Convex Regression: Theory, Practice, and Applications

by

Gábor Balázs

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistical Machine Learning

Department of Computing Science

University of Alberta

Abstract

This thesis explores theoretical, computational, and practical aspects of convex (shape-constrained) regression, providing new excess risk upper bounds, a comparison of convex regression techniques with theoretical guarantee, a novel heuristic training algorithm for max-affine representations, and applications in convex stochastic programming.

The new excess risk upper bound is developed for the general empirical risk minimization setting without any shape constraints, and provides a probabilistic guarantee for cases with unbounded hypothesis classes, targets, and noise models. The strength of the general result is demonstrated by applying it to linear regression under the squared loss both for lasso and ridge regression, as well as for convex nonparametric least squares estimation, in each case allowing one to obtain near-minimax upper bounds on the risk.

Next, cutting plane and alternating direction method of multipliers algorithms are compared for training the max-affine least squares estimators; estimators for which we provide explicit excess risk bounds. These techniques are also extended for the partitioned convex formulation (which is shown to enjoy optimal minimax rates). We also provide an empirical study of various heuristics for solving the non-convex optimization problem underlying the partitioned convex formulation.

A novel max-affine estimator is designed, which scales well for large sample sizes and improves the generalization error of current techniques in many cases. Its training time is proportional to the adaptively set model size, making it computationally attractive for estimation problems where the target can be efficiently approximated by max-affine functions.

Realistic convex regression applications are synthetized for the convex stochastic programming framework such as an energy storage optimization using a solar source with an Economy 7 tariff pricing model, as well as a multi-product assembly problem of operating a beer brewery.

Acknowledgements

I am very grateful to my supervisors, Csaba Szepesvári and Dale Schuurmans, for their endless support, patient teaching, and flexible guidance. I also thank András György for his invaluable help during the hard times.

I have been very lucky to enjoy the support of the Computing Science Department of the University of Alberta, the Reinforcement Learning and Artificial Intelligence (RLAI) group, and the Alberta Innovates Centre for Machine Learning (AICML).

I am also indebted to Russel Greiner, Ivan Mizera, and Bo Pang, for pointing out a few mistakes in the dissertation.

Finally, I thank my family and friends for their patience and unconditional support over this long time.

Table of Contents

List of Figures

List of Algorithms

List of Notations

$\mathbb{N}$	natural numbers (without zero), that is $\mathbb{N} \doteq \{1, 2, 3, \ldots\}$
$\mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{R}_{>0}$	set of real, nonnegative real and positive real numbers
$\{\mathbb{X} \to \mathbb{R}\}$	set of all functions mapping from $\mathbb{X}$ to $\mathbb{Y}$
$\mathbf{0}, \mathbf{1}$	full zero and full one vectors of appropriate dimension
$\mathbf{0}_{m\times n}, \mathbf{1}_{m\times n}$	full zero and full one matrices of size $m \times n$
I_n	identity matrix of size $n \times n$
$0 \preceq M$	square matrix M is (symmetric) positive semi-definite
$0 \prec M$	square matrix M is (symmetric) positive definite
$\|\cdot\|_p$	vector and matrix p-norms for $p \in \mathbb{N} \cup \{\infty\}$
$\|\cdot\|$	vector and matrix Euclidean norm, that is $\|\cdot\| \doteq \|\cdot\|_2$
$\mathcal{B}_p(\boldsymbol{x}, r)$	p-norm ball around vector $\boldsymbol{x}$ of radius r for $p \in \mathbb{N} \cup \{\infty\}$
$\mathbb{I}\{\cdot\}$	indicator function, that is $\mathbb{I}\{E\}$ is 1 if E holds and 0 otherwise
$\mathbb{E}[\,\cdot\,]$	expected value of some random variable
$\mathbb{P}\{\cdot\}$	probability of some random event
$\mathcal{U}(\mathbb{S})$	uniform distribution over the bounded set $\mathbb{S}$
$\mathcal{N}(\boldsymbol{u}, S)$	normal distribution with mean $\boldsymbol{u}$ and covariance matrix S
Ω, Θ, O	asymptotic notations for growth rates
$\mathrm{graph}(C)$	graph of a set–valued function C 9
$\mathcal{N}_\psi(\epsilon, \mathcal{P})$	ϵ-covering number of set $\mathcal{P}$ under distance ψ 14
$\mathcal{H}_\psi(\epsilon, \mathcal{P})$	ϵ-entropy of set $\mathcal{P}$ under distance ψ 14
$\|\cdot\|_{P_\mathbb{X}}, \|\cdot\|_{\mathbb{X}_\infty}$	$L_2(P_\mathbb{X})$ and L_∞ norms for functions in $\{\mathbb{X} \to \mathbb{R}\}$ 14

Chapter 1

Introduction

This thesis considers theoretical and practical aspects of convex regression, where the goal is to recover a hidden convex function from noisy measurements. The discussion includes estimators that map the noisy sample to a convex piecewise linear estimate with a guarantee that the error between the estimate and the hidden convex target decreases as the number of observations in the sample grows. As these methods are computationally too expensive for practical use, their analysis is used for the design of a heuristic training algorithm which is empirically evaluated in various applications.

1.1 Motivation

Convex (or equivalently concave) regression is a machine learning tool with applications in econometrics, engineering, operations research, and possibly more. Its usefulness was recognized early for describing economic relations by imposing concave shape restrictions on utility functions (Afriat, 1967; Varian, 1982, 1984), but its further applications only arrived recently for geometric programming modeling tasks (Magnani and Boyd, 2009; Hoburg and Abbeel, 2014), or stochastic programming planning problems (Cai, 2009; Hannah and Dunson, 2011; Keshavarz, 2012; Nascimento and Powell, 2013; Cai and Judd, 2013; Hannah et al., 2014), although here the importance of convexity was even discussed with the birth of dynamic programming (Bellman, 1957). The computationally intense applications generate a demand for scalable, sample efficient convex regression methods, which motivated this research to advance theoretical understanding and to push practical methods forward.

1.2 Contributions

To understand the theoretical aspects of convex regression, we developed an *excess risk upper bound* (Theorem 3.2) for empirical risk minimization, which extends concentration-based arguments (Pollard, 1990; Dudley, 1999; Györfi et al., 2002; Bartlett et al., 2005) to regression settings with *unbounded* function classes and noise distributions. To fill this gap in the literature, alternative methods appeared recently (Lecué and Mendelson, 2013; Mendelson, 2014; Liang et al., 2015), which, unlike our technique, still have to pose various statistical assumptions and cannot provide near-minimax guarantees for the entire class of sub-Gaussian regression problems (Section 3.3.2). The new result (Theorem 3.2) also extends our previous work (Balázs et al., 2015, Theorem 3.1) by supporting general loss functions, estimates and targets with unbounded magnitude, data-dependent hypothesis classes, and improving the expected value result to a probabilistic guarantee.

Combining our excess risk upper bound (Theorem 3.2) with our results on the *universal function approximation property of max-affine representations* (Lemma 5.2), as shown in Balázs et al. (2015), we provide an analysis for convex nonparametric least squares estimation (Chapter 5) over uniformly Lipschitz convex functions, and construct a *max-affine estimator with near-minimax rate* (Theorem 5.6). This result also serves as the motivation for more practical max-affine training algorithms developed in Chapter 6. Moreover, we apply our excess risk result for linear (convex) regression settings and provide *upper bounds for* widely used practical methods such as *lasso* (Theorem 4.3) and *ridge regression* (Theorem 4.5), extending recent developments (Mendelson, 2014; Hsu et al., 2014) and nearly proving a conjecture on the excess risk rate of slope-bounded linear regression (Shamir, 2015). These new results form the basis of Balázs et al. (2016a).

In order to show that our upper bounds are *tight*, we also provide *excess risk lower bounds* for both linear (Theorem 4.1) and convex nonparametric regression (Theorem 5.1) by constructing examples (Figures 4.1 and 5.1) and extending the probabilistic density estimation lower bound of Yang and Bar-

ron (1999, Proposition 1) to the linear case (Theorem 3.1 and Appendix D) following their discussion in Section 7 of their paper.

For the training of max-affine estimators with an excess risk guarantee, we *demonstrate the overfitting robustness* (Section 6.1.4) of alternating direction methods of multipliers algorithms (Mazumder et al., 2015) compared to cutting plane interior point methods (Lee et al., 2013; Balázs et al., 2015). This observation is used to propose a cross-validation scheme for *learning the Lipschitz factor* (Section 6.1.5). Furthermore, we also extend these training techniques to solve convex formulations of the *partitioned max-affine least squares problem* (Section 6.1.1) which is used to verify the fitting quality of heuristic max-affine training algorithms (Sections 6.2 and 8.3).

To reduce the size of max-affine representations and the computation time for practical applications with large samples, we propose an *adaptive max-affine partitioning algorithm* (Section 6.2.3) by combining the alternating minimization scheme of the least squares partitioning algorithm (Magnani and Boyd, 2009) and the cell splitting technique of the convex adaptive partitioning method (Hannah and Dunson, 2013). While discussing the design of these heuristic approaches (Section 6.2), we compare them empirically to some max-affine estimators that come with excess risk guarantees on a few selected synthetic problems (6.1), and conclude that adaptively tuning the complexity of max-affine representations can significantly improve the generalization error compared to uniform regularization techniques (Figures 6.6, 6.7 and 6.8).

We also provide an extensive *empirical comparison of max-affine estimators* on randomly synthetized convex regression problems (Section 7.1), and on a few applications (Sections 7.2 and 7.3) such as constructing so-called posynomial approximations for geometric programming tasks (Section 7.2.2) for aircraft wing design problems (Hoburg and Abbeel, 2014), and solving convex stochastic programming models by the combination of convex regression and approximate dynamic programming techniques (Hannah and Dunson, 2011; Hannah et al., 2014) for energy storage optimization (Sections 2.3.1 and 7.3.1) and factory operation (Sections 2.3.2 and 7.3.2). The results show that our max-affine partitioning algorithm improves model size adaptation and reduces

the generalization error compared to alternative max-affine training algorithms in many cases.

The adaptive max-affine partitioning algorithm in Section 6.2.3 and the stochastic programming results in Sections 2.3 and 7.3 are also presented in Balázs et al. (2016b).

