\left(\sum_{j\neq k}^{|I|}\left(v_{i_1,j}-d^{-1/2}\right)\left(v_{i_1,k}-d^{-1/2}\right)\right)\left(\sum_{j\neq k}^{|I|}\left(v_{i_2,j}-d^{-1/2}\right)\left(v_{i_2,k}-d^{-1/2}\right)\right)\,\middle|\,|I|\right)\\ &=\mathbf{E}\left(\left(\sum_{j\neq k}^{|I|}\left(v_{i_1,j}-d^{-1/2}\right)\left(v_{i_1,k}-d^{-1/2}\right)\left(v_{i_2,j}-d^{-1/2}\right)\left(v_{i_2,k}-d^{-1/2}\right)\right)\,\middle|\,|I|\right)\\ &=\frac{|I|^2-|I|}{(d^{3/2}+d)^2}.\end{aligned}$$

The last identity leads to

$$\begin{aligned}&\mathbf{E}\left(\left(\sum_{i=1}^{d}\sum_{j\neq k}^{|I|}\left(v_{i,j}-d^{-1/2}\right)\left(v_{i,k}-d^{-1/2}\right)\right)^2\,\middle|\,|I|\right)\\ &=\sum_{i=1}^{d}\mathbf{E}\left(\left(\sum_{j\neq k}^{|I|}\left(v_{i,j}-d^{-1/2}\right)\left(v_{i,k}-d^{-1/2}\right)\right)^2\,\middle|\,|I|\right)+\left(d^2-d\right)\frac{|I|^2-|I|}{(d^{3/2}+d)^2}\\ &=d(|I|^2-|I|)\left(d^{-1/2}(1-d^{-1/2})\right)^2+\left(d^2-d\right)\frac{|I|^2-|I|}{(d^{3/2}+d)^2}\le 2|I|^2.\end{aligned}$$

Finally, using Chebyshev's inequality we have $\|\zeta\|^2\sim|I|\,d^{1/2}\sim n$ with sufficient probability.

Invertibility of A. Observe that A is a sum of $|I|$ independent positive semi-definite random matrices such that each summand has operator norm equal to one. Using the lower tail of the matrix Chernoff bound [196, Theorem 5.1.1] we have

$$\Pr(\lambda_d(A) \le \lambda_d(\mathbf{E}(A \mid |I|))/2 \mid |I|) \le d(2e^{-1})^{|I|(d^{-1}+(d^{3/2}+d)^{-1})/2},$$

which is arbitrary small for $|I| \gg d$. The latter condition holds with high probability, recalling that with sufficiently high probability we have $|I| \gtrsim nd^{-1/2}$, where d and $n \gtrsim d^3 \log d$. Finally, observe that

$$\lambda_d(\mathbf{E}(A \mid |I|)) = \lambda_d\left(|I|\left((d^{-1} - (d^{3/2}+d)^{-1})I_d + (d^{3/2}+d)^{-1}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\right) \sim |I|d^{-1}.$$

Therefore, on an event with sufficiently high probability $\lambda_d(A) > 0$ and hence, on this event, A is invertible.

A lower bound on $\lambda_1(A)$. By (5.13) we have $\mathbf{1}^\mathsf{T} A\mathbf{1} = \mathbf{1}^\mathsf{T}(|I|\, d^{-1/2}\mathbf{1} + \zeta) = |I|\, d^{-1/2}\, \|\mathbf{1}\|^2$, which shows that

$$\lambda_1(A) \ge |I|\, d^{-1/2}. \tag{5.21}$$

An upper bound on $\lambda_2(A)$. We need to prove the following bound stating that with sufficient probability

$$\lambda_2(A) \lesssim |I|d^{-1}. \tag{5.22}$$

By the Courant-Fischer theorem we have

$$\lambda_2(A) = \inf_v \sup_{x \in S^{d-1}, \langle x, v\rangle = 0} x^\mathsf{T} A x \le \sup_{x \in S^{d-1}, \langle x, \mathbf{1}\rangle = 0} x^\mathsf{T} A x.$$

Consider the $d \times d$ partial isometry matrix R defined as follows. Fix an orthonormal basis $w_1, \dots, w_d$ in $\mathbb{R}^d$ such that w_1 is proportional to $\mathbf{1}$. The matrix R has its first row equal to zero and its i-th row for $i \ge 2$ equal to w_i. Observe that $R\,\mathbf{1} = 0$ and for any v such that $\langle u, \mathbf{1}\rangle = 0$ we have $\|Ru\| = \|u\|$ together with $RR^\mathsf{T} = I_d - e_1 e_1^\mathsf{T}$. Next, we show

$$\sup_{x \in S^{d-1}, \langle x, \mathbf{1}\rangle = 0} x^\mathsf{T} A x = \sup_{x \in S^{d-1}} x^\mathsf{T} R A R^\mathsf{T} x. \tag{5.23}$$

Indeed, consider a maximizer $x_0 \in S^{d-1}$ of the right-hand side. We have that $R^\mathsf{T} x_0$ is orthogonal to $\mathbf{1}$ since $\mathbf{1}^\mathsf{T} R^\mathsf{T} x_0 = (R\ \mathbf{1})^\mathsf{T} x_0 = 0$. Finally, we have that for any $x' \in S^{d-1}$ such that $\langle x', \mathbf{1}\rangle = 0$ there is $x \in S^{d-1}$ such that $R^\mathsf{T} x = x'$. This is because

$x' = \alpha_2 w_2 + \ldots \alpha_d w_d = R^\mathsf{T} x$, where $x^\mathsf{T} = (0, \alpha_2, \ldots, \alpha_d) \in S^{d-1}$. Therefore, (5.23) follows.

Further, the matrix RAR^T is non-negative semi-definite as well as each additive term that forms it. We have

$$RAR^\mathsf{T} = \sum_{i \in I} RX_i X_i^\mathsf{T} R^\mathsf{T}$$

and for the operator norm we have $\|RX_i X_i^\mathsf{T} R^\mathsf{T}\| \leq \|R\| \|X_i X_i^\mathsf{T}\| \|R^\mathsf{T}\| = \|R\| \left\|R^\mathsf{T}\right\| \leq 1$. Note that

$$\begin{aligned}\mathbf{E}(RAR^\mathsf{T} \mid |I|) &= R\mathbf{E}(A \mid |I|)R^\mathsf{T} \\ &= |I|\, R\left((d^{-1} - (d^{3/2}+d)^{-1})I_d + (d^{3/2}+d)^{-1}\mathbf{1}\mathbf{1}^\mathsf{T}\right) R^\mathsf{T}.\end{aligned}$$

Using $R\,\mathbf{1} = 0$, the above simplifies to

$$\mathbf{E}(RAR^\mathsf{T} \mid |I|) = R\mathbf{E}\left(A \mid |I|\right) R^\mathsf{T} = |I|\left(d^{-1} - (d^{3/2}+d)^{-1}\right) RR^\mathsf{T}.$$

Since $RR^\mathsf{T} = I_d - e_1 e_1^\mathsf{T}$, we have $\lambda_1(\mathbf{E}(RAR^\mathsf{T} \mid |I|)) = |I|\,(d^{-1} - (d^{3/2}+d)^{-1})$. Applying the matrix Chernoff inequality [196, Theorem 5.1.1] we obtain

$$\begin{aligned}&\Pr\left(\lambda_2(A) \geq 2\,|I|\,(d^{-1} - (d^{3/2}+d)^{-1}) \mid |I|\right) \\ &\leq \Pr\left(\lambda_1(RAR^\mathsf{T}) \geq 2\,|I|\,(d^{-1} - (d^{3/2}+d)^{-1}) \mid |I|\right) \\ &\leq d(e/4)^{|I|(d^{-1}+(d^{3/2}+d)^{-1})}. \end{aligned} \tag{5.24}$$

The above probability is arbitrary small for large enough $|I|$. The desired high probability bound follows for large enough d and using the fact that $n \gtrsim d^3 \log d$.

A lower bound on $\zeta^\mathsf{T} A^{-1} \zeta$. Let $u_1, \ldots, u_d$ be an orthonormal basis of eigenvectors of A. Using the spectral decomposition and $\sum_{i=1}^d \langle u_i, \zeta\rangle^2 = \|\zeta\|^2$, we have

$$\zeta^\mathsf{T} A^{-1} \zeta = \sum_{i=1}^d \langle u_i, \zeta\rangle^2 \lambda_i(A)^{-1} \geq \lambda_2(A)^{-1} \sum_{i=2}^d \langle u_i, \zeta\rangle^2 = \lambda_2(A)^{-1}(\|\zeta\|^2 - \langle u_1, \zeta\rangle^2).$$

By (5.22) we have $\lambda_2(A) \lesssim nd^{-3/2}$ and by above computations $\|\zeta\|^2 \sim n$. Therefore, the claim immediately follows if we prove that $\langle u_1, \zeta\rangle^2 \ll n$. Note that

$$\lambda_1(A)\,\langle u_1, \zeta\rangle = \langle Au_1, \zeta\rangle = \langle (A - \mathbf{E}(A \mid |I|))u_1, \zeta\rangle + \langle \mathbf{E}(A \mid |I|)u_1, \zeta\rangle$$

implies

$$\langle u_1, \zeta\rangle^2 \leq \lambda_1(A)^{-2}(\|A - \mathbf{E}(A \mid |I|)\|\,\|\zeta\| + |\,\langle \mathbf{E}(A \mid |I|)u_1, \zeta\rangle\,|)^2. \tag{5.25}$$

Recall that

$$\mathbf{E}(A \mid |I|) = |I| \left((d^{-1} - (d^{3/2}+d)^{-1}) I_d + (d^{3/2}+d)^{-1} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \quad \text{and} \quad \langle \mathbf{1}, \zeta \rangle = 0.$$

We have

$$|\langle \mathbf{E}(A \mid |I|) u_1, \zeta \rangle| = |I| \, (d^{-1} + (d^{3/2}+d)^{-1}) |\langle u_1, \zeta \rangle| \leq |I| \, (d^{-1} + (d^{3/2}+d)^{-1}) \|\zeta\|.$$

Using that $(X_i X_i^\mathsf{T})^2 = X_i X_i^\mathsf{T}$ for $i \in I$, we have

$$\left\| \sum_{i=1}^{|I|} \mathbf{E} \left(X_i X_i^\mathsf{T} - \mathbf{E} X_i X_i^\mathsf{T} \right)^2 \right\| \leq \left\| \sum_{i=1}^{|I|} \mathbf{E} \left(X_i X_i^\mathsf{T} \right)^2 \right\| = \|\mathbf{E}(A \mid |I|)\| \leq 2 \, |I| \, d^{-1/2}.$$

Applying the matrix Bernstein inequality [196, Theorem 6.6.1] we obtain

$$\Pr \left(\|A - \mathbf{E}(A \mid |I|)\| \geq |I| \, d^{-1} \,\middle|\, |I| \right) \leq d \exp \left(- \frac{|I|^2 \, d^{-2}/2}{2 \, |I| \, d^{-1/2} + d/3} \right), \tag{5.26}$$

where the above probability is arbitrary small for large enough d and $n \gtrsim d^3 \log d$. Hence, (5.25) gives with sufficient probability

$$\langle u_1, \zeta \rangle^2 \leq 2\lambda_1(A)^{-2} (\|A - \mathbf{E}(A \mid |I|)\|^2 + 2 \, |I|^2 \, d^{-2}) \, \|\zeta\|^2 \lesssim \lambda_1(A)^{-2} \, |I|^2 \, d^{-2} \, \|\zeta\|^2 \lesssim \frac{n}{d}.$$

The claim follows.

An upper bound on $\mathbf{1}^\mathsf{T} A^{-1} \mathbf{1}$. As before let $u_1, \ldots, u_d$ be an orthonormal basis of eigenvectors of A. Using the lower bound (5.21) and $\sum_{i=2}^d \langle u_i, \mathbf{1} \rangle^2 = d - \langle u_1, \mathbf{1} \rangle^2$, we have

$$\begin{aligned} \mathbf{1}^\mathsf{T} A^{-1} \mathbf{1} &= \sum_{i=1}^d \langle u_i, \mathbf{1} \rangle^2 \, \lambda_i(A^{-1}) \\ &\leq \langle u_1, \mathbf{1} \rangle^2 / \lambda_1(A) + (d - \langle u_1, \mathbf{1} \rangle^2) / \lambda_d(A) \leq d^{3/2}/|I| + (d - \langle u_1, \mathbf{1} \rangle^2) / \lambda_d(A). \end{aligned} \tag{5.27}$$

We want to provide an upper bound on $d - \langle u_1, \mathbf{1} \rangle^2$. By (5.24) we have that with sufficient probability $\lambda_j(A) \leq 2|I|d^{-1}$ for $j = 2, \ldots, d$ and therefore, for the same values of j we have $\lambda_j(A)/\lambda_1(A) \leq 2d^{-1/2}$. Using the last fact we have

$$\frac{\mathbf{1}^\mathsf{T} A \mathbf{1}}{\lambda_1(A)} = \langle u_1, \mathbf{1} \rangle^2 + \sum_{i=2}^d \langle u_i, \mathbf{1} \rangle^2 \, \frac{\lambda_i(A)}{\lambda_1(A)} \leq \langle u_1, \mathbf{1} \rangle^2 + 2d^{-1/2} (d - \langle u_1, \mathbf{1} \rangle^2). \tag{5.28}$$

Now observe that since $\mathbf{1}$ is the first eigenvector of $\mathbf{E}(A \mid |I|)$, we have

$$d = \mathbf{1}^\mathsf{T} \mathbf{1} = \frac{\mathbf{1}^\mathsf{T} \mathbf{E}(A\mathbf{1} \mid |I|)}{\lambda_1(\mathbf{E}(A \mid |I|))}.$$

Rearranging the inequality (5.28) and combining it with the above identity yields, for $d \geq 16$, the following bound:

$$d - \langle u_1, \mathbf{1} \rangle^2 \leq (1 - 2d^{-1/2})^{-1} \left(d - \frac{\mathbf{1}^\mathsf{T} A \mathbf{1}}{\lambda_1(A)} \right) \leq 2 \left(\frac{\mathbf{1}^\mathsf{T} \mathbf{E}(A \mid |I|)\mathbf{1}}{\lambda_1(\mathbf{E}(A \mid |I|))} - \frac{\mathbf{1}^\mathsf{T} A \mathbf{1}}{\lambda_1(A)} \right). \quad (5.29)$$

Finally, noting that $|\lambda_1(A) - \lambda_1(\mathbf{E}(A \mid |I|))| \leq \|A - \mathbf{E}(A \mid |I|)\|$ and combining this with the lower bound (5.21), it follows that

$$\begin{aligned} &\frac{\mathbf{1}^\mathsf{T} \mathbf{E}(A \mid |I|)\mathbf{1}}{\lambda_1(\mathbf{E}(A \mid |I|))} - \frac{\mathbf{1}^\mathsf{T} A \mathbf{1}}{\lambda_1(A)} \\ &= \mathbf{1}^\mathsf{T} \mathbf{E}(A \mid |I|)\mathbf{1} \left(\frac{1}{\lambda_1(\mathbf{E}(A \mid |I|))} - \frac{1}{\lambda_1(A)} \right) + \frac{\mathbf{1}^\mathsf{T} (\mathbf{E}(A \mid |I|) - A)\mathbf{1}}{\lambda_1(A)} \\ &\leq \frac{2d \|A - \mathbf{E}(A \mid |I|)\|}{\lambda_1(A)} \leq \frac{2d^{3/2} \|A - \mathbf{E}(A \mid |I|)\|}{|I|}. \end{aligned} \quad (5.30)$$

By (5.26), we have that with sufficient probability $\|A - \mathbf{E}(A \mid |I|)\| \lesssim |I| d^{-1}$. Therefore, combining this with (5.29) we have with sufficient probability

$$d - \langle u_1, \mathbf{1} \rangle^2 \lesssim \sqrt{d}.$$

Plugging the above inequality into (5.27) and using our lower bound $\lambda_d(A) \gtrsim |I| d^{-1}$ we prove the claim.