1.3 Overview

The dissertation starts with motivating the convex regression setting by reviewing some applications in Chapter 2, continues with theoretical analysis in Chapters 3 to 5, considers computational aspects in Chapters 6 and 7, and concludes with future research directions in Chapter 8.

The theoretical analysis covers excess risk lower and upper bounds for empirical risk minimization in Chapter 3, which are applied to linear regression in Chapter 4, including lasso and ridge regression, as well as to convex nonparametric least squares estimation in Chapter 5.

Computation of max-affine estimators with a convex training algorithm and a theoretical guarantee on the excess risk is considered in Section 6.1, where cutting plane and alternating direction method of multipliers techniques are used to address the large-scale quadratic programming tasks that arise. Section 6.2 covers more scalable heuristic methods which train compact max-affine representations with reduced size, including the design of a novel state-of-the-art algorithm.

The developed training methods are evaluated in Chapter 7, comparing max-affine estimators and a few non-convex methods such as adaptive regression splines and support vector regression on both synthetic and real data sets. Section 7.3 also compares max-affine estimators as a building block of approximate dynamic programming algorithms in realistic stochastic programming planning tasks such as solar energy production with storage management, and the operation of a beer brewery.

Chapter 2

Applications

In many practical settings, the regression function is known to comply with some shape restrictions, such as monotonicity, convexity, in addition to satisfying some smoothness conditions (continuity, Lipschitzness, differentiability). In this section, we shortly review a few applications, where the convex (or equivalently concave) shape restriction applies.

2.1 Fitting concave utility functions

In economics, demand, production curves, and utility functions representing rational preferences are usually modeled by concave, nondecreasing functions (Afriat, 1967; Varian, 1982, 1984). In finance, portfolio selection and option pricing models often have concavity restrictions (Merton, 1992), where the nondecreasing property might be dropped in order to represent risk aversion.

As a concrete example, consider the estimation of average weekly wages based on years of education and experience (Ramsey and Schafer, 2002, Chapter 10, Exercise 29). A real data set, containing weekly wages in 1987 of 25,632 males between the age of 18 and 70 who worked in the US full-time along with their years of education and experience, can be accessed as `ex1029` in the `Sleuth2` package of the `R` programming language. The average weekly wages of this data set are depicted in Figure 2.1 (the 845 wage averages were computed over a grid with cell size 1×1 years, and 13 worst outliers were dropped). To produce the figure, the transformation $x \mapsto 1.2^x$ was applied to the education variable as suggested by Hannah and Dunson (2013, Sec-

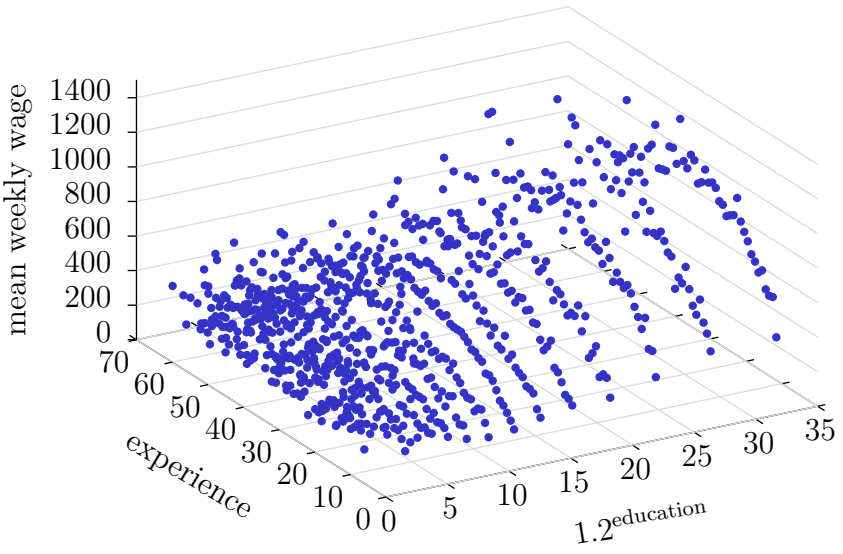

Figure 2.1: Average weekly wage data show a concave quadratic shape based on years of experience and education.

tion 6.2). From the figure, it is apparent that a concave shape restriction is reasonable constraint for this estimation problem (at least, after the said transformation).

2.2 Convex approximation

In engineering applications, one often has to solve an optimization problem of the form

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{s.t.} \quad g_j(\boldsymbol{x}) \leq b_j, \quad j = 1, \ldots, m\,, \tag{2.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is an objective function, $g_j : \mathbb{R}^d \to \mathbb{R}$ are constraint functions and $b_j \in \mathbb{R}$ are some constants. When the functions f and g_j have a "nice" form (for example convex) and can be evaluated in "reasonable time", (2.1) can be solved in poly(m, d) time. When this is not the case, but the functions are "close" to the desired "nice" form, (2.1) can be approximated by replacing each function with its respective convex approximation. A similar scenario occurs when f, $g_1, \ldots, g_m$ are convex or close, but some of the

functions f, $g_1, \ldots, g_m$ are unknown and only noisy observations on them are available.

As an example, consider an aircraft design optimization problem presented by Hoburg and Abbeel (2014, Section VI). As it turns out, this problem "almost" takes the form of a generalized geometric program (GGP), where $\boldsymbol{x} \in \mathbb{R}^d_{>0}$, the functions $f : \mathbb{R}^d_{>0} \to \mathbb{R}_{>0}$, $g_j : \mathbb{R}^d_{>0} \to \mathbb{R}_{>0}$ are generalized posynomials and $b_j = 1$ for all $j = 1, \ldots, m$. (A posynomial is a polynomial of positive variables with positive coefficients. A generalized posynomial is a function of positive variables that can be obtained from posynomials using addition, multiplication, positive (fractional) powers, and maximum. We omit further details here and point the reader to Boyd et al. (2007) for a tutorial on GGP.) An important property of generalized posynomials is that if $\boldsymbol{x} \mapsto f(\boldsymbol{x})$ is a generalized posynomial function then $\boldsymbol{z} \mapsto \ln f(e^{\boldsymbol{z}})$ is convex, where $e^{\boldsymbol{z}}$ denotes a vector obtained by the coordinatewise exponentiation of $\boldsymbol{z}$. This suggests to solve GGP problems using interior point algorithms after transforming (2.1) to a nonlinear convex optimization problem as

$$\min_{\boldsymbol{z}} \ln f(e^{\boldsymbol{z}}) \quad \text{s.t.} \quad \ln g_j(e^{\boldsymbol{z}}) \leq 0, \quad j = 1, \ldots, m\,. \tag{2.2}$$

If the transformed objective $\boldsymbol{z} \mapsto \ln f(e^{\boldsymbol{z}})$ or some constraint function $\boldsymbol{z} \mapsto \ln g_j(e^{\boldsymbol{z}})$ is not convex, but can be approximated by a convex function with "good" accuracy, then convex regression techniques can be used to replace it with its convex approximation.

For the aircraft design example, such a constraint is the drag breakdown inequality (Hoburg and Abbeel, 2014, Equation 38), given as

$$\text{drag coefficient} \geq \text{induced drag} + C_{D_p}(C_L, Re, \tau) + \text{nonwing form drag}\,,$$

where the profile drag coefficient $C_{D_p}(\cdot, \cdot, \cdot)$, depending on the lift coefficient C_L, the Reynolds number Re and the wing thickness ratio τ, is not a generalized posynomial, but "close". This is shown by Figure 2.2 for a fixed τ, where the profile drag coefficient was calculated by the XFOIL simulator (Drela, 1989) as it has no analytical form. Notice that after a logarithmic transformation as needed for (2.2), the profile drag coefficient C_{D_p} for a fixed τ can be well-

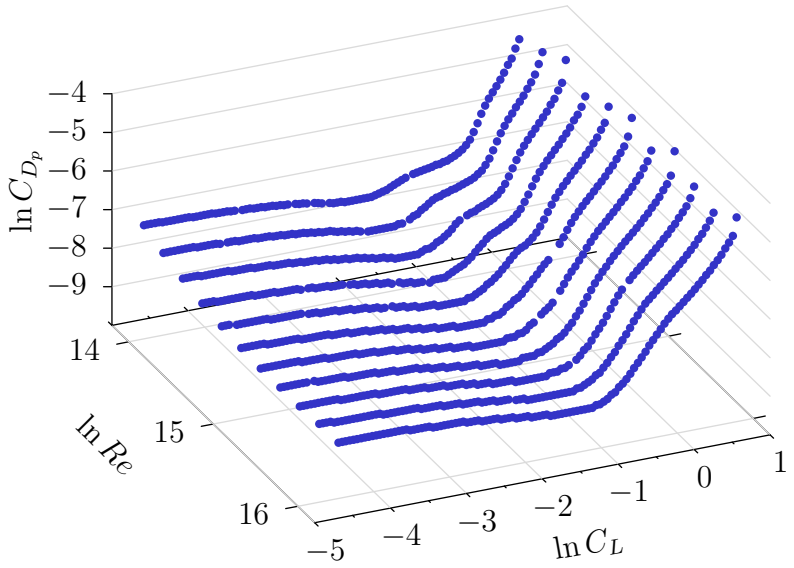

Figure 2.2: The logarithm of the profile drag coefficient (C_{D_p}) shows an almost convex shape in terms of the logarithm of the Reynolds number (Re) and the logarithm of the lift coefficient (C_L) for a fixed thickness ratio ($\tau = 10\%$).

approximated by a convex function to get an approximation of (2.2), which is solvable efficiently by an interior point algorithm.

2.3 Convex stochastic programming

Our next example are T-stage stochastic programming (SP) problems (see for example Ruszczyński and Shapiro, 2003; Shapiro et al., 2009; Birge and Louveaux, 2011), where the goal is to find a decision $\boldsymbol{x}_1^*$ solving the following problem:

$$\begin{aligned} \boldsymbol{x}_1^* \in \operatorname*{argmin}_{\boldsymbol{x}_1 \in \mathbb{X}_1(\boldsymbol{x}_0, \boldsymbol{z}_0)} & J_1(\boldsymbol{x}_1)\,, \\ & J_t(\boldsymbol{x}_t) \doteq \mathbb{E}\Big[c_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) + \min_{\boldsymbol{x}_{t+1} \in \mathbb{X}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t)} J_{t+1}(\boldsymbol{x}_{t+1})\Big]\,, \end{aligned} \tag{2.3}$$

with $t = 1, \ldots, T$, $\boldsymbol{x}_0, \boldsymbol{z}_0$ some fixed initial values, $\mathbb{X}_{T+1}(\boldsymbol{x}_T, \boldsymbol{\mathcal{Z}}_T) \doteq \{\perp\}$, $J_{T+1}(\perp) \doteq 0$, and $\boldsymbol{\mathcal{Z}}_1, \ldots, \boldsymbol{\mathcal{Z}}_T$ a sequence of independent random variables.

We point out that (2.3) includes discrete-time finite-horizon Markov de-

cision process formulations[1] (see for example Puterman, 1994; Sutton, 1998; Bertsekas, 2005; Szepesvári, 2010; Powell, 2011) after the state and action variables are combined into a single decision variable $\boldsymbol{x}_t$, and then reexpressing the environment dynamics and action constraints by the decision constraint functions $\mathbb{X}_t$.

In this text, we consider only a subset of SP problems (2.3) when the cost functions $c_1, \ldots, c_T$ are convex in $\boldsymbol{x}_t$, and $\text{graph}(\mathbb{X}_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t))$ are convex sets for all $t = 1, \ldots, T$ and all $\boldsymbol{\mathcal{Z}}_t$ realizations, where the graph of a set-valued function $C : \mathbb{X} \to 2^{\mathbb{Y}}$ is $\text{graph}(C) \doteq \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X} \times \mathbb{Y} : \boldsymbol{y} \in C(\boldsymbol{x})\}$. In this case, Lemma E.2 (presented in the appendix) implies that the cost-to-go functions $J_t(\cdot)$ are convex for all $t = 1, \ldots, T$, hence we call these SP problems convex.

Numerous specific operations research problems take the form of a convex SP problem including reservoir capacity management (Ruszczyński and Shapiro, 2003, Example 2), the news vendor problem (Shapiro et al., 2009, Section 1.2.1), multi-product assembly problems (Shapiro et al., 2009, Section 1.3.3), portfolio selection (Shapiro et al., 2009, Section 1.4.2), the farmer's problem (Birge and Louveaux, 2011, Section 1.1a) and more. We provide two specific examples in Sections 2.3.1 and 2.3.2.

One approach to deal with such SP problems (2.3) is to use approximate dynamic programming (ADP) methods (see for example Bertsekas, 2005; Powell, 2011; Birge and Louveaux, 2011; Hannah et al., 2014), which construct nested approximations to the cost-to-go functions,

$$\hat{J}_t(\boldsymbol{x}_t) \approx \mathbb{E}\left[c_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) + \min_{\boldsymbol{x}_{t+1} \in \mathbb{X}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t)} \hat{J}_{t+1}(\boldsymbol{x}_{t+1})\right] \approx J_t(\boldsymbol{x}_t)\,,$$

backwards for $t = T, T-1, \ldots, 1$. When the cost-to-go functions $J_t(\cdot)$ are convex (as for the examples below), imposing a convexity constraint for these estimation problems is indeed justified, providing an important application for convex regression. We will evaluate this approach with multiple convex regression procedures later in Section 7.3.

[1] Also, SP problems can arbitrarily approximate infinite-horizon discounted problems by using time-dependent cost functions and a large enough T.

2.3.1 Energy storage optimization

Inspired by a similar example of Jiang and Powell (2015, Section 7.3), we consider an energy storage optimization problem where a renewable energy company makes a decision every hour and plans for two days ($T = 48$). The company owns an energy storage with state s which can be charged with maximum rate r_c, using the company's renewable energy source (E) or the electrical grid that the company can buy electricity from while paying the retail price (p). The goal is to maximize profit by selling electricity to local clients on retail price (p) according to their stochastic demand (D) or selling it back to the electrical grid on wholesale price (w). Electricity can be sold directly from the renewable energy source or from the battery with maximum discharge rate r_d. The energy flow control variables, $f_{\text{es}}, f_{\text{ed}}, f_{\text{eg}}, f_{\text{sd}}, f_{\text{sg}}, f_{\text{gs}}$, are depicted on Figure 2.3.

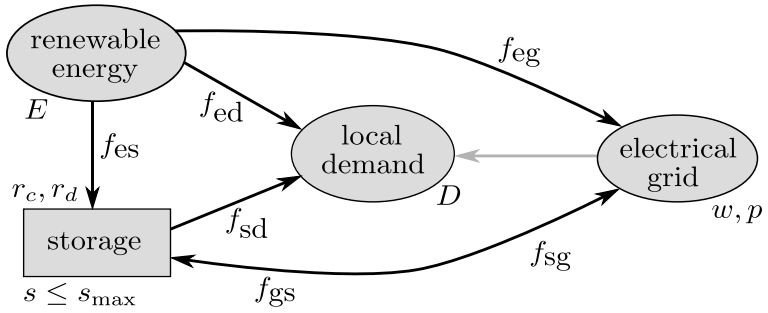

Figure 2.3: Flow diagram of a convex energy storage problem. The storage state is s, its maximum charge and discharge rates are r_c and r_d, the action variables are $f_{\text{es}}, f_{\text{ed}}, f_{\text{eg}}, f_{\text{sd}}, f_{\text{sg}}, f_{\text{gs}}$, the expected retail and wholesale prices are denoted by r, w, and E, D are the stochastic energy production and demand, respectively.

The SP model (2.3) of the energy storage problem can be formulated by setting $\boldsymbol{x}_t \doteq [s_t \ f_{\text{es},t} \ f_{\text{ed},t} \ f_{\text{eg},t} \ f_{\text{sd},t} \ f_{\text{sg},t} \ f_{\text{gs},t}]^\top \in \mathbb{R}^7_{\geq 0}$, and $\boldsymbol{\mathcal{Z}}_t \doteq [E_{t+1} \ D_{t+1}]^\top$. The cost function is defined as

$$c_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) \doteq p_t(f_{\text{gs},t} - f_{\text{ed},t} - f_{\text{sd},t}) - w_t(f_{\text{eg},t} + f_{\text{sg},t}) \,,$$

for all $t = 1, \dots, T$, and the dynamics and control constraints are described by

$$\mathbb{X}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) \doteq \left\{ \begin{bmatrix} s \\ f_{\text{es}} \\ f_{\text{ed}} \\ f_{\text{eg}} \\ f_{\text{sd}} \\ f_{\text{sg}} \\ f_{\text{gs}} \end{bmatrix} \middle| \begin{array}{c} s = s_t + f_{\text{es},t} - f_{\text{sd},t} - f_{\text{sg},t} + f_{\text{gs},t}, \\ f_{\text{es}}, f_{\text{ed}}, f_{\text{eg}}, f_{\text{sd}}, f_{\text{sg}}, f_{\text{gs}} \geq 0, \\ 0 \leq s + f_{\text{es}} - f_{\text{sd}} - f_{\text{sg}} + f_{\text{gs}} \leq s_{\max}, \\ f_{\text{es}} + f_{\text{gs}} \leq r_c,\ f_{\text{sd}} + f_{\text{sg}} \leq r_d, \\ f_{\text{es}} + f_{\text{ed}} + f_{\text{eg}} \leq E_{t+1},\ f_{\text{ed}} + f_{\text{sd}} \leq D_{t+1} \end{array} \right\},$$

for all $t = 0, \dots, T-1$. To initialize the system, define $\boldsymbol{x}_0 \doteq [s_0\ 0 \dots 0]^\top$ and $\boldsymbol{z}_0 \doteq [d_1\ e_1]^\top$, where $s_1 = s_0 \in [0, s_{\max}]$ is the current storage level and $d_1, e_1 \geq 0$ are the currently observed demand and energy production, respectively. This example is further specialized in Section 7.3.1 using a solar energy source and an Economy 7 tariff pricing model.

Notice that the cost function $c_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t)$ is linear in $\boldsymbol{x}_t$ and the dynamics constraint $\boldsymbol{x}_{t+1} \in \mathbb{X}_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t)$ is polyhedral in $(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ for every realization of $\boldsymbol{\mathcal{Z}}_t$, hence the problem is convex if the random variables $\boldsymbol{\mathcal{Z}}_1, \dots, \boldsymbol{\mathcal{Z}}_T$ are independent.[2]

2.3.2 Beer brewery optimization

Inspired by Bisschop (2016, Chapter 17), we consider the multi-product assembly problem of operating a beer brewery which makes a decision in every two weeks and plans for about one year (48 weeks, $T = 24$). The factory has to order ingredients (stratch source, yeast, hops) to produce two types of beers (ale and lager) which have to be fermented (for at least 2 weeks for ale and 6 weeks for lager) before selling. The states and actions of this process are illustrated on Figure 2.4.

The decision variable $\boldsymbol{x}_t$ is a 16 dimensional vector with the following com-

[2]The independence requirement on $\{(E_t, D_t) : t = 1, \dots, T\}$ could be relaxed by introducing extra variables in $\boldsymbol{x}_t$.