An upper bound on $\mathbf{1}^\mathsf{T} A^{-2} \mathbf{1}$. The proof is completely analogous to the case $\mathbf{1}^\mathsf{T} A^{-1} \mathbf{1}$. We have

$$\begin{aligned} \mathbf{1}^\mathsf{T} A^{-2} \mathbf{1} &\leq \langle u_1, \mathbf{1} \rangle^2 / (\lambda_1(A))^2 + (d - \langle u_1, \mathbf{1} \rangle^2) / (\lambda_d(A))^2 \\ &\leq d^2 / |I|^2 + (d - \langle u_1, \mathbf{1} \rangle^2) / (\lambda_d(A))^2. \end{aligned}$$

As before, with sufficient probability we have $d^2/|I|^2 \lesssim n^{-2} d^3$. The only difficulty is that we need a slightly sharper variant of the upper bound on $\|A - \mathbf{E}(A \mid |I|)\|$. Recalling the bound (5.26), by the matrix Bernstein inequality [196, Theorem 6.6.1] we have

$$\Pr\left(\|A - \mathbf{E}(A \mid |I|)\| \geq |I| \, d^{-3/2} \,\middle|\, |I| \right) \leq d \exp\left(- \frac{|I|^2 \, d^{-3}/2}{4 \, |I| \, d^{-1/2} + d/3} \right),$$

which is, with high probability, arbitrarily small provided that d is large enough and $n \gtrsim d^3 \log d$. Observe that this is the step where we have our strongest requirement on n. Note that using that by matrix Chernoff inequality, as shown above in the proof that A is invertible, we have with sufficient probability:

$$\lambda_d(A) \gtrsim |I| \, d^{-1}.$$

Using (5.29), (5.30) and the two inequalities above, we conclude that the following holds with sufficient probability:

$$\frac{d - \langle u_1, \mathbf{1} \rangle^2}{\lambda_d(A)^2} \lesssim \frac{d^{3/2} \|A - \mathbf{E}(A \mid |I|)\| / |I|}{|I|^2 d^{-2}} \lesssim \frac{d^{7/2} |I| d^{-3/2}}{|I|^3} = \frac{d^2}{|I|^2} \lesssim \frac{d^3}{n^2}.$$

The proof of our result is complete. □

5.6 Limitations and Open Directions

Learning linear functions over a bounded Euclidean ball with the quadratic loss is one of the simplest possible settings in Statistical Learning. However, fully characterizing the performance of constrained least squares under boundedness constraints remains an open problem. While Theorem 5.1 provides the sharpest known upper bound, whether one can obtain a matching distribution-dependent lower bound expressed in terms of correlation between leverage scores and the noise remains unknown. In other words, it would be interesting to obtain a complexity measure that would simultaneously provide sharp upper and *lower bounds* on the size of the excess risk for the least squares estimator in the random design regression setting considered in the present chapter. Such a result would be more in the spirit of modern Probability Theory, where both upper and lower bounds on the supremum of stochastic processes are studied in great detail. See the book by Talagrand [193].

Finally, let us remark that along the lines of obtaining sharp upper and lower bound, Chatterjee [50] achieves this goal in the fixed design setting, where the aim is to estimate a vector $\mu \in \mathbb{R}^d$ given an observation $Y = \mu + \xi$, where ξ denotes a vector distributed according to the d-dimensional standard Gaussian distribution. The estimator studied by Chatterjee [50] is the Euclidean projection of the observation vector Y onto a closed and convex subset of $\mathbb{R}^d$ that contains the vector μ. We remark that our setting is different since we study a random design setting, and without distributional assumptions such as Gaussian noise; hence, it is not clear a priori to what extent the results of Chatterjee [50] can answer the question outlined in this section.

5.7 Bibliographic Remarks

This chapter is adapted from the joint work with Nikita Zhivotovskiy [209]. Theorem 5.1 proved in [209] is based on localization arguments via fixed points instead of the offset approach undertaken in the presentation considered in this chapter. A slightly sharper bound on the localized quadratic process is obtained in [209], based on a version

of Rudelson's inequality [183] for sums of rank-one operators due to Oliveira [162]. In particular, it allows to replace the $\log(1+d)$ term appearing in Theorem 5.1 via $\log(\min\{n,d\})$.

Let us now mention some results present in [209] that were omitted in this chapter. First, the work [209] also contains a sharp upper bound for the ridge regression algorithm, closely related to the constrained least squares estimator considered in this chapter. This excess risk bound is obtained via the notion of average stability (see [187, 108, 74] for a detailed account of stability in the context of obtaining distribution-free guarantees under boundedness conditions). In connection to the above-cited works, the novel ingredient in our proof is the exploitation of the curvature of the squared loss in the *stability-fitting trade-off*. As a result, we are able to show that the ridge estimator does not suffer from an excess factor $\log(\min\{n,d\})$ that appears in an upper bound on the localized quadratic process (cf. Theorem 5.1). Moreover, in [209, Proposition 2.3], we demonstrate that this logarithmic term is unavoidable for constrained least squares in some regimes (due to the existence of "bad" empirical risk minimizers), thus showing an interesting performance gap between constrained and penalized least squares. Finally, in [209, Section 2.4], we discuss multiple distributional assumptions, that if granted in addition to boundedness, result in a further upper bound on the multiplier term (the first term in the upper bound of Theorem 5.1) that matches Shamir's lower bound (5.3).

The classical excess risk rate of order d/n for ERM-based algorithms is known to be attainable under various assumptions that are less restrictive than assuming a well-specified model with Gaussian noise. For example, for results under a favorable covariance structure and sub-Gaussian noise assumptions, see [88]; for works that provide guarantees under L_q–L_2 (for some $q>2$) moment equivalence of the marginals $\langle w, X\rangle$ and the noise random variable ξ, see [9, 163, 45, 152]; for bounds that hold under the weaker small-ball assumption, see [141, 114]. Moment equivalence type assumptions allow for modelling heavy-tailed distributions, and, in particular, they have played a crucial role in recent developments in the robust statistics literature (e.g., [45, 55, 128, 148]); however, in some cases, such assumptions only hold with constants that can deteriorate arbitrarily with respect to the parameters of the unknown distribution P, even for light-tailed or bounded distributions. Indeed, the construction of the unfavorable distribution used to prove our main lower bound (Theorem 5.2) is chosen to have ill-behaved moment equivalence assumptions, as discussed at the beginning of Section 5.4.2. In the context of linear regression, the work [45, a discussion following Proposition 4.8] highlights that some of the prior

results on the performance of least squares relying on such assumptions (i.e., moment equivalence assumptions) can have constants in their excess risk bounds that may unintentionally depend on the dimension of the covariates d. Recent literature has further accentuated this problem and witnessed an emerging interest in refining moment equivalence and small-ball assumptions [184, 55, 146].

Many variations of the problem studied in this chapter were previously considered in the literature (e.g., fixed-design regression, distributional assumptions different from boundedness, or performance metrics that differ from the excess risk); see [189] for a detailed comparison of different setups. For comprehensive surveys of existing work, we refer to [8, 9, 88, 152] and the books [78, 106, 216].

Many of the existing upper bounds in the literature hold with high probability. In contrast, in this chapter, we focus on establishing suboptimality of constrained least squares and demonstrating a form of statistical separation between proper and improper estimators; thus, we concentrate on in-expectation analysis to convey our main findings without introducing additional technicalities. In the bounded regime, our upper bound for constrained least squares can likely be translated into high-probability results via standard arguments based on Talagrand's concentration inequality for empirical processes. We believe that our approach for upper bounding the excess risk of constrained ERM can also be used in the case of unbounded (but not heavy-tailed) distributions. Specifically, for bounding the localized quadratic process, there is a version of Oliveira's bound for unbounded matrices [101] that can be used to replace the assumption $\|X\| \leq r$ with a strictly weaker sub-Gaussian tail assumption on the norms $\|X\|$; this could be seen as a step towards incorporating unbounded distributions within our framework while not relying on moment equivalence assumptions discussed above.

To conclude this section, we now turn to a discussion on some prior works in connection with the suboptimality of ERM. Understanding statistical guarantees pertaining to estimators based on ERM has been a subject of intense study in many contexts. Among the simplest problems where ERM is known to incur suboptimal excess risk rates is the model selection aggregation problem; see Section 2.2 for a detailed discussion. When the reference class of functions is finite, ERM performed over this finite dictionary fails due its non-convexity. At the same time, ERM performed over the convex-hull fails to achieve the optimal model selection aggregation rate due to the fact that the convex hull is "too large" (if we keep the original finite class as the reference class with respect to which the excess risk is computed).

Constrained ERM with the squared loss is also actively studied in an on-going line of work concerning the shape restricted regression literature (e.g., [50, 49, 23, 80]), where the least squares projection is performed over constraint sets that may be significantly more complex than Euclidean balls. In particular, when considering some expressive nonparametric classes of functions, ERM can be either optimal [80] or suboptimal [27, 109], depending on some additional properties of these classes. In contrast, our results establish suboptimality of the constrained least squares estimator for a parametric class that has a small intrinsic complexity. The work [50] allows more general convex constraints and shows that ERM can be rate suboptimal. However, establishing suboptimality of ERM in our setting is more complicated: we are not free to choose an arbitrary ill-behaved convex constraint set and also, we study a random design setting and thus cannot choose a fixed set of ill-behaved covariates X_i. We additionally refer to [26] for an extensive discussion on optimality and suboptimality of least squares and maximum likelihood estimators in different setups.

6 Distribution-Free Robust Linear Regression

This chapter is adapted from the joint work with Jaouad Mourtada and Nikita Zhivotovskiy [154]. We study the problem of predicting as well as the best linear function in the absence of any assumptions on the covariates. Our main result is a procedure that attains optimal accuracy/confidence trade-off under minimal assumptions on the conditional distribution of the response variable given the covariate vector, particularly, when no assumptions on the marginal distribution of the covariates are imposed.

The results of this chapter assume familiarity with the mathematical framework of Statistical Learning (Section 2.1) and the concept of improper learning (Section 2.2). We also presuppose familiarity with the results and discussions of Chapter 5. A summary of notation is available in Section 2.6.

6.1 Introduction

We fix the following setup in this chapter:

1. we let the loss function be the quadratic loss $\ell(y, y') = (y - y')^2$;
2. we let the reference class of functions be all d-dimensional linear functions $\mathcal{G}_{\mathrm{lin}} = \{\langle w, \cdot \rangle : w \in \mathbb{R}^d\}$;
3. we aim to obtain excess risk bounds without imposing any assumptions on the marginal distribution of the covariates P_X.

Observe that the setup where we compete against the full linear class $\mathcal{G}_{\mathrm{lin}}$ without distributional assumptions on the covariates falls well outside the scope of the classical tools for obtaining excess risk bounds. For example, the classical local Rademacher complexity machinery (see Section 2.5.1) is only suitable for bounded problems due to the applications of Talagrand's concentration and contraction inequalities. At the same time, we also know from Shamir's lower bound (5.3) (see Section 5.2) that any statistical estimator that returns a linear function is bound to fail. Indeed, proper procedures are sensitive to the scale of the covariate vectors and the Euclidean norm of the optimal linear predictor, and we do not have any control on either of the two quantities under the problem setup outlined above.

For treating unbounded and potentially heavy-tailed distributions of the covariate vectors, over the recent years, L_4–L_2 moment equivalence assumptions and their

variations have become the standard tools in the literature (see, for example, [163, 142, 114, 87, 45, 100, 126, 54, 167]) , which require that the marginal distribution P_X satisfies the following:

$$\Big(\mathbf{E}\,\langle X, w\rangle^4\Big)^{1/4} \leq \kappa\Big(\mathbf{E}\,\langle X, w\rangle^2\Big)^{1/2}, \quad \text{for all} \quad w \in \mathbb{R}^d, \tag{6.1}$$

Indeed, among the first works to highlight the above assumption as a replacement for the more classical boundedness or light-tailedness assumptions is the work of Oliveira [163]. Specifically, in the context of linear regression, Oliveira showed that if $\kappa > 0$ in the above equation is a universal constant, then the classical d/n excess risk rate on the performance of the least squares estimator can be recovered, provided that additional assumptions on the noise are satisfied.

However, as several authors have recently pointed out, the kurtosis constant κ satisfying the inequality (6.1) may depend on the dimension d, leading to suboptimal bounds [45, 163, 114]. In particular, Saumard [184] shows that the slightly weaker small-ball condition fails to hold (with a dimension-free constant) for dictionaries consisting of many classical function bases, such as histograms and wavelets, leading to bounds with a suboptimal dependence on the dimension d. In fact, as we have already seen in Chapter 5 of this thesis, the empirical risk minimization principle can exhibit a highly suboptimal behaviour for data generating distributions that fail to satisfy moment equivalence conditions such as (6.1), despite the otherwise favorable convex and bounded problem setting.

The above discussion naturally brings the question of whether distributional assumptions on X such as the condition (6.1) can be relaxed, and if so, what a corresponding minimal assumption on the distribution of the response variable would be. It is a priori unclear whether non-trivial guarantees are at all possible without imposing any assumptions on the marginal distribution of the covariates P_X. This is the primary subject of investigations in the present chapter.

The starting point of our work is the observation from Section 5.4.3, where it is discussed that the estimator $\widehat{f}^{(\mathrm{FW})}$ due to Forster and Warmuth [66] satisfies the following bound when learning with the quadratic loss for *any* data generating distribution P:

$$\mathbf{E}R(\widehat{f}^{(\mathrm{FW})}(X)) - \inf_{g\in\mathcal{G}_{\mathrm{lin}}} R(g) \leq \frac{2d\,\|Y\|^2_{L_\infty(P)}}{n}. \tag{6.2}$$

Thus, the upper bound on the excess risk of the Forster-Warmuth estimator does not depend on the distribution of covariates. In light of the recent literature related to unbounded and possibly heavy-tailed linear regression, the above bound is far from

trivial – the optimal d/n excess risk rate is attained even for distributions, for which the kurtosis constant κ of (6.1) can be arbitrarily large.

It is important to remark that the Forster-Warmuth estimator [66] originates in the literature on sequential prediction. The estimator $\widehat{f}^{(\mathrm{FW})}$ arises as a refinement of the well-known Vovk-Azoury-Warmuth forecaster [215, 11], a prediction procedure that achieves state of the art performance in the online regret framework when competing with a linear class of functions. Naturally, the tools used to obtain regret bounds in the sequential prediction framework are rather different from those used in the statistical framework considered in this thesis, and even more so when considering robustness to heavy-tailed data.

We now turn to the next section where we discuss why the bound (6.2) does give a fully satisfactory solution to the problem of learning without any assumptions on the covariates, and provide a more precise formulation of the problem that the present chapter addresses.