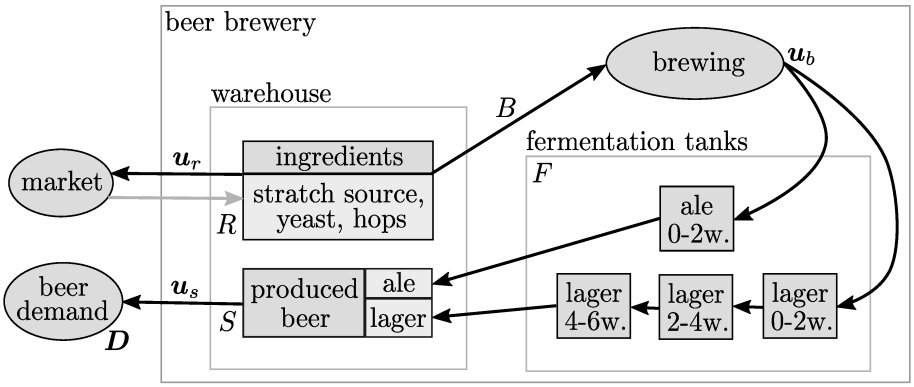

Figure 2.4: Diagram of the beer brewery problem. The state includes three types of ingredients and two types of bottled beer in the warehouse, and the four fermentation tanks. Actions are formed by ingredient orders $\boldsymbol{u}_r$, brewing amounts $\boldsymbol{u}_b$, and beer sales $\boldsymbol{u}_s$.

ponents:

$$\boldsymbol{x}_t \doteq \left[\begin{array}{c} \text{stratch source in storage} \\ \text{yeast in storage} \\ \text{hops in storage} \\ \text{ale beer fermented for less than 2 weeks} \\ \text{produced ale beer} \\ \text{lager beer fermented for less than 2 weeks} \\ \text{lager beer fermented for 2 to 4 weeks} \\ \text{lager beer fermented for 4 to 6 weeks} \\ \text{produced lager beer} \\ \hline \text{stratch source order} \\ \text{yeast order} \\ \text{hops order} \\ \text{ale beer brewing} \\ \text{lager beer brewing} \\ \text{ale beer sales} \\ \text{lager beer sales} \end{array}\right] \in \mathbb{R}^{16}_{\geq 0}\,.$$

Notice that the first 9 coordinates are state variables, while the last 7 coordinates represent actions. The cost functions (which may take negative values) are $c_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) \doteq [\boldsymbol{h}_t^\top \; \boldsymbol{c}_t^\top \; -\boldsymbol{r}_t^\top]\boldsymbol{x}_t$, where $\boldsymbol{h}_t \in \mathbb{R}^9_{\geq 0}$ is the storage cost, $\boldsymbol{c}_t \in \mathbb{R}^5_{\geq 0}$ is the market price of the ingredients with the brewing costs (adjusted by the water price), and $\boldsymbol{r}_t \in \mathbb{R}^2_{\geq 0}$ is the selling price of the beers, at time $t = 1, \ldots, T$.

The constraint on the dynamics of the system is given by

$$\mathbb{X}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) \doteq \left\{ \begin{bmatrix} F\boldsymbol{x}_t + R\boldsymbol{u}_r + B\boldsymbol{u}_b - S\boldsymbol{u}_s \\ \boldsymbol{u}_r \\ \boldsymbol{u}_b \\ \boldsymbol{u}_s \end{bmatrix} \middle| \begin{array}{c} \boldsymbol{u}_r, \boldsymbol{u}_b \geq \mathbf{0},\ \boldsymbol{u}_s \in [\mathbf{0}, \boldsymbol{D}_{t+1}], \\ F\boldsymbol{x}_t + B\boldsymbol{u}_b - S\boldsymbol{u}_s \geq \mathbf{0}, \\ F\boldsymbol{x}_t + R\boldsymbol{u}_r + B\boldsymbol{u}_b \leq \boldsymbol{k}_{t+1} \end{array} \right\},$$

for all $t = 1, \ldots, T-1$, where $\boldsymbol{\mathcal{Z}}_t = \boldsymbol{D}_{t+1} \in \mathbb{R}^2_{\geq 0}$ is the stochastic beer demand, $\boldsymbol{k}_{t+1} \in (\mathbb{R}_{\geq 0} \cup \{\infty\})^9$ is the capacity bound, and the fermentation matrix $F \in \{0,1\}^{9\times 16}$, the brewing matrix $B \in \mathbb{R}^{9\times 2}$, the storage loading matrix $R \in \{0,1\}^{9\times 3}$ and the selling matrix $S \in \{0,1\}^{9\times 2}$ are defined as

$$F \doteq \left[\begin{array}{c|c|c} I_3 & \mathbf{0}_{3\times 13} & \\ \hline & \begin{matrix} 0 & 0 \\ 1 & 1 \end{matrix} \quad \mathbf{0}_{2\times 4} & \\ \mathbf{0}_{6\times 3} & \begin{matrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{matrix} & \mathbf{0}_{6\times 7} \end{array}\right], \quad B \doteq \left[\begin{array}{cc} -\boldsymbol{b}_a & -\boldsymbol{b}_l \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ \hline \multicolumn{2}{c}{\mathbf{0}_{3\times 2}} \end{array}\right], \quad S \doteq \left[\begin{array}{c} \mathbf{0}_{3\times 2} \\ \hline 0\ 0 \\ 1\ 0 \\ 0\ 0 \\ 0\ 0 \\ 0\ 0 \\ 0\ 1 \end{array}\right],$$

$$R \doteq \begin{bmatrix} I_3 \\ \mathbf{0}_{6\times 3} \end{bmatrix},$$

where $\boldsymbol{b}_a, \boldsymbol{b}_l \in \mathbb{R}^3_{\geq 0}$ are the required ingredients for brewing ales and lagers, respectively. To initialize the system, define $\boldsymbol{x}_0 \doteq [\boldsymbol{s}_0\ 0 \ldots 0]^\top$ and $\boldsymbol{z}_0 \doteq \boldsymbol{d}_1$, where $\boldsymbol{s}_1 = \boldsymbol{s}_0 \geq \mathbf{0}$ is the current factory state and $\boldsymbol{d}_1 \geq \mathbf{0}$ is the currently observed beer demand. Further details for the parameter settings of this example is given in Section 7.3.2.

Similar to the previous example above, the cost function is linear and the dynamics constraint is polyhedral, hence the problem is convex if the demand random variables $\boldsymbol{\mathcal{Z}}_1, \ldots, \boldsymbol{\mathcal{Z}}_T$ are independent.

Chapter 3

Regression analysis

In this chapter we discuss the sample complexity of empirical risk minimization (ERM) estimators and analyze the worst-case excess risk convergence rate for general regression settings (not just the convex case). We will apply these results to linear least squares estimation (Chapter 4), and to convex nonparametric regression (Chapter 5).

For the discussion, we need covering numbers and entropies, hence we provide the definitions here. Let $(\mathcal{P}, \psi)$ be a nonempty metric space and $\epsilon \geq 0$. The set $\{p_1, \ldots, p_k\} \subseteq \mathcal{P}$ is called an (internal) *ϵ-cover* of $\mathcal{P}$ under ψ if the ψ-balls of centers $\{p_1, \ldots, p_k\}$ and radius ϵ cover $\mathcal{P}$: for any $q \in \mathcal{P}$, $\min_{i=1,\ldots,k} \psi(q, p_i) \leq \epsilon$. The *$\epsilon$-covering number* of $\mathcal{P}$ under ψ, $\mathcal{N}_\psi(\epsilon, \mathcal{P})$, is the cardinality of the ϵ-cover with the fewest elements:

$$\mathcal{N}_\psi(\epsilon, \mathcal{P}) \doteq \inf \left\{ k \in \mathbb{N} \,\middle|\, \exists p_1, \ldots, p_k \in \mathcal{P} : \sup_{q \in \mathcal{P}} \min_{i=1,\ldots,k} \psi(q, p_i) \leq \epsilon \right\}$$

with $\inf \emptyset = \infty$. Further, the *ϵ-entropy* of $\mathcal{P}$ under ψ is defined as the logarithm of the covering number, $\mathcal{H}_\psi(\epsilon, \mathcal{P}) \doteq \ln \mathcal{N}_\psi(\epsilon, \mathcal{P})$. For convenience, we define these quantities for the empty set to be zero, that is $\mathcal{N}_\psi(\epsilon, \emptyset) \doteq \mathcal{H}_\psi(\epsilon, \emptyset) \doteq 0$.

When the function class $\mathcal{F} \subseteq \{\mathbb{X} \to \mathbb{R}\}$, over some set $\mathbb{X}$, is square integrable with respect to some distribution $P_\mathbb{X}$, that is $\sup_{f \in \mathcal{F}} \|f\|_{P_\mathbb{X}} < \infty$ holds with $\|f\|_{P_\mathbb{X}}^2 \doteq \int_\mathbb{X} f^2(\boldsymbol{x})\, dP_\mathbb{X}(\boldsymbol{x})$, we use a shorthand notation to denote the ϵ-entropy of $\mathcal{F}$ under $\|\cdot\|_{P_\mathbb{X}}$ as $\mathcal{H}_{P_\mathbb{X}}(\epsilon, \mathcal{F}) \doteq \mathcal{H}_{\|\cdot\|_{P_\mathbb{X}}}(\epsilon, \mathcal{F})$.

3.1 The regression problem

We start with the formal definition of a regression problem, which is given by a probability distribution μ over some set $\mathbb{X} \times \mathbb{R}$ with some *domain* $\mathbb{X}$ being a separable Hilbert space,[1] a *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, and a *reference class* $\mathcal{F}_* \subseteq \{\mathbb{X} \to \mathbb{R}\}$.

Then the task of a regression estimator is to produce a function $f : \mathbb{X} \to \mathbb{R}$ based on a *training sample* $\mathcal{D}_n \doteq \{(\boldsymbol{\mathcal{X}}_1, \mathcal{Y}_1), \ldots, (\boldsymbol{\mathcal{X}}_n, \mathcal{Y}_n)\}$ of $n \in \mathbb{N}$ pairs $(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) \in \mathbb{X} \times \mathbb{R}$ independently sampled from μ (in short $\mathcal{D}_n \sim \mu^n$), such that the prediction error, $\ell(\mathcal{Y}, f(\boldsymbol{\mathcal{X}}))$, is "small" on a new instance $(\boldsymbol{\mathcal{X}}, \mathcal{Y}) \sim \mu$ with respect to ℓ.

The *risk* of function $f : \mathbb{X} \to \mathbb{R}$ is defined as $R_\mu(f) \doteq \mathbb{E}[\ell(\mathcal{Y}, f(\boldsymbol{\mathcal{X}}))]$ and the cost of using a fixed function f is measured by the *excess risk*, $L_\mu(f, f_*) \doteq R_\mu(f) - R_\mu(f_*)$, where $f_* \doteq f_{\mu,\mathcal{F}_*} \in \operatorname{argmin}_{f \in \mathcal{F}_*} R_\mu(f)$ is a *reference function*.[2] When f_* also satisfies $f_* \in \operatorname{argmin}_{f \in \{\mathbb{X} \to \mathbb{R}\}} R_\mu(f)$, it is also called the *regression function*. Clearly, the two concepts coincide when $\mathcal{F}_* = \{\mathbb{X} \to \mathbb{R}\}$. Also notice that not every $f \in \mathcal{F}_*$ is a reference function.