6.2 Problem Formulation

While the excess risk bound attained by the Forster-Warmuth estimator (6.2) achieves a statistical performance guarantee that circumvents the dependence on the moment equivalence kurtosis constant κ of (6.1), a few natural objections may be raised to challenge the above bound, both of which are related to its high-probability performance.

First, whether the Forster-Warmuth estimator attains the optimal excess risk rate in deviations is unclear because of its improperness. In relation to improperness, the principal example of deviation-suboptimality for an expectation-optimal algorithm is due to Audibert [6]; see Section 2.2 of this thesis for an extended discussion.

Second, in statistical learning regimes that admit heavy-tailed data distributions, expected excess risk is not representative of the true performance. This observation was emphasized by Olivier Catoni in his seminal work [44], which essentially started the line of work concerning robustness to heavy-tailed data (see Section 6.7). Indeed, even for the problem of scalar mean estimation for a data-generating distribution with bounded variance, the ERM estimator (i.e., in this case, the sample mean estimator) achieves optimal performance on average; however, on a set of probability δ, its confidence interval may scale as $\sqrt{1/\delta}$[1], which should be contrasted with confidence intervals that scale as $\sqrt{\log(2/\delta)}$ if the underlying data generating distribution is

[1]Just take a distribution for which Chebyshev's inequality is tight.

subGaussian. At the same time, Catoni [44] shows that confidence intervals scaling as $\sqrt{\log(2/\delta)}$ are *always* attainable for distributions with bounded variance, but such performance is only achievable via different procedures from the classical sample mean estimator.

Thus, different algorithmic principles from ERM are required to handle learning setups that admit heavy-tailed distributions, and it is not clear if the Forster-Warmuth estimator satisfies the optimal accuracy/confidence trafe-off. By accuracy of an estimator $\widehat{f}$, we mean its excess risk $\mathcal{E}(\widehat{f}, \mathcal{G}_{\text{lin}})$. The confidence of $\widehat{f}$ for an error rate ε is equal to $\mathbf{P}(\mathcal{E}(\widehat{f}, \mathcal{G}_{\text{lin}}) \leq \varepsilon)$. Robust statistical learning aims to design procedures with optimal accuracy/confidence trade-off under minimal distributional assumptions. Regarding the minimality of assumptions, in order that the bound (6.2) is non-vacuous, it is necessary to restrict ourselves to data generating distributions with bounded response variable Y. A priori, it is not clear whether such an assumption can be relaxed in the absence of any assumptions on the covariates.

We may now formulate the question that the present chapter aims to address.

> Is it possible to predict as well as the best linear predictor in $\mathcal{G}_{\text{lin}}$ without any assumptions on the distribution of the covariates X, while maintaining the optimal accuracy/confidence trade-off? If so, what is the minimal assumption on the response variable Y allowing this?

In the context of linear regression, the optimal accuracy/confidence trade-off is usually achieved via bounds on excess risk $\mathcal{E}(\widehat{f}, \mathcal{G}_{\text{lin}})$ of order $(d + \log(1/\delta))/n$ that hold with probability at least $1 - \delta$. In particular, such bounds match the performance of ERM for subGaussian data generating distributions. Using either PAC-Bayesian truncations [9, 45] or the median-of-means tournaments [126], it has been shown that the optimal accuracy/confidence trade-off can be achieved under the L_4–L_2 moment equivalence assumption (6.1) together with some additional assumptions on the noise variable $\xi = Y - \langle w^*, X \rangle$, where $\langle w^*, \cdot \rangle$ denotes the optimal linear predictor in $\mathcal{G}_{\text{lin}}$. We remark that the existing procedures for heavy-tailed regression select a function within the class $\mathcal{G}_{\text{lin}}$; however, in view of Shamir's lower bound (5.3) (see Section 5.2), none of such procedures can address our question above, since such procedures are sensitive to the scale of the covariate vectors in the absence of restrictions on the marginal distribution P_X.

6.3 Summary of Contributions

As discussed in Section 6.2, this chapter aims to obtain a procedure that attains the optimal accuracy/confidence trade-off without any assumptions on the marginal distribution P_X. Since the joint distribution $P_{(X,Y)}$ factorizes as $P_X \otimes P_{Y|X}$, we are left to investigate what is the minimal constraint on the probability kernel $(P_{Y|X=x})_{x\in\mathbb{R}^d}$ under which it is possible to obtain non-trivial excess risk guarantees. We identify such an assumption to be the following.

Assumption 6.1. The conditional distribution of Y given X satisfies, for some $m > 0$,

$$\sup_{x\in\mathbb{R}^d} \mathbf{E}[Y^2|X = x] \leq m^2.$$

Upon inspection of the proof of the excess risk bound (6.2) for the Forster-Warmuth estimator (we omit the details, see [66]), it is easy to see that the bounded response variable assumption can be replaced with the above assumption while maintaining the same *expected* excess risk bound of order dm^2/n. However, let us remark that Assumption 6.1 is significantly weaker than assuming that the response variable is bounded, and in fact, it allows for modelling heavy-tailed distributions. For example, we may let $Y = Y' + Z$, where Y' is almost surely bounded by $m/\sqrt{4}$ and Z is a random variable independent of X such that $\mathbf{E}Z^2 \leq m^2/4$ and $\mathbf{E}Z^{2+\varepsilon} = \infty$ for any $\varepsilon > 0$. In this case, the random variable Y satisfies Assumption 6.1 even though the random variable $E[Y|X]$ has only two moments (in particular, it is heavy-tailed).

Our first contribution is establishing that the truncated least squares estimator satisfies the expected excess risk bound of order dm^2/n under Assumption 6.1. This result strengthens the well-known result of Györfi, Kohler, Krzyżak, and Walk [78, Theorem 11.3], who obtain an *inexact* oracle inequality for the truncated least squares (in particular, it does not establish that the excess risk converges to zero with increasing sample size). See Section 6.4.1 for more details.

Our second contribution, presented in Section 6.4.2, demonstrates that both procedures achieving optimal expected excess risk bound under Assumption 6.1, namely, the Forster-Warmuth and the truncated least squares estimators, do not achieve the optimal accuracy/confidence trade-off. In particular, both estimators can incur constant excess risk with constant probability.

Section 6.4.3 contains our third contribution, where we show that in the absence of assumptions on the covariates, Assumption 6.1 cannot be relaxed any further, thus establishing its minimality.

Finally, we present our main result in Section 6.4.4, where we demonstrate that robust learning of linear classes is possible with no restriction on the distribution of the covariates and under the minimal assumption on the conditional distribution of the response variable Y given covariates X (i.e., Assumption 6.1). The construction of our statistical estimator naturally leverages the ideas of the analysis of truncated linear functions [78, Chapter 11], skeleton estimators [61, Section 28.3], [171], the deviation optimal model selection aggregation procedures [6, 112, 143], min-max estimators [9, 111], and the median-of-means tournaments [126]. An extended discussion is deferred to Section 6.4.4.

6.4 Main Results

This section contains the main results. Section 6.4.1 shows that the truncated least squares estimator matches the expected excess risk rate of the Forster-Warmuth estimator. Section 6.4.2 demonstrates that both estimators, however, fail in the high-probability regime, despite the optimality of their average excess risk. In Section 6.4.3, we show that Assumption 6.1 stated in the previous section is minimal if no restrictions are imposed on the marginal distribution of the covariates. Finally, Section 6.4.4 presents a procedure that attains optimal (up to a logarithmic factor) accuracy/confidence trade-off for learning linear functions under Assumption 6.1, thus answering the main question investigated in this chapter.

The proofs are deferred to Section 6.5.

6.4.1 An Improved Bound for Truncated Least Squares

In addition to the Forster-Warmuth estimator discussed in Section 5.4.3, to the best of our knowledge, the only other known result providing excess risk bound guarantees without any assumptions on covariates is the classical upper bound for the truncated linear least squares due to Györfi, Kohler, Krzyżak, and Walk [78, Theorem 11.3], to be improved in this section.

The truncated linear least squares estimator $\widehat{f}_m^{(\mathrm{ERM})}$ is defined as follows. First, consider the linear least squares estimator

$$\widehat{f}^{(\mathrm{ERM})}(\cdot) = \operatorname*{arg\,min}_{f \in \mathcal{G}_{\mathrm{lin}}} R(f) = \langle \widehat{w}^{(\mathrm{ERM})}, \cdot \rangle,$$

where we choose the uniquely defined $\widehat{w}^{(\mathrm{ERM})}$ (among potentially other existing empirical risk minimizers) given by

$$\widehat{w}^{(\mathrm{ERM})} = \Bigg(\sum_{i=1}^{n} X_i X_i^\top \Bigg)^{\dagger} \Bigg(\sum_{i=1}^{n} Y_i X_i \Bigg) = \widehat{\Sigma}_n^{\dagger} \cdot \frac{1}{n} \sum_{i=1}^{n} Y_i X_i \,,$$

with the sample covariance matrix $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ and $\widehat{\Sigma}_n^\dagger$ denoting its Moore-Penrose inverse. Given a threshold $m > 0$, the truncated least squares estimator $\widehat{f}_m^{(\mathrm{ERM})}$ returns the prediction of the linear function $\langle \widehat{w}^{(\mathrm{ERM})}, \cdot \rangle$, truncated to the interval $[-m, m]$. That is,

$$\widehat{f}_m^{(\mathrm{ERM})}(x) = \max(-m, \min(m, \langle \widehat{w}^{(\mathrm{ERM})}, x \rangle)). \tag{6.3}$$

Under Assumption 6.1, the result for the performance of $\widehat{f}_m^{(\mathrm{ERM})}$ [78, Theorem 11.3] states that for some absolute constant $c > 0$ the following holds:

$$\mathbf{E}\, R(\widehat{f}_m^{(\mathrm{ERM})}) - \inf_{g \in \mathcal{G}_{\mathrm{lin}}} R(g) \leq c\, \frac{m^2 d(\log n + 1)}{n} + 7\Big(\inf_{f \in \mathcal{G}_{\mathrm{lin}}} R(f) - R(f_{\mathrm{reg}}) \Big), \tag{6.4}$$

where f_{reg} is the regression function defined as $f_{\mathrm{reg}}(x) = \mathbf{E}[Y|X = x]$. The bound (6.4) is a standard benchmark for several communities. Applications of this result are known in mathematical finance [224], optimal control [24] and variance reduction [73, 72]; there are known improvements of this result under different assumptions [56, 57].

As discussed in Section 5.4.3, the non-linearity introduced by the Forster-Warmuth estimator $\widehat{f}^{(\mathrm{FW})}$ mitigates the instability of ERM predictions at high-leverage points, yielding an excess risk guarantee not achievable via any proper prediction procedure. Compared to the truncated least squares estimator $\widehat{f}_m^{(\mathrm{ERM})}$, the more sophisticated Forster-Warmuth procedure, which relies on an explicit leverage correction, however, leads to a better excess risk guarantee than (6.4). Indeed, the risk guarantee of $\widehat{f}_m^{(\mathrm{ERM})}$ takes the form of an inexact oracle inequality, suffering from the approximation error term $\inf_{f \in \mathcal{G}_{\mathrm{lin}}} R(f) - R(f_{\mathrm{reg}})$. This type of guarantee only ensures that the procedure approaches the performance of the best linear function in the nearly well-specified case, where the true regression function is almost linear. While reasonable in nonparametric estimation [78] with appropriate linear spaces, such an assumption is generally restrictive and is not satisfied in our setting. Unfortunately, the proof technique employed in [78] can only yield inexact oracle inequalities, and hence, no straightforward modification of their argument can match guarantees of the Forster-Warmuth estimator $\widehat{f}^{(\mathrm{FW})}$, which under Assumption 6.1 satisfies an expected excess risk bound of order $m^2 d/n$.

A natural question remains of whether the gap between the existing in-expectation performance guarantees satisfied by the Forster-Warmuth and truncated least squares estimators is intrinsic, or whether it is a byproduct of suboptimal analysis of the performance of the simpler procedure $\widehat{f}_m^{(\mathrm{ERM})}$ (although let us note that it relies on

the knowledge of m). In the theorem below, we show that truncated least squares estimator indeed matches the statistical performance of the Forster-Warmuth algorithm by removing the excess approximation term $7\left(\inf_{f\in\mathcal{G}_{\mathrm{lin}}} R(f) - R(f_{\mathrm{reg}})\right)$ as well as the excess $\log n$ factor appearing in the bound (6.4). Our proof is based on a leave-one-out argument akin to the one used to prove the optimal upper bound for the Forster-Warmuth estimator (see [66, Section 3]). We remark that leave-one-out arguments have a long history; see, for example, the references [203, Chapter 6] and [82].

Theorem 6.1. *Suppose that Assumption 6.1 holds and let $\widehat{f}_m^{(\mathrm{ERM})}$ denote the truncated least squares estimator (6.3). Then, we have*

$$\mathbf{E}R(\widehat{f}_m^{(\mathrm{ERM})}) - \inf_{f\in\mathcal{G}_{\mathrm{lin}}} R(f) \le \frac{8m^2 d}{n+1}.$$

The proof of the above theorem is presented in Section 6.5.1.

6.4.2 Failure of Previous Estimators With Constant Probability

As discussed in Section 6.3, Assumption 6.1 suffices to ensure that the Forster-Warmuth estimator $\widehat{f}^{(\mathrm{FW})}$ [66] achieves an expected excess risk bound of order $m^2 d/n$ irrespective of the distribution of X. Our results established in Section 6.4.1 demonstrate the same conclusion for the truncated least squares estimator of [78, Theorem 11.3]. In addition to the guarantees in expectation, high-probability or tail bounds are desirable, as they provide a control on the probability of failure of the estimator. The following theorem shows that in fact, none of the two procedures satisfy meaningful high-probability guarantees, in a rather strong sense. The proof is deferred to Section 6.5.2.

Theorem 6.2. *Fix the dimension $d = 1$. There exist absolute constants $c > 0$ and $n_0 \ge 2$ such that the following holds. For any sample size $n \ge n_0$, there exists a distribution $P = P(n)$ of (X, Y) with $\|Y\|_{L_\infty} \le m$, such that if $\widehat{f}$ is either the truncated least squares estimator $\widehat{f}_m^{(\mathrm{ERM})}$ (6.3) or the Forster-Warmuth estimator $\widehat{f}^{(\mathrm{FW})}$ (5.8), computed on an i.i.d. sample S_n, then*

$$\mathbf{P}\left(R(\widehat{f}) - \inf_{g\in\mathcal{G}_{\mathrm{lin}}} R(g) \ge c\, m^2\right) \ge c\,.$$

Note that under Assumption 6.1, the trivial, identically zero function has risk at most $\mathbf{E}Y^2 \le m^2$. Theorem 6.2 states that, with constant probability, the truncated least squares and the Forster-Warmuth estimators incur a constant excess risk of the same order. At the first sight, this property may seem incompatible with expected excess risk bounds of order d/n. However, one should keep in mind that the estimators

in question are improper (returning predictors outside of the class $\mathcal{G}_{\text{lin}}$), so that the excess risk may well take negative values; the expected excess risk remains small due to the fact that positive and negative values essentially compensate in expectation, regardless of the distribution. Let us remark that in our context the above-established in-deviation failure is even more severe than that of deviation-suboptimality of progressive mixture rules discussed in Section 2.2. Indeed, above we establish excess risk of constant order, while the suboptimality of progressive mixture rules established by Audibert [6] exhibit the suboptimal slow rate $1/\sqrt{n}$.

6.4.3 The Necessity of Assumption 6.1

The *expected* excess risk guarantees that hold for the Forster-Warmuth and the truncated least squares estimators (cf. Section 6.4.1) only require Assumption 6.1 and in particular, no assumptions on the marginal distribution P_X are imposed. This section is dedicated to establishing that Assumption 6.1 is, in fact, necessary to obtain non-trivial guarantees on the excess risk without restrictions on P_X.

Proposition 6.1. *Fix any $n \geq 1$, $\delta \in (e^{-n}, 1)$ and any measurable function $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(0) = 0$ and $\sup_{x \in \mathbb{R}} f(x)^2 \geq 1$. Then, there exists a distribution P_X of X such that for any estimator $\widehat{f}$ (possibly improper and P_X-dependent), setting $Y = f_{\text{reg}}(X)$ (where $f_{\text{reg}} \in \{f, -f\}$) the following three conditions hold:*

- *there exists $w^* \in \mathbb{R}$ such that $R(\langle w^*, \cdot \rangle) = 0$;*
- $\mathbf{E}[Y^2] \leq 1$;
- *denoting $\|f_{\text{reg}}\|_\infty = \sup_{x \in \mathbb{R}} |f_{\text{reg}}(x)| = \|f\|_\infty \in [1, +\infty]$ we have*

$$\mathbf{P}\left(R(\widehat{f}) \geq \min\left(\frac{\|f_{\text{reg}}\|_\infty^2 \cdot \log(1/\delta)}{4n}, 1 \right) \right) \geq \delta.$$

The proof is deferred to Section 6.5.3. We now comment on the implications of the above lower bound. First, note that if the conditional second moment bound $\mathbf{E}[Y^2|X] \leq 1$ of Assumption 6.1 is relaxed to the weaker unconditional bound $\mathbf{E}Y^2 \leq 1$, then (taking $\delta = 0.9$, and any f such that $\|f\|_\infty \geq \sqrt{n}$) the worst-case excess risk of any estimator $\widehat{f}$ is lower-bounded by an absolute constant c with probability 0.9, matching up to constants the risk of at most 1 trivially achieved by the identically zero function. Second, without Assumption 6.1, the above lower bound shows that the upper bound obtained in the next section (Section 6.4.4) cannot be improved even in the "realizable" case where the linear class $\mathcal{G}_{\text{lin}}$ contains a perfect predictor

(that is, when $R(g) = 0$ for some $g \in \mathcal{G}_{\text{lin}}$) and in particular, $\text{Var}(Y|X) = 0$ almost surely. As a result, the quantity $\sup_{x \in \mathbb{R}^d} \mathbf{E}[Y^2|X = x]$ in Assumption 6.1 cannot be replaced by $\sup_{x \in \mathbb{R}^d} \text{Var}[(]\, Y|X = x)$. Finally, when $Y = f_{\text{reg}}(X)$, then the worst-case dependence on f_{reg} can be no better than $\|f_{\text{reg}}\|_\infty^2$, as shown in the last part of the above proposition. The dependence on m^2 in our upper bounds is thus unavoidable, recalling that $m^2 = \|f_{\text{reg}}\|_\infty^2$ when $Y = f_{\text{reg}}(X)$, as in the lower bound above.

6.4.4 Deviation-Optimal Robust Estimator

In Section 6.4.3, we showed that Assumption 6.1 is *necessary* to ensure the existence of estimators satisfying non-trivial excess risk guarantees in the regimes where the distribution of the covariates is not constrained. The theorem below is the main positive result of this chapter. It demonstrates that Assumption 6.1 is also a *sufficient* condition for the existence of linear regression estimators satisfying an excess risk deviation inequality with logarithmic dependence on the confidence parameter.

Theorem 6.3. *There is an absolute constant $c > 0$ such that the following holds. Assume that $n \geq d$. Suppose that Assumption 6.1 holds and fix any $\delta \in (0, 1)$. Then, there exists an estimator $\widehat{f}$ depending on δ and m such that the following holds:*

$$\mathbf{P}\left(R(\widehat{f}) - \inf_{g \in \mathcal{G}_{\text{lin}}} R(g) \leq c\,\frac{m^2(d\log(n/d) + \log(1/\delta))}{n}\right) \geq 1 - \delta.$$

Moreover, the above bound also holds if the reference class $\mathcal{G}_{\text{lin}}$ is replaced by an arbitrary VC-subgraph class $\mathcal{G}$ of dimension d.

Before presenting our estimator, we briefly comment on the above theorem. First, in contrast to existing work on robust linear regression, our estimator $\widehat{f}$ is improper, even though the underlying linear class is convex. Second, unlike our previous results presented in this paper, the bound of Theorem 6.3 is not specific to the linear class. In particular, our proof extends without changes to the family of VC-subgraph classes (see [71, Definition 3.6.8]). Some recent results in the robust statistics literature apply to more general classes of functions, including non-parametric classes (see, for example, [126, 145, 55]). However, as discussed in Section 6.7, such results are only known to be valid under additional assumptions on P_X. Finally, we note that our estimator depends on the value of m. This assumption simplifies the analysis and is standard in similar contexts (see, for example, [78, Theorem 11.3] and [147]).

We now introduce some additional notation needed to define our estimator. For any $\varepsilon > 0$ and any class of real-valued functions $\mathcal{G}$ let $\mathcal{G}_\varepsilon$ denote the smallest ε-net of

$\mathcal{G}$ with respect to the empirical L_1 distance $\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-g(X_i)|$. We only consider ε-nets that are subsets of $\mathcal{G}$. For the standard definition of an ε-net we refer to [212, Section 4.2]. Assume that we have a sample $S=(X_i,Y_i)_{i=1}^{3n}$ of size $3n$ and denote $S_1=(X_i,Y_i)_{i=1}^{n}$, $S_2=(X_i,Y_i)_{i=n+1}^{2n}$ and $S_3=(X_i,Y_i)_{i=2n+1}^{3n}$. Fix any $1\le k\le n$, and assume without loss of generality that n/k is integer. Split the set $\{1,\dots,n\}$ into k blocks $I_1,\dots,I_k$ of equal size such that $I_j=\{1+(j-1)(n/k),\dots,j(n/k)\}$. Fix any function $\ell:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}$, any sample S' of size n, and denote the i-th element of S' by $Z_i=(X_i,Y_i)$. The median-of-means estimator (see also [124, Section 2.1], [157]) is defined as follows:

$$\mathrm{MOM}^k_{S'}(\ell)=\mathrm{Median}\left(\frac{k}{n}\sum_{i\in I_1}\ell(Z_i),\dots,\frac{k}{n}\sum_{i\in I_k}\ell(Z_i)\right).$$

Finally, for any predictor $f:\mathbb{R}^d\to\mathbb{R}$, recall that the associated loss function is denoted by $\ell_f(Z_i)=(f(X_i)-Y_i)^2$. We are now ready to present our estimator.

The estimator of Theorem 6.3

1. Split the sample S of size $3n$ into three equal parts S_1,S_2 and S_3 as defined above. Use the value m to construct the truncated class
$$\overline{\mathcal{F}}=\left\{f_m : f\in\mathcal{G}_{\mathrm{lin}}\right\},$$
where recall that f_m denotes the truncation of a function f (see (6.3)).

2. Fix $\varepsilon=\frac{md}{n}$. Using the first sample S_1, construct an ε-net of $\overline{\mathcal{F}}$ with respect to the empirical L_1 distance and denote it by $\overline{\mathcal{F}}_\varepsilon$.

3. Let $c_1,c_2>0$ be some specifically chosen absolute constants. Fix the number of blocks $k=\lceil c_1 d(\log(n/d)+\log(1/\delta))\rceil$ and set $\alpha=c_2\sqrt{\frac{m^2(d\log(n/d)+\log(1/\delta))}{n}}$. If $k>n$, then set $\widehat{f}=0$. Otherwise, using the second sample S_2 define a random subset of $\overline{\mathcal{F}}_\varepsilon$ as follows:
$$\widehat{\mathcal{F}}=\left\{f\in\overline{\mathcal{F}}_\varepsilon : \forall g\in\overline{\mathcal{F}}_\varepsilon,\ \mathrm{MOM}^k_{S_2}(\ell_f-\ell_g)\le\alpha\sqrt{\frac{1}{n}\sum_{X_i\in S_2}(f(X_i)-g(X_i))^2+\alpha^2}\right\}.$$

4. Define the set $\widehat{\mathcal{F}}_+$ consisting of all the mid-points of $\widehat{\mathcal{F}}$, that is, $\widehat{\mathcal{F}}_+=(\widehat{\mathcal{F}}+\widehat{\mathcal{F}})/2$. Using the third sample S_3, define our estimator $\widehat{f}$ as
$$\widehat{f}=\operatorname*{arg\,min}_{f\in\widehat{\mathcal{F}}_+}\max_{g\in\widehat{\mathcal{F}}_+}\mathrm{MOM}^k_{S_3}(\ell_f-\ell_g).$$

5. Return $\widehat{f}$.

Our estimator involves a combination of several seemingly disconnected ideas in the literature. The truncation step is inspired by the analysis in [78, Chapter 11], with the difference that we use the truncation as a preliminary step, rather than as a post-processing of the ERM prediction (see Theorem 6.1). The second step replaces the original class by an empirical L_1 ε-net of the truncated class. In many situations, such a construction leads to suboptimal results. However, since we work with a particular parametric class, this step does not affect the resulting performance. The use of the ε-net $\overline{\mathcal{F}}_\varepsilon$ is needed for technical reasons.

If we worked under the assumption of almost surely bounded response variable Y, we could replace the third and fourth steps with any deviation-optimal model selection aggregation procedure. Instead, due to the use of much weaker Assumption 6.1, we need to handle potentially heavy-tailed distributions. Thus, the third and fourth steps can be seen as an adaptation of the midpoint estimator discussed in Section 3.5 of this thesis. In order to adapt the midpoint estimator to heavy-tailed regimes, we draw inspiration from the median-of-means tournaments introduced in [126], and the idea of min-max formulation of robust estimators [9, 111]. We remark that the idea of combining model selection aggregation techniques with the median-of-means tournaments has also recently been explored by Mendelson [145], but under different assumptions. The key distinction from our setting therein is that the suggested learning procedure collapses to a proper estimator for convex classes of functions, such as $\mathcal{G}_{\text{lin}}$ considered in our work; as already discussed, for such procedures some restrictions on the distribution of covariates are required to obtain non-trivial performance bounds.

The proof of the above theorem is presented in Section 6.5.4, where we discuss the technical reasons motivating the definition of our estimator in greater detail.

6.5 Proofs

This section contains the proofs.

6.5.1 Proof of Theorem 6.1

To simplify the presentation, we introduce additional notation. Let $S_{n+1} = (X_i, Y_i)_{i=1}^{n+1}$ denote an i.i.d. sample of size $n+1$. For any $j \in \{1, \dots, n+1\}$, let $S_{n+1}^{(j)} = (X_i, Y_i)_{i=1, i\neq j}^{n+1}$ be the dataset obtained by removing the j-th sample. On the sample S_{n+1} (respectively $S_{n+1}^{(j)}$), we define the minimal norm empirical risk minimizer $\widetilde{f}$ (respectively $\widetilde{f}^{(j)}$) and its truncated variant $\widetilde{f}_m$ (respectively $\widetilde{f}_m^{(j)}$).

Since S_{n+1} is an i.i.d. sample, for every $j \in \{1, \dots, n+1\}$, $S_{n+1}^{(j)}$ has the same distribution as $S_n = S_{n+1}^{(n+1)}$ (so that $\widetilde{f}_m^{(j)}$ has the same distribution as the truncated

least squares estimator $\widehat{f}_m^{(\mathrm{ERM})} = \widetilde{f}_m^{(n+1)}$), and is independent of $Z_j = (X_j, Y_j)$. This implies that the expected excess risk of $\widehat{f}_m^{(\mathrm{ERM})}$ can be bounded as follows:

$$\begin{aligned}
&\mathbf{E}\,\mathcal{E}(\widehat{f}_m^{(\mathrm{ERM})}, \mathcal{G}_{\mathrm{lin}}) \\
&= \mathbf{E}_{S_{n+1}}\Big(\widetilde{f}_m^{(n+1)}(X_{n+1}) - Y_{n+1}\Big)^2 - \inf_{g\in\mathcal{G}_{\mathrm{lin}}} \mathbf{E}_{Z_{n+1}}\left(g(X_{n+1}) - Y_{n+1}\right)^2 \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\Big(\widetilde{f}_m^{(j)}(X_j) - Y_j\Big)^2\right] - \inf_{g\in\mathcal{G}_{\mathrm{lin}}} \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(g(X_j) - Y_j\right)^2\right] \\
&\le \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\Big(\widetilde{f}_m^{(j)}(X_j) - Y_j\Big)^2 - \Big(\widetilde{f}(X_j) - Y_j\Big)^2\right], \qquad (6.5)
\end{aligned}$$

where the last line follows from the definition of $\widetilde{f}$. Now, define the leverage h_j of the point X_j among $X_1, \ldots, X_{n+1}$ by

$$h_j = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top\right)^\dagger X_j, X_j \right\rangle \in [0,1].$$

An explicit computation—postponed to the end of the proof—shows that for every j,

$$\widetilde{f}(X_j) = (1-h_j)\,\widetilde{f}^{(j)}(X_j) + h_j Y_j. \qquad (6.6)$$

Plugging (6.6) into the bound (6.5), we obtain

$$\mathbf{E}\,\mathcal{E}(\widehat{f}_m^{(\mathrm{ERM})}, \mathcal{G}_{\mathrm{lin}}) \le \mathbf{E}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\Big(\widetilde{f}_m^{(j)}(X_j) - Y_j\Big)^2 - (1-h_j)^2\Big(\widetilde{f}^{(j)}(X_j) - Y_j\Big)^2\right]. \qquad (6.7)$$

By Assumption 6.1 and Jensen's inequality we have $\sup_{x\in\mathbb{R}^d}|f_{\mathrm{reg}}(x)| \le m$. It follows that $(\widetilde{f}_m^{(j)}(X_j) - f_{\mathrm{reg}}(X_j))^2 \le (\widetilde{f}^{(j)}(X_j) - f_{\mathrm{reg}}(X_j))^2$, so that

$$\begin{aligned}
&\mathbf{E}\Big[(1-h_j)^2\Big(\widetilde{f}^{(j)}(X_j) - Y_j\Big)^2 \,\Big|\, S_{n+1}^{(j)}, X_j\Big] \\
&= (1-h_j)^2\Big((\widetilde{f}^{(j)}(X_j) - f_{\mathrm{reg}}(X_j))^2 + \mathbf{E}\Big[(f_{\mathrm{reg}}(X_j) - Y_j)^2 | S_{n+1}^{(j)}, X_j\Big]\Big) \\
&\ge \mathbf{E}\Big[(1-h_j)^2\Big(\widetilde{f}_m^{(j)}(X_j) - Y_j\Big)^2 \,\Big|\, S_{n+1}^{(j)}, X_j\Big].
\end{aligned}$$