An *estimator* h_n is defined as a sequence of mappings $h \doteq (h_n)_{n \in \mathbb{N}}$, where $h_n : (\mathbb{X} \times \mathbb{R})^n \to \{\mathbb{X} \to \mathbb{R}\}$ maps the data $\mathcal{D}_n$ into an estimate $f_n \doteq h_n(\mathcal{D}_n)$. These estimates lie within some *hypothesis class* $\mathcal{F}_n \subseteq \{\mathbb{X} \to \mathbb{R}\}$, that is $f_n \in \mathcal{F}_n$, where $\mathcal{F}_n$ might depend on the random sample $\mathcal{D}_n$.

Finally, for a regression problem specified by $(\ell, \mu, \mathcal{F}_*)$, the goal of an estimator h_n is to minimize the excess risk,

$$L_\mu\big(h_n(\mathcal{D}_n), f_*\big)\,,$$

with high-probability or in expectation, where the random event is induced by the random sample $\mathcal{D}_n$ and the possible randomness of the estimator h_n.

[1] All sets and functions considered are assumed to be measurable as necessary. To simplify the presentation, we omit these conditions by noting here that all the measurability issues can be overcome using standard techniques as we work with separable Hilbert spaces (see for example, Dudley, 1999, Chapter 5).

[2] A straightforward limiting argument can be used if the minimums are not attained.

3.2 Lower bounds on minimax rates

To measure the performance of an estimator, we need a baseline, for which we use minimax rates. Formally, for a family of regression problems, represented by a set of probability distributions $\mathbb{M} = \{\mu \,|\, \mu \text{ is a distribution on } \mathbb{X} \times \mathbb{R}\}$, a loss function ℓ, and a reference class $\mathcal{F}_*$, we define the *minimax rate* as

$$\mathcal{R}_n(\mathbb{M}, \ell, \mathcal{F}_*) \doteq \sup \left\{ r_n \,\middle|\, \inf_{h_n} \sup_{\mu \in \mathbb{M}} \mathbb{P}\left\{ L_\mu\big(h_n(\mathcal{D}_n), f_{\mu,\mathcal{F}_*}\big) \geq r_n \right\} \geq 1/2 \right\},$$

where the infimum over h_n scans through all estimators mapping to $\{\mathbb{X} \to \mathbb{R}\}$. We also say that an estimator has a *near-minimax rate* if it achieves the minimax rate up to a polylogarithmic factor in the sample size n.

We only consider deriving lower bounds for the minimax rate with the squared loss $\ell_{\text{sq}}(y, \hat{y}) \doteq |y - \hat{y}|^2$. For these settings, we use information theoretic developments based on Fano's lemma and the Kullback-Leibler (KL) divergence (Yang and Barron, 1999). This is well-suited for squared loss regression settings with Gaussian noise, because the KL divergence of two Gaussian random variables is equal to the squared distance between their means. For a more thorough treatment of minimax lower bounds, we point to the work of Guntuboyina (2011) or Chapter 2 of Tsybakov (2009).

The following result builds on a slightly modified version of the density estimation lower bound of Yang and Barron (1999, Theorem 1), by making it capable to handle linear settings as well. For this, we also use local entropies (Yang and Barron, 1999, Section 7). For a $(\mathcal{F}, \psi)$ metric space, the ϵ_*-*local* ϵ-*entropy* of $\mathcal{F}$ under ψ is defined by

$$\mathcal{H}^*_\psi(\epsilon, \epsilon_*, \mathcal{F}) \doteq \sup_{f \in \mathcal{F}} \mathcal{H}_\psi\big(\epsilon, \{g \in \mathcal{F} : \psi(f, g) \leq \epsilon_*\}\big),$$

and $\mathcal{H}^*_{P_\mathbb{X}}(\epsilon_*, \epsilon, \mathcal{F}) \doteq \mathcal{H}^*_{\|\cdot\|_{P_\mathbb{X}}}(\epsilon_*, \epsilon, \mathcal{F})$ for short. An important property of local entropies is that they can be often upper bounded independently of the sample size n for linear function classes as long as ϵ_* and ϵ are kept on the same scale, for example $\epsilon_* \approx \epsilon \approx 1/\sqrt{n}$ is used to prove Theorem 3.1a below.

Here, we present only the final result on the lower bound (Theorem 3.1). The full proof is given in Appendix D to isolate some notation, which we do not use elsewhere.

Theorem 3.1. *Let $P_{\mathbb{X}}$ be a distribution on $\mathbb{X}$, and for some $\sigma > 0$ define*

$$\mathbb{M}_{gs}^{\sigma}(\mathcal{F}_*, P_{\mathbb{X}}) \doteq \Big\{ \mu \,\Big|\, (\boldsymbol{\mathcal{X}}, \mathcal{Y}) \sim \mu,\ \mathcal{Y} = f_*(\boldsymbol{\mathcal{X}}) + \mathcal{Z},\ f_* \in \mathcal{F}_*,\ \boldsymbol{\mathcal{X}} \sim P_{\mathbb{X}},\ \mathcal{Z} \sim \mathcal{N}\big(0, \sigma^2\big) \Big\}.$$

Then for all distribution classes $\mathbb{M} \supseteq \mathbb{M}_{gs}^{\sigma}(\mathcal{F}_*, P_{\mathbb{X}})$, function sets $\mathcal{F} \supseteq \mathcal{F}_*$, and $v > 0$, $c_2 \geq c_1 > 0$, the following claims hold:

(a) If for some $c_0 > 0$, $v \ln(c_1/\epsilon) \leq \mathcal{H}_{P_{\mathbb{X}}}(\epsilon, \mathcal{F}_*)$, $\mathcal{H}^*_{P_{\mathbb{X}}}(\epsilon_*, \epsilon, \mathcal{F}_*) \leq v \ln(c_2 \epsilon_*/\epsilon)$ for all $\epsilon_* \in (0, c_0]$ and $\epsilon \in (0, c_2\epsilon_*]$, then $\mathcal{R}_n(\mathbb{M}, \ell_{sq}, \mathcal{F}) \geq 2\sigma^2 v/(c_2^2\, n)$ holds for all $n \geq (4\sigma^4\, v/c_2^2) \max\big\{32 \cdot 2^{4/v}/c_1^2, 1/c_0^2\big\}$.

(b) If for some $\epsilon_0 > 0$, $c_1 \epsilon^{-v} \leq \mathcal{H}_{P_{\mathbb{X}}}(\epsilon, \mathcal{F}_*) \leq c_2 \epsilon^{-v}$ for all $\epsilon \in (0, \epsilon_0]$, then for all $n \in \mathbb{N}$, we have $\mathcal{R}_n(\mathbb{M}, \ell_{sq}, \mathcal{F}) \geq c_*\, n^{-2/(v+2)}$ with

$$c_* \doteq \big(\sigma^2 c_1^2/18\big)^{1/v} \max\Big\{1, \epsilon_0^2, 2\sigma^2 \big(2\sigma^2 c_2^2\big)^{1/v}/\epsilon_0^2\Big\}^{-1}.$$

Proof. See Appendix D (page 125). ■

We will apply Theorem 3.1a for linear settings (Section 4.1), and Theorem 3.1b to derive lower bounds for convex nonparametric regression problems (Section 5.2).

3.3 Upper bounds for ERM estimators

Now we state our general excess risk upper bound for empirical risk minimization (ERM) estimators. An estimate is called an *α-approximate β-penalized ERM estimate* with respect to the function class $\mathcal{F}_n$, in short (α, β)-ERM$(\mathcal{F}_n)$, when its estimate $f_n \in \mathcal{F}_n$ satisfies

$$R_n(f_n) + \beta(f_n) \leq \inf_{f \in \mathcal{F}_n} R_n(f) + \beta(f) + \alpha\,, \tag{3.1}$$

where $R_n(f) \doteq \frac{1}{n}\sum_{i=1}^n \ell(\mathcal{Y}_i, f(\boldsymbol{\mathcal{X}}_i))$ is the *empirical risk* of function $f : \mathbb{X} \to \mathbb{R}$, $\beta : \mathcal{F}_n \to \mathbb{R}_{\geq 0}$ is a penalty function and $\alpha \geq 0$ is an error term. All α, β, and $\mathcal{F}_n$ might depend on the sample $\mathcal{D}_n$. When the penalty function is zero (that is $\beta \equiv 0$), we simply call the estimator α-ERM$(\mathcal{F}_n)$.

Our result will require a few conditions to hold on the random variables $\mathcal{Z}(f,g) \doteq \ell(\mathcal{Y}, f(\boldsymbol{\mathcal{X}})) - \ell(\mathcal{Y}, g(\boldsymbol{\mathcal{X}}))$, $f, g : \mathbb{X} \to \mathbb{R}$, which are related to the excess risk through $L_\mu(f, f_*) = \mathbb{E}[\mathcal{Z}(f, f_*)]$. Similarly, we use the empirical excess risk defined as $L_n(f, f_*) \doteq R_n(f) - R_n(f_*) = \frac{1}{n}\sum_{i=1}^n \mathcal{Z}_i(f, f_*)$, where $\mathcal{Z}_i(f, f_*) \doteq \ell(\mathcal{Y}_i, f(\boldsymbol{\mathcal{X}}_i)) - \ell(\mathcal{Y}_i, f_*(\boldsymbol{\mathcal{X}}_i))$.

We also need sub-Gaussian random variables (see for example, Buldygin and Kozachenko, 2000, Section 1.1). A real-valued random variable $\mathcal{W}$ is called *B-sub-Gaussian* (or sub-Gaussian with B) when $\sup_{s\in\mathbb{R}} \mathbb{E}\big[e^{s(\mathcal{W}-\mathbb{E}\mathcal{W})-s^2B^2/2}\big] \le 1$ holds with some $B \ge 0$. To simplify the calculations and extend the sub-Gaussian property to random vectors $\boldsymbol{\mathcal{W}} \in \mathbb{R}^d$, we use the Ψ_2 Orlicz norm defined as $\|\boldsymbol{\mathcal{W}}\|_{\Psi_2} \doteq \inf\{B > 0 : \Psi_2(\boldsymbol{\mathcal{W}}/B) \le 1\}$, where $\Psi_2(\boldsymbol{x}) \doteq e^{\|\boldsymbol{x}\|^2} - 1$ and $\inf\emptyset \doteq \infty$. The norm $\|\cdot\|_{\Psi_2}$ provides an alternative characterization of sub-Gaussian random variables, because $\mathcal{W} \in \mathbb{R}$ is sub-Gaussian if and only if $\|\mathcal{W}\|_{\Psi_2} < \infty$.[3] Furthermore, if $\|\boldsymbol{\mathcal{W}}\|_{\Psi_2} < \infty$, then the coordinates of $\boldsymbol{\mathcal{W}} \in \mathbb{R}^d$ are sub-Gaussian random variables. Sub-Gaussian random vectors and the properties of the norm $\|\cdot\|_{\Psi_2}$ are reviewed in Appendix A.

To state our main result, we use the scaled cumulant-generating function of a random variable $\mathcal{W}$, which is defined as $\mathbb{C}_t[\mathcal{W}] \doteq (1/t)\ln\mathbb{E}\big[\exp(t\mathcal{W})\big]$ for any $t > 0$. Its properties are reviewed in Appendix B.

Finally, we are ready to state the promised result:

Theorem 3.2. *Consider a regression problem $(\ell, \mu, \mathcal{F}_*)$, an arbitrary reference function f_* in $\mathcal{F}_*$, and an i.i.d. training sample $\mathcal{D}_n \sim \mu^n$. Let $\mathcal{F}_n \subseteq \{\mathbb{X} \to \mathbb{R}\}$ be a hypothesis class which might depend on the data $\mathcal{D}_n$, and let f_n be an (α, β)-ERM($\mathcal{F}_n$) estimate (3.1). Furthermore, let $\hat{\mathcal{F}}_n, \mathcal{F} \subseteq \{\mathbb{X} \to \mathbb{R}\}$ be two function classes, where $\hat{\mathcal{F}}_n$ might depend on $\mathcal{D}_n$, but $\mathcal{F}$ might depend on the sample only through its size n. Suppose that $\hat{\mathcal{F}}_n$ and $\mathcal{F}$ enclose $\mathcal{F}_n$ as $\mathbb{P}\big\{\hat{\mathcal{F}}_n \subseteq \mathcal{F}_n \subseteq \mathcal{F}\big\} \ge 1 - \gamma/2$ with some $\gamma \in (0,1)$, and the approximation error of $\hat{\mathcal{F}}_n$ to f_* is bounded as $\mathbb{P}\big\{\inf_{f\in\hat{\mathcal{F}}_n} L_n(f, f_*) + \beta(f) + \alpha \le B_*\big\} \ge 1 - \gamma/4$ with some $B_* \in \mathbb{R}$. Finally, suppose that the following conditions are satisfied for some metric $\psi : \mathcal{F}\times\mathcal{F} \to \mathbb{R}_{\ge 0}$ and $\mathcal{F}(r_0) \doteq \{f \in \mathcal{F} : L_\mu(f, f_*) > r_0/n\}$*

[3] The sub-Gaussian parameter and $\|\cdot\|_{\Psi_2}$ are not equal. For example, a centered Gaussian random variable $\mathcal{W} \sim \mathcal{N}(0, \sigma^2)$ is σ-sub-Gaussian, but only satisfies $\mathbb{E}\big[e^{3\mathcal{W}^2/(8\sigma^2)}\big] = 2$.

with some $r_0 \geq 0$*:*

(C1) there exists $G : \mathbb{X} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ *such that* $\mathcal{Z}(f,g) \leq G(\boldsymbol{\mathcal{X}}, \mathcal{Y})\, \psi(f,g)$ *a.s. for all* $f, g \in \mathcal{F}(r_0)$*,*

(C2) there exists $S \in (0, \infty]$ *such that the random variable* $\mathcal{Z}(f,g)$ *is sub-Gaussian with* $S\psi(f,g)$ *for all* $f, g \in \mathcal{F}(r_0)$*,*

(C3) there exists $r \in (0,1]$ *and* $\theta > 0$ *such that*

$$\sup_{f \in \mathcal{F}(r_0)} \mathbb{E}\Big[\exp\Big((r/\theta)\mathbb{E}[\mathcal{Z}(f,f_*)] - (1/\theta)\mathcal{Z}(f,f_*) \Big) \Big] \leq 1\,.$$

Then for all $\epsilon \geq \delta > 0$*, with probability at least* $1-\gamma$*,*

$$L_\mu(f_n, f_*) \leq \frac{1}{r}\Bigg(\frac{\theta\, \mathcal{H}_\psi(\epsilon, \mathcal{F})}{n} + 16S \int_\delta^\epsilon \max\left\{ \sqrt{\frac{\mathcal{H}_\psi(z,\mathcal{F})}{n}},\, \frac{4S}{\theta\delta}\, z^2 \right\} dz \\ + 16\delta\, \mathbb{C}_{\frac{1}{\theta}}\big[G(\boldsymbol{\mathcal{X}}, \mathcal{Y}) \big] + B_* \Bigg) + \frac{r_0 + 4\theta \ln(4/\gamma)}{n}\,.$$

Furthermore, the result holds without Condition (C2) *(that is when* $S = \infty$*) with* $\epsilon = \delta$ *defining* $\infty \cdot 0 = 0$*.*

The proof of Theorem 3.2 is presented in Section 3.3.3.

We point out that if $f_* \in \hat{\mathcal{F}}_n$, the approximation term is upper bounded by zero, that is $\inf_{f \in \hat{\mathcal{F}}_n} L_n(f, f_*) \leq 0$ a.s., and then Theorem 3.2 is an exact oracle inequality. Otherwise, when $f_* \notin \hat{\mathcal{F}}_n$, Theorem 3.2 becomes an inexact oracle inequality. We use exact oracle inequalities to prove ERM upper bounds for linear regression problems (Chapter 4), and inexact ones for convex nonparametric cases (Chapter 5).

Notice that Conditions (C1) and (C2) are immediately satisfied when the loss function ℓ is Lipschitz and $\boldsymbol{\mathcal{X}}$ is bounded. Furthermore, Condition (C1) is a generalization of the usual Lipschitz condition of the loss ℓ (see for example the 2nd condition in Section 5.2 of Bartlett et al. 2005), which allows Theorem 3.2 to deliver excess risk upper bounds for the (non-Lipschitz) squared loss over an (unbounded) sub-Gaussian range (support of $\mathcal{Y}$ and the range of the functions in the hypothesis class). For Condition (C3), we provide a detailed analysis below in Section 3.3.1.

The entropy integral in Theorem 3.2 is presented only for completeness, but later we use only $\delta = \epsilon$ and ignore Condition (C2). This integral can be used for nonparametric settings to prove near-minimax bounds for ERM estimators over infinite dimensional function spaces as long as the integral converges (as $\delta \to 0$), which happens only up to a few dimensions d, where $\mathbb{X} \subseteq \mathbb{R}^d$. We discuss this in more detail for convex nonparametric least squares regression over convex, uniformly Lipschitz functions in Section 5.1, where the entropy integral converges for $d \in \{1, 2, 3\}$, and diverges for $d = 4$ with a logarithmic rate, which is still enough to prove a near-minimax rate. For $d > 4$, the divergence is too fast and ERM is not able to deliver the near-minimax rate.

Finally, we mention that an upper bound on the expected excess risk $\mathbb{E}[L_\mu(f_n, f_*)]$ can be obtained by integration. Transforming the probabilistic bound $\mathbb{P}\big\{L_\mu(f_n, f_*) \le b + \frac{4\theta \ln(4/\gamma)}{n}\big\} \ge 1 - \gamma$ with some $b \ge 0$ and $\gamma = 4e^{-n\,t/(4\theta)}$, we get

$$\begin{aligned} \mathbb{E}[L_\mu(f_n, f_*)] - b &\le \mathbb{E}\big[\max\{0, L_\mu(f_n, f_*) - b\}\big] \\ &= \int_0^\infty \mathbb{P}\{L_\mu(f_n, f_*) > b + t\}\, dt \\ &\le \int_0^\infty 4e^{-n\,t/(4\theta)}\, dt = \frac{16\,\theta}{n}\,. \end{aligned} \tag{3.2}$$

As the $O(\theta/n)$ rate cannot be exceeded by Theorem 3.2, our probabilistic results also imply the same rate for the expected excess risk.

3.3.1 Analysis of the moment condition

Here we provide an analysis for Condition (C3) of Theorem 3.2. For this, consider a regression problem $(\mu, \ell, \mathcal{F}_*)$ with a reference function f_* in $\mathcal{F}_*$, a function class $\mathcal{F} \subseteq \{\mathbb{X} \to \mathbb{R}\}$, and some $r_0 \ge 0$, as needed for Theorem 3.2.

Then, we say that $(\mu, \ell, \mathcal{F}(r_0), f_*)$ satisfies the *Bernstein condition* (Bartlett and Mendelson, 2006, Definition 2.6) if there exists some $C > 0$ such that for all $f \in \mathcal{F}(r_0)$, we have

$$\mathbb{E}\big[\mathcal{W}_f^2\big] \le C\,\mathbb{E}\big[\mathcal{Z}(f, f_*)\big]\,, \quad \mathcal{W}_f \doteq f(\boldsymbol{\mathcal{X}}) - f_*(\boldsymbol{\mathcal{X}})\,. \tag{3.3}$$

When, next to the Bernstein condition (3.3) the loss function ℓ has a "sub-Gaussian Lipschitz" property, Condition (C3) is satisfied as detailed by Lemma 3.3.

For this, we also use the kurtosis about the origin of a random variable $\mathcal{W}$ which is defined as $\mathbb{K}_0[\mathcal{W}] \doteq \mathbb{E}[\mathcal{W}^4]/\mathbb{E}[\mathcal{W}^2]^2$.