Plugging the above in the upper bound (6.7), we proceed as follows

$$\begin{aligned}
&\mathbf{E}\,\mathcal{E}(\widehat{f}_m^{(\mathrm{ERM})}, \mathcal{G}_{\mathrm{lin}}) \\
&\le \mathbf{E}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\Big(\widetilde{f}_m^{(j)}(X_j) - Y_j\Big)^2 - (1-h_j)^2\Big(\widetilde{f}_m^{(j)}(X_j) - Y_j\Big)^2\right] \\
&\le \mathbf{E}\left[\frac{1}{n+1}\sum_{j=1}^{n+1} 2h_j\Big(\widetilde{f}_m^{(j)}(X_j) - Y_j\Big)^2\right] \\
&\le 8m^2\mathbf{E}\left[\frac{1}{n+1}\sum_{j=1}^{n+1} h_j\right] \le 8\frac{m^2 d}{n+1},
\end{aligned}$$

where the penultimate step follows from Jensen's inequality combined with Assumption 6.1 and the last step follows from the bound

$$\sum_{j=1}^{n+1} h_j = \mathrm{Trace}\left[\Big(\sum_{i=1}^{n+1} X_i X_i^\top\Big)^\dagger \Big(\sum_{i=1}^{n+1} X_i X_i^\top\Big)\right] \le d.$$

We now conclude by showing the identity (6.6). First, define

$$\widetilde{\Sigma} = \sum_{i=1}^{n+1} X_i X_i^\intercal, \quad \widetilde{\Sigma}^{(j)} = \widetilde{\Sigma} - X_j X_j^\top, \quad b = \sum_{i=1}^{n+1} Y_i X_i, \text{ and } b^{(j)} = b - Y_j X_j,$$

so that

$$\widetilde{f}(X_j) = \langle \widetilde{\Sigma}^\dagger b, X_j\rangle, \quad \widetilde{f}^{(j)}(X_j) = \langle \Big(\widetilde{\Sigma}^{(j)}\Big)^\dagger b^{(j)}, X_j\rangle, \quad \text{and} \quad h_j = \langle \widetilde{\Sigma}^\dagger X_j, X_j\rangle.$$

Note that (6.6) is an identity, and up to restricting to the linear span of $(X_1, \dots, X_{n+1})$ we may assume that $\widetilde{\Sigma}$ is invertible. In addition, if X_j does not belong to the linear span of $(X_i)_{i=1, i\neq j}^{n+1}$, namely, if $\widetilde{\Sigma}^{(j)}$ is singular, then it can be shown that $h_j = 1$ and $\widetilde{f}(X_j) = Y_j$ (since $\widetilde{f}$ minimizes the empirical risk on S_{n+1}, and $g(X_j)$ can be set freely without affecting the other predictions), so that (6.6) holds. Therefore, we may assume that $\widetilde{\Sigma}^{(j)}$ is invertible. Using the definition and the Sherman-Morrison formula, as $h_j \in [0, 1)$, we obtain

$$\begin{aligned}
\widetilde{f}^{(j)}(X_j) &= \left\langle \left(\widetilde{\Sigma}^{-1} + \frac{\widetilde{\Sigma}^{-1} X_j X_j^\top \widetilde{\Sigma}^{-1}}{1 - h_j}\right)(b - Y_j X_j), X_j \right\rangle \\
&= \widetilde{f}(X_j) + \frac{h_j}{1 - h_j}\widetilde{f}(X_j) - h_j Y_j - \frac{h_j^2}{1 - h_j} Y_j \\
&= \frac{1}{1 - h_j}\widetilde{f}(X_j) - \frac{h_j}{1 - h_j} Y_j;
\end{aligned}$$

rearranging the last equality yields (6.6), concluding the proof. □

6.5.2 Proof of Theorem 6.2

For any $n \ge n_0$, let $P = P(n)$ be the distribution of (X, Y) satisfying

$$(X, Y) = \begin{cases} (1, m) & \text{with probability } 1 - \frac{1}{n}; \\ (\sqrt{n}, 0) & \text{with probability } \frac{1}{n}. \end{cases}$$

By homogeneity, we may assume that $m = 1$. For any $w \in \mathbb{R}$, set $f_w(x) = w \cdot x$. We have

$$R(f_w) = \left(1 - \frac{1}{n}\right)(w - 1)^2 + \frac{1}{n}(w\sqrt{n})^2 = \left(1 - \frac{1}{n}\right)(w - 1)^2 + w^2.$$

It follows that the risk of the best linear predictor is equal to

$$\inf_{w\in\mathbb{R}} R(f_w) = \frac{1-1/n}{2-1/n} \le \frac{1}{2}\,. \tag{6.8}$$

In addition, let $K = K_n$ denote the number of indices $i = 1, \dots, n$ such that $X_i = \sqrt{n}$. The empirical risk writes

$$R_n(f_w) = \Big(1 - \frac{K}{n}\Big)(w-1)^2 + Kw^2,$$

and so

$$\widehat{w}^{(\mathrm{ERM})} = \operatorname*{arg\,min}_{w\in\mathbb{R}} R_n(f_w) = \frac{1-K/n}{K+1-K/n}\,.$$

In particular, $0 \le \widehat{w}^{(\mathrm{ERM})} \le 1/(K+1)$. Now, note that if $\widehat{f}$ denotes either the truncated least squares or the Forster-Warmuth estimator, then $\widehat{f}(1) \le \widehat{w}^{(\mathrm{ERM})} \cdot 1 \le 1/(K+1) \le 1$, and thus, denoting the sample $(X_i, Y_i)_{i=1}^n$ by S_n, we have

$$R(\widehat{f}) \ge \mathbf{E}\big[(\widehat{f}(X) - Y)^2 \mathbf{1}(X=1) \,|\, S_n\big] \ge \Big(1 - \frac{1}{n}\Big) \cdot \Big(\frac{K}{K+1}\Big)^2. \tag{6.9}$$

Thus, under the event $E_n = \{K_n \ge 4\}$, it follows from (6.8) and (6.9) that for $n \ge 16$,

$$R(\widehat{f}) - \inf_{g\in\mathcal{G}_{\mathrm{lin}}} R(g) \ge \Big(1 - \frac{1}{n}\Big) \cdot \Big(\frac{K}{K+1}\Big)^2 - \frac{1}{2} = \Big(1 - \frac{1}{16}\Big) \cdot \frac{16}{25} - \frac{1}{2} = \frac{1}{10}\,.$$

Finally, since K_n follows the binomial distribution $\mathsf{Binomial}(n, 1/n)$, the probability $\mathbf{P}(E_n)$ is positive for $n \ge 16 \ge 4$. Further, since K_n converges in distribution to the Poisson distribution $\mathsf{Poisson}(1)$ as $n \to \infty$, $\mathbf{P}(E_n) \to \mathbf{P}(\widetilde{K} \ge 4) > 0$ with $\widetilde{K} \sim \mathsf{Poisson}(1)$, so that setting $p_0 = \inf_{n\ge 16} \mathbf{P}(E_n)$, we have $p_0 > 0$. This concludes the proof with $c = \min(p_0, 1/10)$ and $n_0 = 16$. □

6.5.3 Proof of Proposition 6.1

Let $p \in (0,1)$ be such that $(1-p)^n = \delta$; using that $1 - e^{-u} \ge (1-e^{-1})u \ge u/2$ for $u = \log(1/\delta)/n \in [0,1]$, we have

$$p = 1 - \delta^{1/n} \ge \frac{\log(1/\delta)}{2n}. \tag{6.10}$$

Let $x_0 \in \mathbb{R} \setminus \{0\}$ be such that $|f(x_0)|$ is larger than $\min(\|f\|_\infty/\sqrt{2}, 1/\sqrt{p})$ and let $p_0 = \min(p, 1/f(x_0)^2)$. Fix the distribution of the covariates P_X as follows:

$$X = \begin{cases} 0 & \text{with probability } 1-p_0, \\ x_0 & \text{with probability } p_0. \end{cases}$$

Up to replacing f by $-f$, assume that $f(x_0) > 0$. For $\varepsilon \in \{-1, 1\}$, let P_ε denote the joint distribution of the random pair $(X, \varepsilon f(X))$ (where the marginal distribution of X is given by P_X defined above), and let R_ε denote the risk functional associated to the distribution P_ε. Note that P_ε satisfies the first condition of the proposition with $w^* = \varepsilon f(x_0)/x_0$. Also, the second condition holds since $\mathbf{E}Y^2 = p_0 f(x_0)^2 \leq 1$.

We now turn to proving the third condition of this proposition. Let $\widehat{f}$ be an arbitrary procedure, possibly improper and depending on P_X. Let $S_0 = \big((0,0), \ldots, (0,0)\big)$ denote a sample of n points equal to $(0,0)$. Since the quadratic loss function is convex, we may assume without loss of generality that $\widehat{f}$ is a deterministic procedure and let $f' : \mathbb{R} \to \mathbb{R}$ denote the output of $\widehat{f}$ on the sample S_0, that is, $f' = \widehat{f}(S_0)$. By symmetry of the problem, assume that $f'(x_0) \leq 0$ and fix the distribution P of (X, Y) to P_1 (if $f'(x_0) \geq 0$, we may fix $P = P_{-1}$ instead). Consider the event $E = \{X_1 = \cdots = X_n = 0\}$ and note that $\mathbf{P}(E) = (1 - p_0)^2 \geq (1 - p)^n = \delta$. Since $f(0) = 0$, on the event E the observed sample is S_0, so that by (6.10) we have

$$\begin{aligned} R(\widehat{f}) &\geq \mathbf{E}[(f'(X) - Y)^2 \mathbf{1}(X = x_0)] = p_0 \cdot (f'(x_0) - f(x_0))^2 \geq p_0\, f(x_0)^2 \\ &= \min\Big(p\, f(x_0)^2, 1\Big) \geq \min\left(\frac{p\|f\|_\infty^2}{2}, 1\right) \geq \min\left(\frac{\|f_{\mathrm{reg}}\|_\infty^2 \cdot \log(1/\delta)}{4n}, 1\right), \end{aligned}$$

which completes our proof.

6.5.4 Proof of Theorem 6.3

This section is devoted to proving Theorem 6.3. We first state and comment on some technical lemmas, the proofs for which will be provided in Section 6.5.5.

First, observe that the truncation at the level m can only make the risk smaller whenever Assumption 6.1 is satisfied. Indeed, this follows from the identity

$$R(g) = \mathbf{E}(g(X) - f_{\mathrm{reg}}(X))^2 + \mathbf{E}(f_{\mathrm{reg}}(X) - Y)^2,$$

and the fact that f_{reg} is absolutely bounded by m. Therefore, we may focus on bounding the excess risk computed with respect to the reference class of truncated linear functions $\overline{\mathcal{F}}$ instead of $\mathcal{G}_{\mathrm{lin}}$:

$$R(\widehat{f}) - \inf_{g \in \overline{\mathcal{F}}} R(g).$$

The following lemma provides a uniform deviation bound on the L_1 distances between the elements of $\overline{\mathcal{F}}$. This lemma will be used to reduce the statistical problem of competing with the *infinite class* $\overline{\mathcal{F}}$, to the problem of competing with the *finite class* $\overline{\mathcal{F}}_\varepsilon$.

Lemma 6.1. *Assume that $n \geq d$. There is a constant $c > 0$ such that simultaneously for all $f, g \in \overline{\mathcal{F}}$, with probability at least $1 - \delta$, it holds that*

$$\mathbf{E}|f(X) - g(X)| \leq \frac{2}{n}\sum_{i=1}^{n}|f(X_i) - g(X_i)| + c\left(\frac{md\log(n/d) + m\log(3/\delta)}{n}\right).$$

To simplify the statements of the lemmas to follow, for any finite class $\mathcal{G}$ and for any confidence parameter $\delta \in (0,1)$ define:

$$\alpha(\mathcal{G}, \delta) = 32\sqrt{\frac{m^2(\log(2|\mathcal{G}|) + \log(4/\delta))}{n}}, \tag{6.11}$$

where the sample size n and the value m (of Assumption 6.1) will always be clear from the context. The next technical lemma provides basic concentration properties of the median-of-means estimators, the proof of which follows from a combination of uniform Bernstein's inequality and a median-of-means deviation inequality for mean estimation [124, Theorem 2].

Lemma 6.2. *Suppose that Assumption 6.1 holds and let $S_n = (X_i, Y_i)_{i=1}^n$ denote an i.i.d. sample. Let $\mathcal{G}$ be any finite class of functions whose absolute value is bounded by m. Fix any $\delta \in (0,1)$, let $k = \lceil 8\log\frac{2|\mathcal{G}|^2}{\delta}\rceil$ and let α denote any upper bound on $\alpha(\mathcal{G}, \delta)$ defined in (6.11). Then, with probability at least $1 - \delta$, the following inequalities hold simultaneously for any $f, g \in \mathcal{G}$:*

$$\left|R(f) - R(g) - \mathrm{MOM}_{S_n}^k(\ell_f - \ell_g)\right| \leq \alpha\sqrt{\mathbf{E}(f(X) - g(X))^2},$$

$$\left|R(f) - R(g) - \mathrm{MOM}_{S_n}^k(\ell_f - \ell_g)\right| \leq \sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 + \alpha^2},$$

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 \leq 2\mathbf{E}(f(X) - g(X))^2 + \alpha^2.$$

For any class $\mathcal{G}$, define its L_2 diameter by:

$$\mathcal{D}(\mathcal{G}) = \sup_{f,g\in\mathcal{G}}\sqrt{\mathbf{E}(f(X) - g(X))^2}.$$

As a corollary of the above lemma, we are able to derive some basic properties of the random set $\widehat{\mathcal{F}}$ defined in the third step of our statistical procedure. In particular, we show that with high probability the set $\widehat{\mathcal{F}}$ contains the population risk minimizer over the ε-net $\overline{\mathcal{F}}_\varepsilon$. At the same time, we establish a uniform Bernstein-type bound on the excess risk of the elements of $\widehat{\mathcal{F}}$, with the role of the variance term played by $\mathcal{D}(\widehat{\mathcal{F}})$.

Lemma 6.3. *Suppose that Assumption 6.1 holds and let $S_n = (X_i, Y_i)_{i=1}^n$ denote an i.i.d. sample. Let $\mathcal{G}$ be any finite class of functions whose absolute value is bounded by m. Fix any $\delta \in (0,1)$, $k = \lceil 8 \log \frac{2|\mathcal{G}|^2}{\delta} \rceil$ and let α denote any upper bound on $\alpha(\mathcal{G}, \delta)$ defined in (6.11). Define the random subset of $\mathcal{G}$:*

$$\widehat{\mathcal{G}} = \left\{ f \in \mathcal{G} : \text{for every } g \in \mathcal{G},\ \mathrm{MOM}_{S_n}^k (\ell_f - \ell_g) \le \sqrt{2}\alpha \sqrt{\frac{1}{n}\sum_{i=1}^n (f(X_i) - g(X_i))^2 + \alpha^2} \right\},$$

Then, the following two conditions hold simultaneously, with probability at least $1 - \delta$:

1. *The function $g^* = \arg\min_{g \in \mathcal{G}} R(g)$ belongs to the class $\widehat{\mathcal{G}}$.*
2. *For any $f, g \in \widehat{\mathcal{G}}$, we have $R(f) - R(g^*) \le 4\alpha \mathcal{D}(\widehat{\mathcal{G}}) + 5\alpha^2$.*

Finally, we prove an excess risk bound for the min-max estimator in terms of the L_2 diameter of the set over which the estimator is computed. The intuitive implications of the following lemma are the following. First, if $\mathcal{D}(\widehat{\mathcal{F}})$ is of order $1/\sqrt{n}$, the lemma below immediately yields the fast rate of convergence for our estimator $\widehat{f}$. If, on the other hand, the diameter of $\mathcal{D}$ is much larger than $1/\sqrt{n}$, then we can exploit the curvature of the quadratic loss and the gain in the approximation error (due to considering the larger class $\widehat{\mathcal{F}}_+$ instead of $\widehat{\mathcal{F}}$) to prove the desired rate of convergence. See also the work of Lecué and Mendelson [112] for related discussions in the context of model selection aggregation for bounded problems.