Lemma 3.3. *Let $r_0 > 0$ and suppose that the Bernstein condition (3.3) holds for $(\mu, \ell, \mathcal{F}(r_0), f_*)$ with some $C > 0$. Further, suppose $\sup_{f\in\mathcal{F}} \|\mathcal{W}_f\|_{\Psi_2} \le B$ holds with some $B > 0$, and $|\mathcal{Z}(f, f_*)| \le L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y})|\mathcal{W}_f|$ a.s. for all $f \in \mathcal{F}(r_0)$ with some $L : \mathcal{F} \times \mathbb{X} \times \mathbb{R} \to \mathbb{R}_{\ge 0}$ such that $\sup_{f\in\mathcal{F}} \|L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y})\|_{\Psi_2} \le R < \infty$. Then $(\mu, \ell, \mathcal{F}(r_0), f_*)$ also satisfies Condition (C3) for any $r \in (0, 1)$ with $\theta \ge 2tQ_n \max\{\frac{3R^2C}{(t-1)(1-r)}, 4BR\}$, $t > 1$, and*

$$Q_n \doteq \ln\left(3\min\left\{\sup_{f\in\mathcal{F}(r_0)} \mathbb{K}_0[\mathcal{W}_f]^{\frac{1}{4}}, \frac{nBR}{r_0}\right\}\right).$$

Proof. Let $\mathcal{Z}_f \doteq \mathcal{Z}(f, f_*)$ and fix any $f \in \mathcal{F}(r_0)$. Then by the definition of $\mathcal{F}(r_0)$, the Cauchy-Schwartz inequality, and Lemma A.2b for $k = 1$, we get

$$\frac{r_0}{n} \le \mathbb{E}[\mathcal{Z}_f] \le \mathbb{E}\big[|L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y})\mathcal{W}_f|\big] \le \mathbb{E}\big[L^2(f, \boldsymbol{\mathcal{X}}, \mathcal{Y})\big]^{\frac{1}{2}} \mathbb{E}\big[\mathcal{W}_f^2\big]^{\frac{1}{2}} \le R\,\mathbb{E}\big[\mathcal{W}_f^2\big]^{\frac{1}{2}},$$

which implies $\mathbb{E}\big[\mathcal{W}_f^2\big] \ge \big(r_0/(nR)\big)^2$. Combining this with Lemma A.2b for $k = 2$, we obtain $4\sqrt{\mathbb{K}_0[\mathcal{W}_f]} \le 4\,\mathbb{E}\big[\mathcal{W}_f^4\big]^{1/2}\big(r_0/(nR)\big)^{-2} \le (3nBR/r_0)^2$. Then, by using Lemma A.5 with $\ln(4\sqrt{\mathbb{K}_0[\mathcal{W}_f]}) \le 2Q_n$, we get for all $2 \le k \in \mathbb{N}$ that

$$\mathbb{E}\big[|\mathcal{Z}_f|^k\big] \le \mathbb{E}\big[|L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y})\mathcal{W}_f|^k\big] \le (k!/2)\big(12Q_n\mathbb{E}\big[\mathcal{W}_f^2\big]R^2\big)\big(8Q_nBR\big)^{k-2}.$$

Hence, the conditions of Bernstein's lemma (Lemma A.4) hold for $\mathcal{Z}_f$ and $\theta \ge 8tQ_nBR$ with $t > 1$, so by $(1 - 8Q_nBR/\theta)^{-1} \le t/(t-1)$, we obtain

$$\begin{aligned}
\mathbb{E}\big[e^{(r/\theta)\mathbb{E}[\mathcal{Z}_f]-(1/\theta)\mathcal{Z}_f}\big] &= e^{(r-1)\mathbb{E}[\mathcal{Z}_f]/\theta}\,\mathbb{E}\big[e^{(\mathbb{E}[\mathcal{Z}_f]-\mathcal{Z}_f)/\theta}\big] \\
&\le \exp\left(\frac{r-1}{\theta}\mathbb{E}[\mathcal{Z}_f] + \frac{6tQ_nR^2}{(t-1)\theta^2}\mathbb{E}[\mathcal{W}_f^2]\right) \\
&\le \exp\left(\frac{\mathbb{E}[\mathcal{Z}_f]}{\theta}\left(r - 1 + \frac{6tQ_nR^2C}{(t-1)\theta}\right)\right) \le 1,
\end{aligned}$$

where in the last step we applied the Bernstein condition (3.3) and the bound $\theta \ge \frac{6tQ_nR^2C}{(t-1)(1-r)}$. ∎

In the following paragraphs, we present a few important examples which satisfy the requirements of Lemma 3.3.

Lipschitz losses

First, observe that if $\sup_{f\in\mathcal{F}} \|\mathcal{W}_f\|_{\Psi_2} \leq B$, then the Bernstein condition (3.3) always holds for any $\mathcal{F}(r_0)$ with $r_0 > 0$ and $C = nB^2/r_0$, as by Lemma A.2b with $k = 1$, and $1 < (n/r_0)L_\mu(f, f_*)$ for any $f \in \mathcal{F}(r_0)$, we have

$$\mathbb{E}\big[\mathcal{W}_f^2\big] \leq B^2 < \frac{nB^2}{r_0} L_\mu(f, f_*) = \frac{nB^2}{r_0}\mathbb{E}[\mathcal{Z}(f, f_*)]\,.$$

Now consider a loss function ℓ, which is R-Lipschitz in its second argument satisfying $|\ell(y, \hat{y}_1) - \ell(y, \hat{y}_2)| \leq R\,|\hat{y}_1 - \hat{y}_2|$ for all $y, \hat{y}_1, \hat{y}_2 \in \mathbb{R}$. Then, we clearly have $\mathcal{Z}(f, f_*) \leq R\,|\mathcal{W}_f|$ for any $f \in \mathcal{F}$, so the requirements of Lemma 3.3 hold and we obtain the following result:

Lemma 3.4. *Let $r_0 > 0$, $\sup_{f\in\mathcal{F}} \|\mathcal{W}_f\|_{\Psi_2} \leq B$, and the loss function ℓ be R-Lipschitz in its second argument. Then $(\mu, \ell, \mathcal{F}(r_0), f_*)$ satisfies Condition (C3) for any $r \in (0, 1)$, $t > 1$, and $\theta \geq 2tQ_n \max\big\{\frac{3n(BR)^2}{r_0(t-1)(1-r)}, 4BR\big\}$.*

Proof. Based on the discussion above, apply Lemma 3.3 with $C = nB^2/r_0$ and $L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y}) = R$. ∎

For a successful combination of Theorem 3.2 and Lemma 3.4, one should choose r_0 to balance $\frac{\theta}{n}\mathcal{H}_\psi(\epsilon, \mathcal{F}) = \Theta\big(\frac{Q_n(BR)^2}{r_0}\mathcal{H}_\psi(\epsilon, \mathcal{F})\big)$ with $\frac{r_0}{n}$, which is satisfied if $r_0 = \Theta\big(BR\sqrt{n\,Q_n\mathcal{H}_\psi(\epsilon, \mathcal{F})}\big)$. This way the bound of Theorem 3.2 scales with $BR\sqrt{Q_n\mathcal{H}_\psi(\epsilon, \mathcal{F})/n}$, which cannot be improved in general.

For example, consider the estimation problem of a constant regression function f_* in $\mathcal{F}_{\text{const}} \doteq \big\{f \,|\, \exists c \in [0, 1] : f(x) = c, \forall x \in \mathbb{X}\big\}$ on the unit domain $\mathbb{X} = [0, 1]$ sampled uniformly $\boldsymbol{\mathcal{X}} \sim \mathcal{U}(\mathbb{X})$ using a standard Gaussian noise model $\mathcal{Y} = f_*(\boldsymbol{\mathcal{X}}) + \xi$ with $\xi \sim \mathcal{N}(0, 1)$. Then, for the absolute value norm $\psi(f, g) = |f(x) - g(x)|$, we have Condition (C1) by $G(\boldsymbol{\mathcal{X}}, \mathcal{Y}) = R$. This is matching for the 1-Lipschitz loss $\ell(y, \hat{y}) = |y - \hat{y}|$, for which Lemma 3.4 provides Condition (C3). Hence, as $\mathcal{H}_\psi(\epsilon, \mathcal{F}_{\text{const}}) = O\big(\ln(1/\epsilon)\big)$ by Lemma C.1 and $Q_n = \Theta(1)$ for constant functions $\mathcal{F}_{\text{const}}$, the bound of Theorem 3.2 with $\epsilon = \delta = \Theta(n^{-1/2})$ and $r_0 = \Theta\big(BR\sqrt{n}\big)$ scales by $n^{-1/2}$ up to logarithmic factors. This rate is near-minimax, as this problem is equivalent to estimating the mean of a Gaussian random variable from i.i.d. samples for which the $n^{-1/2}$ error guarantee is the best possible.

Strongly-convex losses

Now consider a loss function ℓ, which is η-*strongly convex* in its second argument, that is

$$\ell(y, \lambda\hat{y}_1 + (1-\lambda)\hat{y}_2) \leq \lambda\ell(y, \hat{y}_1) + (1-\lambda)\ell(y, \hat{y}_2) - \frac{\eta\lambda(1-\lambda)}{2}|\hat{y}_1 - \hat{y}_2|^2$$

holds for all $y, \hat{y}_1, \hat{y}_2 \in \mathbb{R}$ and $\lambda \in (0,1)$. Then, if $R_\mu(f_*) \leq R_\mu\big((f+f_*)/2\big)$ is satisfied for all $f \in \mathcal{F}(r_0)$ with some $r_0 \geq 0$, the Bernstein condition (3.3) holds with $C = 4/\eta$. To see this, proceed similarly to Bartlett et al. (2006, Lemma 7) by using the strong convexity property of ℓ to get for all $f \in \mathcal{F}$ that

$$\begin{aligned}\mathbb{E}\big[\mathcal{W}_f^2\big] &\leq (4/\eta)\Big(R_\mu(f) + R_\mu(f_*) - 2R_\mu\big((f+f_*)/2\big)\Big) \\ &\leq (4/\eta)\big(R_\mu(f) - R_\mu(f_*)\big) = (4/\eta)\mathbb{E}\big[\mathcal{Z}(f, f_*)\big]\,.\end{aligned} \tag{3.4}$$

Furthermore, notice that if $f_* \in \mathcal{F} \subseteq \mathcal{F}_*$ and $\mathcal{F}$ is midpoint convex, that is $f, g \in \mathcal{F}$ implies $(f+g)/2 \in \mathcal{F}$, or when f_* is a regression function, then the definition of f_* implies $R_\mu(f_*) \leq R_\mu\big((f+f_*)/2\big)$ for all $f \in \mathcal{F}$, which makes the Bernstein condition valid with $C = 4/\eta$ for any $r_0 \geq 0$.