Lemma 6.4. *Suppose that Assumption 6.1 holds and let $S_n = (X_i, Y_i)_{i=1}^n$ denote an i.i.d. sample. Let $\mathcal{G}$ be any finite class of functions whose absolute value is bounded by m. Fix any $\delta \in (0,1)$, let $k = \lceil 8 \log \frac{2|\mathcal{G}|^2}{\delta} \rceil$ and let α denote any upper bound on $\alpha(\mathcal{G}, \delta)$ defined in (6.11). Let $\widehat{f}$ be any estimator satisfying*

$$\widehat{f} \in \arg\min_{f \in \mathcal{G}} \max_{g \in \mathcal{G}} \mathrm{MOM}_{S_n}^k (\ell_f - \ell_g).$$

Let $g^* \in \arg\min_{g \in \mathcal{G}} R(g)$. Then, with probability at least $1 - \delta$, it holds that

$$R(\widehat{f}) \le R(g^*) + 2\alpha \mathcal{D}(\mathcal{G}).$$

We are now ready to prove Theorem 6.3.

Proof of Theorem 6.3. Our proof is split into two parts. First, we approximate the truncated linear class $\overline{\mathcal{F}}$ with a finite class, namely, an empirical L_1 ε-net constructed using the first third of the dataset denoted by S_1. Then, conditionally on S_1, we show that our estimator $\widehat{f}$ achieves the optimal rate of model selection aggregation over the finite class $\overline{\mathcal{F}}_\varepsilon$, in spite of the lack of assumptions on the covariates and the presence

of heavy-tailed labels. Finally, we note that if the number of median-of-means blocks k is equal to 0 (i.e., $n \lesssim d(\log(n/d) + \log(1/\delta))$), then we may output the 0 function which satisfies the desired bound for such sample sizes. Thus, in what follows we assume that $n \gtrsim d(\log(n/d) + \log(1/\delta))$.

The approximation step. Recall that $\overline{\mathcal{F}}_\varepsilon$ is an empirical L_1 ε-net of the truncated linear class $\overline{\mathcal{F}}$ constructed using the sample S_1. Let $f^* = \arg\min_{f \in \overline{\mathcal{F}}} R(f)$ and let f_ε^* be any element of $\overline{\mathcal{F}}_\varepsilon$ minimizing the empirical L_1 distance to f^*, that is, we have

$$\frac{1}{n} \sum_{X_i \in S_1} |f^*(X_i) - f_\varepsilon^*(X_i)| \le \varepsilon. \tag{6.12}$$

Let E_1 denote the event of Lemma 6.1 applied with respect to the sample S_1 (that contains n points) with the choice of the confidence parameter set to $\delta/3$ (thus, $\mathbf{P}(E_1) \ge 1 - \delta/3$). It follows that on the event E_1 we have

$$\begin{aligned}
&R(f_\varepsilon^*) - R(f^*) \\
&= 2\mathbf{E}Y(f^*(X) - f_\varepsilon^*(X)) + \mathbf{E}(f_\varepsilon^*(X)^2 - f^*(X)^2) \\
&\le 2\mathbf{E}(\mathbf{E}[Y|X](f^*(X) - f_\varepsilon^*(X))) + 2m\mathbf{E}|f_\varepsilon^*(X) - f^*(X)| \\
&\qquad \text{(since } |f_\varepsilon^*(X) + f^*(X)| \le 2m\text{)} \\
&\le 2\mathbf{E}(\sqrt{\mathbf{E}[Y^2|X]}|f^*(X) - f_\varepsilon^*(X)|) + 2m\mathbf{E}|f_\varepsilon^*(X) - f^*(X)| \\
&\qquad \text{(by Jensen's inequality)} \\
&\le 4m\mathbf{E}|f_\varepsilon^*(X) - f^*(X)| \\
&\qquad \text{(by Assumption 6.1)} \\
&\le 8m\varepsilon + 4mc_1 \left(\frac{md\log(n/d) + m\log(9/\delta)}{n} \right) \\
&\qquad \text{(by (6.12) and Lemma 6.1)} \\
&\le 12c_1 \left(\frac{m^2 d\log(n/d) + m^2\log(9/\delta)}{n} \right), \\
&\qquad \text{(by the definition of } \varepsilon\text{)}
\end{aligned}$$

where c_1 is an absolute constant. Observe that on the event E_1, any estimator $\widehat{f}$ satisfies

$$\begin{aligned}
R(\widehat{f}) - R(f^*) &\le R(\widehat{f}) - \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) + R(f_\varepsilon^*) - R(f^*) \\
&\le R(\widehat{f}) - \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) + 12c_1 \left(\frac{m^2 d\log(n/d) + m^2\log(9/\delta)}{n} \right).
\end{aligned}$$

From this point onward, we work on the event E_1. It thus remains to prove that with probability $1 - 2\delta/3$, the estimator $\widehat{f}$ computed using the remaining $2n$ points split into samples S_2 and S_3 satisfies

$$R(\widehat{f}) - \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) \lesssim \frac{m^2 d \log(n/d) + m^2 \log(1/\delta)}{n}. \tag{6.13}$$

Since $\overline{\mathcal{F}}_\varepsilon$ is a finite class of functions, we now turn to the aggregation part of this proof.

The aggregation step. By the L_2 covering number bound stated in [78, Theorem 9.4, Theorem 9.5], which also holds for the empirical L_1 distances, we have (see the proof of Lemma 6.1)

$$\log |\overline{\mathcal{F}}_\varepsilon| \lesssim d \log \frac{me}{\varepsilon} \lesssim d \log(n/d).$$

Note that $|\widehat{\mathcal{F}}_+|$ and $|\widehat{\mathcal{F}}|$ are simultaneously upper bounded by $|\overline{\mathcal{F}}_\varepsilon|^2$. For an arbitrary finite class $\mathcal{G}$, recall the definition of $\alpha(\mathcal{G}, \delta)$ stated in (6.11). It follows that there exists some absolute constant $c_2 > 0$ such that $\overline{\alpha}$ defined below satisfies

$$\max\Big(\alpha(\widehat{\mathcal{F}}, \delta/3), \alpha(\widehat{\mathcal{F}}_+, \delta/3)\Big) \le \overline{\alpha} = c_2 \sqrt{\frac{m^2 d \log(n/d) + m^2 \log(1/\delta)}{n}}. \tag{6.14}$$

Thus, $\overline{\alpha}$ defined above will be used in the applications of Lemmas 6.2, 6.3 and 6.4 to follow.

Let E_2 be the event of Lemma 6.3 applied for the set $\widehat{\mathcal{F}}$ with confidence parameter $\delta/3$. In particular, on the event E_2 we have

$$\arg\min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) \in \widehat{\mathcal{F}}, \text{ and for any } f \in \widehat{\mathcal{F}} \text{ we have } R(f) \le \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) + 4\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}) + 5\overline{\alpha}^2. \tag{6.15}$$

Conditionally on the sample S_2, let the set $\widehat{\mathcal{F}}$ defined in the third step of our algorithm be fixed. Denote $g^* = \arg\min_{g \in \widehat{\mathcal{F}}_+} R(f)$, where recall that $\widehat{\mathcal{F}}_+ = (\widehat{\mathcal{F}} + \widehat{\mathcal{F}})/2$. Observe that the L_2 diameters of $\widehat{\mathcal{F}}$ and $\widehat{\mathcal{F}}_+$ are equal, that is $\mathcal{D}(\widehat{\mathcal{F}}_+) = \mathcal{D}(\widehat{\mathcal{F}})$. Let E_3 be the event of Lemma 6.4 applied to the third part of our sample S_3 and the finite class $\widehat{\mathcal{F}}_+$ with the confidence parameter set to $\delta/3$. Thus, on E_3 our estimator $\widehat{f}$ satisfies:

$$R(\widehat{f}) \le R(g^*) + 2\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}). \tag{6.16}$$

Now choose any $g, h \in \widehat{\mathcal{F}}$ such that $\sqrt{\mathbf{E}(g(X) - h(X))^2} \ge \mathcal{D}(\widehat{\mathcal{F}})/2$ (such a choice always exists by definition of the diameter). Since $(g + h)/2 \in \widehat{\mathcal{F}}_+$, the parallelogram

identity yields

$$\begin{aligned} R(g^*) &\le R((g+h)/2) \\ &= \frac{1}{2}R(g) + \frac{1}{2}R(h) - \frac{1}{4}\mathbf{E}(g(X) - h(X))^2 \\ &\le \frac{1}{2}R(g) + \frac{1}{2}R(h) - \frac{1}{16}\mathcal{D}(\widehat{\mathcal{F}})^2. \end{aligned} \tag{6.17}$$

On the event E_2, applying (6.15) for the functions g and h we obtain

$$\frac{1}{2}R(g) + \frac{1}{2}R(h) \le \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) + 4\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}) + 5\overline{\alpha}^2.$$

Combining the above with equations (6.16) and (6.17) we have

$$R(\widehat{f}) - \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) \le 6\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}) + 5\overline{\alpha}^2 - \frac{1}{16}\mathcal{D}(\widehat{\mathcal{F}})^2 \le 149\overline{\alpha}^2,$$

where the last step follows by maximizing the quadratic equation with respect to $\mathcal{D}(\widehat{\mathcal{F}})$. Plugging in the definition of $\overline{\alpha}$ (see (6.14)) we obtain the desired inequality (6.13). The proof is complete by taking the union bound over the events E_1, E_2 and E_3 defined above. □

6.5.5 Proofs of Technical Lemmas

This section contains the proofs of lemmas appearing in Section 6.5.4. Note that rescaling the response Y by $1/m$ affects the excess risk by a multiplicative factor $1/m^2$. Thus, without loss of generality, in all the proofs of this section we may assume that Assumption 6.1 holds with $m = 1$.

6.5.5.1 Proof of Lemma 6.1

The proof of this lemma is based on a combination of the classical localization via empirical Rademacher complexities argument of [19] and the covering number bounds for truncated VC-subgraph classes due to [78].

First, define the star-hull of $|\overline{\mathcal{F}} - \overline{\mathcal{F}}| = \{|f - g| : f, g \in \overline{\mathcal{F}}\}$ by $\mathcal{H}$, and for $r \ge 0$, define its localized subsets by $\mathcal{H}_r$:

$$\mathcal{H} = \Big\{\beta|f-g| : \beta \in [0,1],\ \ f, g \in \overline{\mathcal{F}}\Big\}, \quad \mathcal{H}_r = \Big\{h \in \mathcal{H} :\ \frac{1}{n}\sum_{i=1}^n |h(X_i)|^2 \le 4r\Big\}.$$

Let $\widehat{\psi}_n(r) : [0, \infty) \to \mathbb{R}$ denote any sub-root function with unique positive fixed-point $\widehat{r}^*$ (that is, a positive solution to the equation $\widehat{\psi}_n(\widehat{r}^*) = \widehat{r}^*$ (see [19, Definition 3.1, Lemma 3.2]). Suppose that $\widehat{\psi}_n$ satisfies the following inequality for any $r \ge \widehat{r}^*$:

$$\frac{1}{n}\mathbf{E}_{\varepsilon_1,\ldots,\varepsilon_n} \sup_{h \in \mathcal{H}_r} \left(\sum_{i=1}^n \varepsilon_i h(X_i)\right) + \frac{\log(3/\delta)}{n} \lesssim \widehat{\psi}_n(r), \tag{6.18}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is a sequence of i.i.d. Rademacher random variables. Notice that for any $r \geq 0$ and any $h \in H_r$ we have $\sup_x |h(x)| \leq 2$ and $\mathbf{E}h(X)^2 \leq 4\mathbf{E}h(X)$. Hence, by the first part of [19, Theorem 4.1], with probability at least $1 - \delta$, the following holds simultaneously for all $f, g \in \overline{\mathcal{F}}$:

$$\mathbf{E}|f(X) - g(X)| \leq \frac{2}{n} \sum_{i=1}^{n} |f(X_i) - g(X_i)| + c\left(\widehat{r}^* + \frac{\log(3/\delta)}{n}\right), \tag{6.19}$$

where $c > 0$ is some universal constant.

In the rest of the proof we show that a suitable value of $\widehat{r}^*$ can be obtained by upper bounding the empirical Rademacher complexity terms via Dudley's entropy integral. To do so, we first need to obtain an upper bound on the covering numbers of the class $\mathcal{H}$ with respect to the empirical L_2 distance, defined between any $h, h' \in \mathcal{H}$ by $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(X_i) - h'(X_i))^2}$. In what follows, for any class $\mathcal{G}$ and any $\gamma > 0$, an empirical L_2 γ-net of $\mathcal{G}$ will be denoted by $N(\mathcal{G}, \gamma) \subseteq \mathcal{G}$. Thus, the covering number of $\mathcal{G}$ with respect to the empirical L_2 distance at scale γ is at most $|N(\mathcal{G}, \gamma)|$.

Since $\mathcal{H}$ is a star-hull of the class $|\overline{\mathcal{F}} - \overline{\mathcal{F}}|$, it follows from [139, Lemma 4.5] that for any $\gamma > 0$ we have

$$|N(\mathcal{H}, \gamma)| \leq \left|N(\overline{\mathcal{F}} - \overline{\mathcal{F}}, \gamma/2)\right| \cdot \frac{4}{\gamma}. \tag{6.20}$$

Further, noting that the Minkowski sum of $\gamma/4$ covers of $\overline{\mathcal{F}}$ forms a $\gamma/2$ cover of $\overline{\mathcal{F}} - \overline{\mathcal{F}}$ it follows that

$$|N(\overline{\mathcal{F}} - \overline{\mathcal{F}}, \gamma/2)| \leq |N(\overline{\mathcal{F}}, \gamma/4)|^2. \tag{6.21}$$

Let $\overline{\mathcal{F}}_+ = \{x \mapsto \max(0, f(X)) : f \in \overline{\mathcal{F}}\}$ and $\overline{\mathcal{F}}_- = \{x \mapsto \min(0, f(X)) : f \in \overline{\mathcal{F}}\}$. By the same argument, it holds that

$$|N(\overline{\mathcal{F}}, \gamma/4)| \leq |N(\overline{\mathcal{F}}_+, \gamma/8)| \cdot |N(\overline{\mathcal{F}}_-, \gamma/8)|. \tag{6.22}$$

Finally, plugging in the upper bounds on the covering numbers of $\overline{\mathcal{F}}_+$ and $\overline{\mathcal{F}}_-$ due to [78, Theorem 9.4, Theorem 9.5][2], the chain of inequalities (6.20), (6.21) and (6.22) yields

$$\log|N(\mathcal{H}, \gamma)| \lesssim d\log(e/\gamma).$$

[2]See also the proof of [78, Theorem 11.3] where the same bound on covering numbers is used.

Plugging in the above inequality into Dudley's entropy integral [71, Theorem 3.5.1] upper bound on Rademacher complexities, we obtain

$$\begin{aligned}\frac{1}{n}\mathbf{E}_{\varepsilon_1,\dots,\varepsilon_n}\sup_{h\in\mathcal{H}_r}\left(\sum_{i=1}^n \varepsilon_i h(X_i)\right) &\lesssim \frac{1}{\sqrt{n}}\int_0^{2\sqrt{r}}\sqrt{d\log(e/\gamma)}d\gamma \\ &\lesssim \sqrt{\frac{d}{n}}\sqrt{r\log(e/r)}\left(\mathbb{1}_{\{r\geq d/n\}}+\mathbb{1}_{\{r<d/n\}}\right) \\ &\lesssim \sqrt{\frac{dr\log(n/d)}{n}}+\frac{d\sqrt{\log(n/d)}}{n}.\end{aligned}$$

In particular, the inequality (6.18) is satisfied by the choice:

$$\widehat{\psi}_n(r) = c\left(\sqrt{\frac{dr\log(n/d)}{n}}+\frac{d\sqrt{\log(n/d)}+\log(3/\delta)}{n}\right).$$

Solving the fixed-point equation $\widehat{\psi}_n(\widehat{r}^*)=\widehat{r}^*$ yields $\widehat{r}^*\lesssim \frac{d\log(n/d)+\log(1/\delta)}{n}$. The claim follows by the localization theorem stated in (6.19).

6.5.5.2 Proof of Lemma 6.2

Fix any $f,g\in\mathcal{G}$ and recall that $\mathbf{E}(\ell_f-\ell_g)=R(f)-R(g)$. By the standard bound [124, Theorem 2], for any $\delta'\in(0,1)$, the choice $k(\delta')=\lceil 8\log(1/\delta')\rceil$ guarantees that with probability at least $1-\delta'$ we have

$$\left|R(f)-R(g)-\mathrm{MOM}_{S_n}^{k(\delta')}(\ell_f-\ell_g)\right| \leq \sqrt{\frac{32\mathrm{Var}(\ell_f-\ell_g)\log(1/\delta')}{n}}. \tag{6.23}$$

To upper bound the variance term, first notice that

$$\ell_f(X,Y)-\ell_g(X,Y) = 2Y(g(X)-f(X))+f(X)^2-g(X)^2.$$

Combining the above identity with the inequality $(a+b)^2\leq 2a^2+2b^2$ for any a,b, Assumption 6.1 (with $m=1$) and the boundedness of f,g, we obtain

$$\begin{aligned}\mathrm{Var}(\ell_f-\ell_g) &\leq 8\mathbf{E}Y^2(g(X)-f(X))^2+2\mathbf{E}(f(X)^2-g(X)^2)^2 \\ &\leq 8\mathbf{E}(g(X)-f(X))^2+2\mathbf{E}(f(X)-g(X))^2(f(X)+g(X))^2 \\ &\leq 16\mathbf{E}(g(X)-f(X))^2.\end{aligned}$$

Since the class $\mathcal{G}$ is finite, taking $\delta'=\delta/(2|\mathcal{G}|^2)$ the upper bound (6.23) extend uniformly to all pairs $f,g\in\mathcal{G}$, with probability at least $1-\delta/2$. In particular, for any

$f, g \in \mathcal{G}$ it holds that

$$\begin{aligned}
&\left|R(f) - R(g) - \mathrm{MOM}_{S_n}^{k(\delta')}(\ell_f - \ell_g)\right| \\
&\leq \sqrt{\frac{512\mathbf{E}(g(X) - f(X))^2 \cdot (2\log(|\mathcal{G}|) + \log(2/\delta))}{n}} \\
&\leq \alpha\sqrt{\mathbf{E}(g(X) - f(X))^2}. \qquad (6.24)
\end{aligned}$$

This completes the proof of the first inequality.

We will now simultaneously prove the second and the third inequalities appearing in the statement of this lemma. Note that $m = 1$ ensures that for any $f, g \in \mathcal{G}$ we have $(f(X) - g(X))^2 \leq 4$ and $\mathbf{E}(f(X) - g(X))^4 \leq 4\mathbf{E}(f(X) - g(X))^2$. Hence, for any $\delta'' \in (0, 1)$ and any $f, g \in \mathcal{G}$, Bernstein's inequality ensures that with probability at least $1 - 2\delta''$ it holds simultaneously that

$$\mathbf{E}(f(X) - g(X))^2 \leq \frac{2}{n}\sum_{i=1}^{n}(f(X) - g(X))^2 + \frac{12\log(1/\delta'')}{n}, \qquad (6.25)$$

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 \leq 2\mathbf{E}(f(X) - g(X))^2 + \frac{12\log(1/\delta'')}{n}. \qquad (6.26)$$

Setting $\delta'' = \delta/(4|\mathcal{G}|^2)$ the above inequalities extend uniformly to all pairs $f, g \in \mathcal{G}$ with probability at least $1 - \delta/2$. Noting that $\frac{12\log(1/\delta'')}{n} \leq \alpha^2$, the inequality (6.26) completes the proof of the third inequality of this lemma. Finally, the second inequality appearing in the statement of this lemma is implied (on the event of the first and third inequalities) by plugging in (6.25) into (6.24) together with the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ valid for any $a, b \geq 0$. The proof of this lemma is thus complete.

6.5.5.3 Proof of Lemma 6.3

Let E denote the event of Lemma 6.2 (thus, $\mathbf{P}(E) \geq 1 - \delta$). By the definition of g^*, for any $g \in \mathcal{G}$ we have $R(g^*) - R(g) \leq 0$. Hence, on the event E it holds simultaneously for all $g \in \mathcal{G}$ that

$$\begin{aligned}
\mathrm{MOM}_{S_n}^{k}(\ell_{g^*} - \ell_g) &\leq R(g^*) - R(g) + |R(g^*) - R(g) - \mathrm{MOM}_{S_n}^{k}(\ell_{g^*} - \ell_g)| \\
&\leq \sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^{n}(g^*(X_i) - g(X_i))^2} + \alpha^2.
\end{aligned}$$

In particular, on the event E the function $g^* \in \widehat{\mathcal{G}}$, which completes the first part of the proof.

We now turn to proving the second part of this lemma. Since $g^* \in \widehat{\mathcal{G}}$, by the definition of $\widehat{\mathcal{G}}$, for any $g \in \widehat{\mathcal{G}}$ we have

$$\mathrm{MOM}^k_{S_n}(\ell_g - \ell_{g^*}) \le \sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^n (g(X_i) - g^*(X_i))^2} + \alpha^2.$$

Hence, on the event E, by the third inequality of Lemma 6.2, for any $g \in \widehat{\mathcal{G}}$ it holds that

$$\begin{aligned} R(g) - R(g^*) &\le \left|R(g) - R(g^*) - \mathrm{MOM}^k_{S_n}(\ell_g - \ell_{g^*})\right| + \mathrm{MOM}^k_{S_n}(\ell_g - \ell_{g^*}) \\ &\le 2\sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^n (g(X_i) - g^*(X_i))^2} + 2\alpha^2 \\ &\le 4\alpha\sqrt{\mathbf{E}(g(X) - g^*(X))^2} + 5\alpha^2. \end{aligned}$$

By the definition of the L_2 diameter of the class $\widehat{\mathcal{G}}$ and by the fact that $g^*, g \in \widehat{\mathcal{G}}$, it follows that $\sqrt{\mathbf{E}(g(X) - g^*(X))^2} \le \mathcal{D}(\widehat{\mathcal{G}})$ and hence our proof is complete.

6.5.5.4 Proof of Lemma 6.4

First observe that

$$\begin{aligned} R(\widehat{f}) &= R(g^*) + \left(R(\widehat{f}) - R(g^*) - \mathrm{MOM}^k_{S_n}(\ell_{\widehat{f}} - \ell_{g^*})\right) + \mathrm{MOM}^k_{S_n}(\ell_{\widehat{f}} - \ell_{g^*}) \\ &\le R(g^*) + \sup_{g\in\mathcal{G}} \left|R(g) - R(g^*) - \mathrm{MOM}^k_{S_n}(\ell_g - \ell_{g^*})\right| + \mathrm{MOM}^k_{S_n}(\ell_{\widehat{f}} - \ell_{g^*}) \\ &\le R(g^*) + \alpha\mathcal{D}(\mathcal{G}) + \mathrm{MOM}^k_{S_n}(\ell_{\widehat{f}} - \ell_{g^*}), \end{aligned} \tag{6.27}$$

where the last line follows via an application of Lemma 6.2. Further, notice that by the definition of $\widehat{f}$ we have

$$\mathrm{MOM}^k_{S_n}(\ell_{\widehat{f}} - \ell_{g^*}) \le \max_{g\in\widehat{\mathcal{G}}} \mathrm{MOM}^k_{S_n}(\ell_{\widehat{f}} - \ell_g) \le \max_{g\in\widehat{\mathcal{G}}} \mathrm{MOM}^k_{S_n}(\ell_{g^*} - \ell_g).$$

At the same time, on the event of Lemma 6.2, for all $g \in \mathcal{G}$ we have

$$\mathrm{MOM}^k_{S_n}(\ell_{g^*} - \ell_g) \le R(g^*) - R(g) + \alpha\mathcal{D}(\mathcal{G}) \le \alpha\mathcal{D}(\mathcal{G}).$$

Combining the above inequality with (6.27) concludes our proof.

6.6 Limitations and Open Directions

The results presented in this chapter offer two natural directions for further extensions.

The first direction concerns the computational challenges associated with the deviation-optimal robust estimator presented in Section 6.4.4. Indeed, the statistical

estimator proposed therein is computationally intractable for two reasons. The first reason is the pre-processing step which replaces the convex class $\mathcal{G}_{\text{lin}}$ with the non-convex class of truncated functions $\overline{\mathcal{F}} = \{f_m : f \in \mathcal{G}_{\text{lin}}\}$. Due to the non-convexity of $\overline{\mathcal{F}}$, even far simpler procedures than computing median of means, such as computing the empirical risk minimizer over $\mathcal{F}$, become computationally intractable. The second reason is that we employ a model selection aggregation procedure over a class $\overline{\mathcal{F}}_\varepsilon$, whose cardinality is exponential in the dimension d. Observe that both of these challenges extend beyond the known issues associated with the computation of statistical estimators based on the median of means tournaments principle. That is, the approach adopted in Section 6.4.4 would suffer from computational issues even if we replaced Assumption 6.1 with an assumption of almost surely bounded response variable Y (and if we modified the procedure of Section 6.4.4 accordingly).

The second direction for extending our results concerns the question of adaptivity to the constant m appearing in Assumption 6.1. In particular, the deviation-optimal robust estimator of Section 6.4.4, as well as the expectation-optimal truncated least squares estimator of Section 6.4.1, both rely on the knowledge of m. In contrast, the shrinkage mechanism employed by the Forster-Warmuth procedure attains the expectation-optimal bound $m^2 d/n$ without the knowledge of m. Thus, computational issues aside, it would be interesting to obtain an estimator attaining the optimal accuracy/confidence trade-off without the knowledge of m, perhaps exploiting the ideas present in the analysis of the Forster-Warmuth algorithm.

6.7 Bibliographic Remarks

This chapter is adapted from the joint paper [154] written in collaboration with Jaouad Mourtada and Nikita Zhivotovskiy. For the purposes of this thesis, some material present in [154] was omitted. We refer the interested reader to [154] for results in connection to deviation-optimal linear regression under an additional assumption that the covariance structure $\mathbf{E}[XX^\top]$ is known to the statistician, and to some technical discussions of open problems related to extending the scope of Theorem 6.3 beyond VC-subgraph classes.

To situate the results of this chapter in a broader context, we conclude with a summary of related work taken from [154, Section 1.2]. We split the remaining discussion into three separate paragraphs. The first paragraph discusses work related to the analysis of least squares estimators. However, least squares estimators are not deviation-optimal in heavy-tailed settings, which leads us to the second paragraph

discussing literature on robustness to heavy-tailed distributions. Finally, in the third paragraph, we comment on some works related to distribution-free bounds.

Analysis of least squares estimators. The most standard approach to regression problems is the least squares principle, where one selects the predictor achieving the best fit to data within some predefined class of functions. A large body of work is devoted to analyzing and obtaining guarantees on its performance, in its most classical form, relying on the fact that the empirical risk provides a good approximation of its population counterpart. This is typically established when the underlying distribution is sufficiently well-behaved (for instance, bounded or light-tailed), using tools from empirical process theory. For this point of view to statistical learning, we refer to the standard textbooks [202, 133, 106, 216]. It should be noted that statistical analysis of linear regression has also been treated via a complementary approach of stochastic approximation; see, for instance, the works [217, 77, 64] and references therein.

A recent line of research has established that empirical minimization can perform well under significantly weaker assumptions. Our starting point is the work of Oliveira [163], where in the context of linear regression the usual sub-Gaussian assumption on X is replaced by a significantly weaker L_4–L_2 moment equivalence assumption (6.1). In particular, such an assumption does not even require the existence of any moments of X higher than the fourth. Variations of this assumption have become the standard tool in the recent literature on linear regression [114, 87, 45, 126, 100, 152, 54, 167]. The seminal work of Mendelson [142] introduced a more general condition, called the *small-ball* assumption. In most of the aforementioned papers, the analysis is performed for empirical risk minimization, which usually does not lead to the optimal accuracy/confidence trade-off. The papers [8, 9] provide the optimal confidence for ERM, albeit under stronger moment equivalence assumptions than that of (6.1). The L_4–L_2 moment equivalence is also important in the robust covariance estimation problem [45, 148, 165].

It has been recently observed that the absolute constants involved in the moment equivalence and the small-ball assumptions can behave badly in some cases. First, Saumard [184] shows that the small-ball condition is unsuitable for some important classes leading to suboptimal performance of ERM. Further, the work [45] (see also the discussion in [163] and [114]) discusses that the kurtosis constant κ in the moment assumptions similar to (6.1) can depend on the dimension and affect the bounds negatively. Indeed, in Chapter 5 of this thesis, we discussed this suboptimal behavior in the context of linear regression, even in a favorable setup where both X and Y

are almost surely bounded. There is a growing interest in further relaxing these assumptions and refining the underlying methods [184, 46, 146, 145, 55, 152, 147]. In particular, the works [46, 147] replace moment equivalence assumptions by the bounds on the L_p moments for $p \geq 4$. This is closer to the setting targeted at this chapter.

Robustness to heavy-tailed distributions. In a broad sense, robustness encompasses the study and design of statistical estimation procedures exhibiting certain stability properties under the existence of "outlier" points in the observed sample. For a classical perspective on robustness, originating from the work of Tukey [200] and building on the ideas of contaminated models, influence functions and breakdown points, we refer to the standard books [79, 89, 182].

In contrast to the classical perspective, our work falls within the recent body of work initiated by Catoni [44], where the term robustness is to be understood specifically as robustness to heavy-tailed distributions (rather than, for example, adversarial contamination of the sample). The starting point of this direction is the question of mean estimation, where informally, one aims to construct statistical estimators performing as well as the sample mean does for Gaussian samples, all while making as weak distributional assumptions as possible. Several ways of constructing such estimators (called sub-Gaussian estimators) have been proposed in the literature. The most widespread approach is based on the median-of-means estimators, which appeared first independently in [157, 91, 4] and were further developed in the works of [149, 62, 125, 127]. Other techniques include the Catoni's estimator and its extensions [44, 46] or the trimmed means [129]. We refer to the survey [124] for further details and references. For a complementary survey focusing on the computational aspects see [63].

The central ideas behind the robust mean estimation found their applications in many related problems such as regression [87, 37, 55, 126, 111, 55, 150, 145], covariance estimation [45, 148, 165] and clustering [37, 103]. In the context of linear regression, the first works showing the optimal accuracy/confidence trade-off under weak assumptions are attributed to Audibert and Catoni [8, 9] and were further extended in [45, 46]; these papers are based on PAC-Bayesian truncations.

Distribution-free linear regression. Distribution-free non-asymptotic excess risk bounds take their roots in the PAC-learning framework [203, 201], where historically the binary loss is studied the most. Because of its boundedness, excess risk bounds in such setups can be obtained without any assumptions on the distribution of (X, Y).

In the context of non-parametric regression with the squared loss, only asymptotic consistency results are possible under truly minimal assumptions on the underlying distribution (see the book [78]). In fact, the standard notions of universal consistency [78, Section 1.6 and Chapter 10] involve only the assumption $\mathbf{E}Y^2 < \infty$ and no assumptions on the distribution of X. The distribution-free nature of this notion is one of our motivations. A notable non-asymptotic result in this direction is [78, Theorem 11.3], where an inexact oracle inequality (6.4) is proved without any explicit assumptions on the distribution of X.

Another direction originates from the online learning literature (see [47] for background on this topic). For instance, when both X and Y are bounded, the renowned Vovk-Azoury-Warmuth forecaster [215, 11] can be used to provide excess risk bounds of order d/n in our setup even when the aforementioned moment equivalence constants behave badly with respect to the dimension. This observation has been explored in Chapter 5 of this thesis. For linear regression, the Forster-Warmuth algorithm [66], which is in turn a modification of the Vovk-Azoury-Warmuth forecaster, leads to the only known exact oracle inequality (prior to the results presented in [154], on which this chapter is based) without imposing any assumptions on X.

References

[1] Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, 2008.

[2] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1370–1378, 2019.

[3] Alnur Ali, Edgar Dobriban, and Ryan J Tibshirani. The implicit regularization of stochastic gradient flow for least squares. *arXiv preprint arXiv:2003.07802*, 2020.

[4] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58 (1):137–147, 1999.

[5] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7411–7422, 2019.

[6] Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

[7] Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

[8] Jean-Yves Audibert and Olivier Catoni. Linear regression through pac-bayesian truncation. *arXiv preprint arXiv:1010.0072*, 2010.

[9] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

[10] Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *International Conference on Learning Representations*, 2019.

[11] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43 (3):211–246, 2001.

[12] Gábor Balázs, András György, and Csaba Szepesvári. Chaining bounds for empirical risk minimization. *arXiv preprint arXiv:1609.01872*, 2016.

[13] Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019.

[14] Andrew R Barron. Are bayes rules consistent in information? In *Open Problems in Communication and Computation*, pages 85–91. Springer, 1987.

[15] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3 (Nov):463–482, 2002.

[16] Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

[17] Peter L Bartlett and Mikhail Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368, 2007.

[18] Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

[19] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[20] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[21] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3): 167–175, 2003.

[22] Pierre C Bellec. Optimal exponential bounds for aggregation of density estimators. *Bernoulli*, 23(1):219–248, 2017.

[23] Pierre C Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

[24] Denis Belomestny and John Schoenmakers. *Advanced Simulation-Based Methods for Optimal Stopping and Control: With Applications in Finance*. Springer, 2018.

[25] Peter J Bickel, Ya'acov Ritov, Alon Zakai, and Bin Yu. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7(5), 2006.

[26] Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pages 273–325, 2006.

[27] Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

[28] Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

[29] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. A sharp concentration inequality with applications. *Random Structures & Algorithms*, 16(3):277–292, 2000.

[30] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9: 323–375, 2005.

[31] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[32] Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[33] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[34] Olivier Bousquet and Nikita Zhivotovskiy. Fast classification rates without standard margin assumptions. *Information and Inference: A Journal of the IMA*, 2021.

[35] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[36] Leo Breiman and David Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381): 131–136, 1983.

[37] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

[38] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[39] Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, James R Lee, and Aleksander Mądry. K-server via multiscale entropic regularization. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pages 3–16, 2018.

[40] Sébastien Bubeck, Michael B Cohen, James R Lee, and Yin Tat Lee. Metrical task systems on trees via mirror descent and unfair gluing. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 89–97. SIAM, 2019.

[41] Peter Bühlmann and Bin Yu. Boosting with the l2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[42] Olivier Catoni. The mixture approach to universal model selection. In *École Normale Supérieure*. Citeseer, 1997.

[43] Olivier Catoni. Pac-bayesian supervised classification: The thermodynamics of statistical learning, 2008.

[44] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

[45] Olivier Catoni. Pac-bayesian bounds for the gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

[46] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

[47] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[48] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[49] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015.

[50] Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.

[51] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[52] Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

[53] Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

[54] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020.

[55] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and related fields*, pages 1–44, 2019.

[56] Albert Cohen, Mark A Davenport, and Dany Leviatan. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13(5):819–834, 2013.

[57] Fabienne Comte and Valentine Genon-Catalot. Regression function estimation on non compact support in an heteroscesdastic model. *Metrika*, 83(1):93–128, 2020.

[58] Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

[59] Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2): 202–207, 1979.

[60] Luc Devroye, László Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 1997.

[61] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[62] Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[63] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[64] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[65] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[66] Jürgen Forster and Manfred K Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.

[67] Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference On Learning Theory*, volume 75, pages 167–208, 2018.

[68] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[69] Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. *arXiv preprint arXiv:1902.01903*, 2019.

[70] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in deep linear neural networks. *arXiv preprint arXiv:1904.13262*, 2019.

[71] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40. Cambridge University Press, 2016.

[72] E Gobet and P Turkedjiev. Adaptive importance sampling in least-squares Monte Carlo algorithms for backward stochastic differential equations. *Stochastic Processes and their Applications*, 127(4):1171–1203, 2017.

[73] Emmanuel Gobet. *Monte-Carlo Methods and Stochastic processes: from Linear to Non-linear*. CRC Press, 2016.

[74] Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning: Implications to exp-concave empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):8245–8257, 2017.

[75] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[76] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[77] László Györfi and Harro Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.

[78] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media, 2002.

[79] Frank R Hampel, Peter J Rousseeuw, Elvezio M Ronchetti, and Werner A Stahel. *Robust statistics: the approach based on influence functions*. Wiley, 1980.

[80] Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

[81] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[82] David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting {0, 1}-functions on randomly drawn points. *Information and Computation*, 115(2): 248–292, 1994.

[83] David Haussler, Jyrki Kivinen, and Manfred K Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

[84] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[85] Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209, 2014.

[86] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[87] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

[88] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

[89] Peter J Huber. Robust statistics. *Wiley Series in Probability and Mathematical Statistics*, 1981.

[90] Valentin Konstantinovich Ivanov. On linear problems which are not well-posed. In *Doklady akademii nauk*, volume 145, pages 270–272. Russian Academy of Sciences, 1962.

[91] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[92] Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Conference on Learning Theory*, 2020.

[93] Wenxin Jiang. Process consistency for adaboost. *The Annals of Statistics*, 32 (1):13–29, 2004.

[94] Anatoli Juditsky, Philippe Rigollet, and Alexandre B Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[95] Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

[96] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: risk bounds, margin bounds, and regularization. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 793–800, 2008.

[97] Michael J Kearns and Umesh V Vazirani. An introduction to computational learning theory, 1994.

[98] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and computation*, 132(1):1–63, 1997.

[99] Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[100] Adam Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *(Extended abstract) Proceedings of the 31st Conference On Learning Theory*, pages 1420–1430, 2018.

[101] Yegor Klochkov and Nikita Zhivotovskiy. Uniform Hanson-Wright type concentration inequalities for unbounded entries via the entropy method. *Electronic Journal of Probability*, 25, 2020.

[102] Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 34, 2021.

[103] Yegor Klochkov, Alexey Kroshnin, and Nikita Zhivotovskiy. Robust k-means clustering for distributions with two moments. *The Annals of Statistics (forthcoming)*, 2020.

[104] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[105] Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[106] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[107] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

[108] Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 1477–1485, 2015.

[109] Gil Kur, Alexander Rakhlin, and Adityanand Guntuboyina. On suboptimality of least squares with application to estimation of convex bodies. In *Conference on Learning Theory*, pages 2406–2424, 2020.

[110] Louis Landweber. An iteration formula for fredholm integral equations of the first kind. *American journal of mathematics*, 73(3):615–624, 1951.

[111] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, 48(2):906–931, 2020.

[112] Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613, 2009.

[113] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

[114] Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[115] Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.

[116] Johannes Lederer and Sara van de Geer. New concentration inequalities for suprema of empirical processes. *Bernoulli*, 20(4):2020–2038, 2014.

[117] Michel Ledoux. On talagrand's deviation inequalities for product measures. *ESAIM: Probability and statistics*, 1:63–87, 1997.

[118] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[119] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

[120] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

[121] Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *Proceedings of the 35th International Conference on Machine Learning*, number CONF, 2018.

[122] Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

[123] Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative regularization for learning with convex loss functions. *The Journal of Machine Learning Research*, 17(1):2718–2755, 2016.

[124] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[125] Gábor Lugosi and Shahar Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 175:957–973, 2019.

[126] Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019.

[127] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

[128] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc*, 22:925–965, 2020.

[129] Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

[130] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697, 2004.

[131] Pascal Massart. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

[132] Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

[133] Pascal Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer-Verlag Berlin Heidelberg, 2007.

[134] Simon Matet, Lorenzo Rosasco, Silvia Villa, and Bang Long Vu. Don't relax: early stopping for convex regularization. *arXiv preprint arXiv:1707.05422*, 2017.

[135] Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

[136] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

[137] David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3): 355–363, 1999.

[138] Brendan McMahan and Matthew Streeter. Open problem: Better bounds for online logistic regression. In *Conference on Learning Theory*, pages 44–1, 2012.

[139] Shahar Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

[140] Shahar Mendelson. Empirical processes with a bounded ψ 1 diameter. *Geometric and Functional Analysis*, 20(4):988–1027, 2010.

[141] Shahar Mendelson. Learning without concentration. *Journal of the ACM*, 62(3), 2015. ISSN 0004-5411. doi: 10.1145/2699439.

[142] Shahar Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):1–25, 2015.

[143] Shahar Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3-4):641–674, 2017.

[144] Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1):459–502, 2018.

[145] Shahar Mendelson. An unrestricted learning procedure. *Journal of the ACM (JACM)*, 66(6):1–42, 2019.

[146] Shahar Mendelson. Extending the scope of the small-ball method. *Studia Mathematica*, 2020.

[147] Shahar Mendelson. Learning bounded subsets of L_p. *arXiv preprint arXiv:2002.01182*, 2020.

[148] Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under L_4-L_2 norm equivalence. *Annals of Statistics*, 48(3):1648–1664, 2020.

[149] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[150] Stanislav Minsker and Timothée Mathieu. Excess risk bounds in robust empirical risk minimization. *arXiv preprint arXiv:1910.07485*, 2019.

[151] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2018.

[152] Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics (forthcoming); arXiv preprint arXiv:1912.10754*, 2019.

[153] Jaouad Mourtada and Stéphane Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research (forthcoming); arXiv preprint arXiv:1912.10784*, 2019.

[154] Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 2022.

[155] Arkadi Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

[156] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[157] Arkadiĭ Nemirovsky and David Yudin. *Problem complexity and method efficiency in optimization.* Wiley, New York, 1983.

[158] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[159] Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. *arXiv preprint arXiv:1802.08009*, 2018.

[160] Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In *Conference on Learning Theory*, pages 3030–3048. PMLR, 2020.

[161] Albert B Novikoff. On convergence proofs for perceptrons. Technical report, STANFORD RESEARCH INST MENLO PARK CA, 1963.

[162] Roberto Oliveira. Sums of random hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15:203–212, 2010.

[163] Roberto Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3-4): 1175–1194, 2016.

[164] Stanley Osher, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016.

[165] Dmitrii M Ostrovskii and Alessandro Rudi. Affine invariant covariance estimation for heavy-tailed distributions. In *Conference on Learning Theory*, pages 2531–2550, 2019.

[166] Nicolò Pagliana and Lorenzo Rosasco. Implicit regularization of accelerated methods in hilbert spaces. In *Advances in Neural Information Processing Systems*, pages 14454–14464, 2019.

[167] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

[168] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[169] Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3806–3832. PMLR, 15–19 Aug 2021.

[170] Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

[171] Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[172] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

[173] Dominic Richards and Patrick Rebeschini. Optimal statistical rates for decentralised non-parametric regression with linear speed-up. In *Advances in Neural Information Processing Systems*, pages 1214–1225, 2019.

[174] Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(34):1–44, 2020.

[175] Philippe Rigollet. Kullback-leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, pages 639–665, 2012.

[176] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[177] R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[178] William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

[179] Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[180] F Rosenblatt. Principles of neurodynamics. *Perceptrons and the Theory of Brain Mechanisms*, 1962.

[181] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[182] Peter J Rousseeuw and Annick M Leroy. *Robust Regression and Outlier Detection*, volume 589. John Wiley & Sons, 2005.

[183] Mark Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

[184] Adrien Saumard. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203, 2018.

[185] Robert E Schapire and Yoav Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

[186] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[187] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[188] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[189] Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *The Journal of Machine Learning Research*, 16(1):3475–3486, 2015.

[190] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pages 10608–10619, 2018.

[191] Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

[192] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

[193] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[194] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[195] Andrei Nikolaevich Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.

[196] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[197] A. B. Tsybakov. Discussion: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2681 – 2687, 2006. doi: 10.1214/009053606000001064.

[198] Alexandre B Tsybakov. Optimal rates of aggregation. *Conference on Learning Theory*, pages 303–313, 2003.

[199] Alexandre B Tsybakov. Introduction to nonparametric estimation., 2009.

[200] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

[201] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27 (11):1134–1142, 1984.

[202] Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[203] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka. Moscow., 1974.

[204] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[205] Vladimir Vapnik and Alexey Chervonenkis. Uniform convergence of frequencies of occurence of events to their probabilities. In *Dokl. Akad. Nauk SSSR*, volume 181, pages 781–783, 1968.

[206] Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

[207] Vladimir Vapnik and Alexey Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.

[208] VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

[209] Tomas Vaškevičius and Nikita Zhivotovskiy. Suboptimality of constrained least squares and improvements via non-linear predictors. *Bernoulli (forthcoming); arXiv preprint arXiv:2009.09304*, 2020.

[210] Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pages 2968–2979, 2019.

[211] Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 253–264, 2020.

[212] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[213] Suhas Vijaykumar. Localization, convexity, and star aggregation. *Advances in Neural Information Processing Systems*, 34, 2021.

[214] Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

[215] Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

[216] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[217] H Walk and L Zsidó. Convergence of the robbins-monro method for linear problems in a banach space. *Journal of Mathematical Analysis and Applications*, 139(1):152–177, 1989.

[218] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction and estimation in unbounded domain. In *Uncertainty in Artificial Intelligence*, 2018.

[219] Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Transactions on Information Theory*, 65(10):6685–6703, 2019.

[220] Olivier Wintenberger. Optimal learning with bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

[221] Blake Woodworth, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Kernel and deep regimes in overparametrized models. *arXiv preprint arXiv:1906.05827*, 2019.

[222] Yuhong Yang. Combining different procedures for adaptive regression. *Journal of multivariate analysis*, 74(1):135–161, 2000.

[223] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[224] Daniel Z Zanger. Quantitative error estimates for a least-squares Monte Carlo algorithm for American option pricing. *Finance and Stochastics*, 17(3):503–534, 2013.

[225] Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

[226] Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via Hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2019.

[227] Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *arXiv preprint arXiv:2108.08198*, 2021.

[228] Nikita Zhivotovskiy and Steve Hanneke. Localization of vc classes: Beyond local rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.