When the loss ℓ is also Lipschitz, we have Lemma 3.4 with C and r_0 which are independent of n. However, Lipschitzness and strong convexity hold simultaneously only for bounded problems when both the range of $\mathcal{Y}$ and the estimators are bounded. Next, we review such an example by the cross-entropy loss, and relax the Lipschitz requirement for the squared loss.

Cross-entropy loss

For regression problems with $\mathcal{Y} \in (0,1)$ a.s., a reasonable choice might be the cross-entropy loss defined as $\ell_{\text{ce}}(y, \hat{y}) \doteq y\ln(y/\hat{y}) + (1-y)\ln\big((1-y)/(1-\hat{y})\big)$, for $y, \hat{y} \in (0,1)$. Fix $\lambda \in (0, 1/2)$ and consider estimates in $\mathcal{F} \subseteq \{\mathbb{X} \to [\lambda, 1-\lambda]\}$, which are λ-away from the boundary of $(0,1)$.

By having $\partial_z \ell_{\text{ce}}(y,z) = \frac{z-y}{z(1-z)}$ and $\partial_{zz}\ell_{\text{ce}}(y,z) = \frac{y}{z^2} + \frac{1-y}{(1-z)^2}$, we get that ℓ_{ce} over $(0,1) \times [\lambda, 1-\lambda]$ is $1/\lambda$-Lipschitz and $(1-\lambda)^{-2}$-strongly convex in its second argument. Hence, the conditions of Lemma 3.3 are satisfied with $B = 1-\lambda$, $C = 4(1-\lambda)^2$ due to (3.4), and $L(\boldsymbol{\mathcal{X}}, \mathcal{Y}) = 1/\lambda$.

Squared loss

Finally, we consider the squared loss $\ell = \ell_{\text{sq}}$, which is 2-strongly convex, hence (3.4) provides the Bernstein condition (3.3) with $C = 2$ when either $\mathcal{F}$ is midpoint convex, or if f_* is a regression function. Furthermore, the Bernstein condition also holds with $C = 1$ if $\mathcal{F}$ is a closed, convex class (Lecué and Mendelson, 2013, Theorem 6.1).

Because the squared loss is not uniformly Lipschitz over $\mathbb{R}$, we have to exploit the relaxed Lispchitz condition of Lemma 3.3 for sub-Gaussian problems, where $\sup_{f\in\mathcal{F}} \|\mathcal{W}_f\|_{\Psi_2} \le B$ and $\|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})\|_{\Psi_2} \le \sigma$ hold for some $B, \sigma \ge 0$. Then, write $\mathcal{Z}(f, f_*) = \big(\mathcal{W}_f - 2(\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})\big)\mathcal{W}_f$ for all $f \in \mathcal{F}$, and observe that $|\mathcal{Z}(f, f_*)| \le L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y})|\mathcal{W}_f|$ holds with $L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y}) = \big|\mathcal{W}_f - 2\big(\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})\big)\big|$. Further, notice that Lemma A.2d implies $\sup_{f\in\mathcal{F}} \|L(f, \boldsymbol{\mathcal{X}}, \mathcal{Y})\|_{\Psi_2} \le B + 2\sigma$. Putting these together, we obtain the following result:

Lemma 3.5. *Consider a regression problem, for which either f_* is a regression function and $\mathcal{F}$ is an arbitrary class, or $f_* \in \mathcal{F}$ and $\mathcal{F}$ is midpoint convex. Further, suppose that $\sup_{f\in\mathcal{F}} \|\mathcal{W}_f\|_{\Psi_2} \le B$ and $\|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})\|_{\Psi_2} \le \sigma$ hold with some $B, \sigma \ge 0$. Then $(\mu, \ell_{sq}, \mathcal{F}(r_0), f_*)$ satisfies Condition (C3) for any $r_0 > 0$ and $r = 1/2$ with $\theta \ge 240 Q_n \max\{B, \sigma\}^2$.*

Proof. Based on the discussion above, simply apply Lemma 3.3 with $C = 2$, $t = 10$, and $R = 3\max\{B, \sigma\}$. ∎

Later, we use Lemma 3.5 for Theorem 3.2 to derive near-minimax rates by an exact oracle inequality for linear regression settings (Lemma 4.2) such as lasso and ridge regression, and by an inexact oracle inequality for convex nonparametric sieved least squares estimation (Theorem 5.6).

3.3.2 Connection to the literature

Here we relate our main regression result (Theorem 3.2) to the literature.

For bounded problems, when both the hypothesis class $\mathcal{F}$ and the response $\mathcal{Y}$ are uniformly bounded, the most common approach is to localize and bound the Rademacher complexity around the optimum f_*. Such results, including

Bartlett et al. (2005, Section 5.2) and Koltchinskii (2011, Chapter 5), provide exact oracle inequalities for strongly convex, uniformly Lipschitz losses (including the squared loss over a bounded domain) by using the Bernstein condition (3.3) with $r_0 = 0$, and uniform boundedness on $\mathcal{F}$ and $\mathcal{Y}$. Our result, Theorem 3.2 combined with Lemma 3.3, relaxes the uniform Lipschitz property of the loss function, and extends these developments to sub-Gaussian classes $\mathcal{F}$ (that is when $\mathcal{W}_f$ is sub-Gaussian) and $\mathcal{Y}$.

We mention that there are extensions to sub-Gaussian noise (that is when $\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})$ is sub-Gaussian) for uniformly bounded hypothesis classes $\mathcal{F}$ and regression function f_*. Consider the following random-fixed design decomposition for the squared loss and an ERM($\mathcal{F}$) estimate f_n,

$$L_\mu(f_n, f_*) = \|f_n - f_*\|^2_{P_\mathbb{X}} \leq \sup_{f\in\mathcal{F}} \left\{ \|f - f_*\|^2_{P_\mathbb{X}} - 2\|f - f_*\|^2_{P_n} \right\} + 2\|f_n - f_*\|^2_{P_n},$$

where $\|g\|^2_{P_n} \doteq \frac{1}{n}\sum_{i=1}^n g^2(\boldsymbol{\mathcal{X}}_i)$ is the empirical L_2-norm based on the sample $\mathcal{D}_n$. Then the random design part (left subexpression) could be bounded by Theorem 3.3 of Bartlett et al. (2005), while the fixed design part (right) could be handled by Theorem 10.11 of van de Geer (2000). Alternatively, Corollary 1 of Györfi and Wegkamp (2008) could be also used to derive an inexact oracle inequality (when the bound depends on the approximation error of $\mathcal{F}$ to the regression function and its coefficient is larger than 1) for the expected excess risk $\mathbb{E}[L_\mu(f_n, f_*)]$ using the squared loss with sub-Gaussian noise and uniformly bounded $\mathcal{F}$. However, these results cannot handle unbounded classes $\mathcal{F}$, data-dependent classes (they use $\hat{\mathcal{F}}_n = \mathcal{F}_n = \mathcal{F}$), and do not provide exact oracle inequalities.

There has been a lot of emphasis lately to develop exact oracle inequalities without the uniform boundedness restriction on $\mathcal{F}$, especially for linear regression settings. Lecué and Mendelson (2013) introduced an alternative technique to local Rademacher complexities and proved exact oracle inequalities with the squared loss for sub-Gaussian classes and noise. Recently, Liang et al. (2015) modified the local Rademacher complexity results, still for the squared loss, making them capable to even reach beyond the sub-Gaussian case, including some heavy-tailed distributions. However, these results depend on a uniform

bound on the kurtosis, $\sup_{f \in \mathcal{F}} \mathbb{K}_0[\mathcal{W}_f] \leq K < \infty$. This dependence is either ignored (Lecué and Mendelson, 2013, Theorem A) or linear (Liang et al., 2015, Theorem 7). As K can grow arbitrarily large, even for bounded problems, these results cannot provide near-minimax rates for all sub-Gaussian (or even bounded) problems for which K might scale with the sample size n linearly (see the example in Section 4.1.2). Compared to these results, the bound of Theorem 3.2 grows only logarithmically in K (see θ in Lemma 3.3), which allows us to eliminate this dependence altogether by suffering a $\ln(n)$ penalty for the excess risk bound in the worst case.

We also mention the work of Mendelson (2014, Theorem 2.2), which proves an upper bound on the squared deviation of an ERM estimate f_n using the squared loss $\ell = \ell_{\text{sq}}$ and f_*, that is $\mathbb{E}\big[\mathcal{W}_{f_n}^2\big]$ with $\mathcal{W}_{f_n} \doteq f_n(\boldsymbol{\mathcal{X}}) - f_*(\boldsymbol{\mathcal{X}})$, instead of the excess risk $L_\mu(f_n, f_*)$. However, as pointed out by Shamir (2015, Section 1), the squared deviation can be arbitrarily smaller than the excess risk.

Finally, we mention that Theorem 3.2 is a significant extension of our previous result (Balázs et al., 2015, Theorem 3.1) by supporting many loss functions beyond the squared loss, sub-Gaussian settings for unbounded, data-dependent hypothesis classes, and improving the expected value result to a probabilistic guarantee.

3.3.3 Proof of the upper bound

In this section, we finally prove our main result, Theorem 3.2, our upper bound on the excess risk of ERM estimators.

The strategy is to first decompose the excess risk to "supremal" and approximation error terms. Then we reduce the former to a general concentration inequality (Lemma 3.6) and upper bound the latter using the ERM property (3.1).

First, use the union bound with $\mathbb{P}\{\hat{\mathcal{F}}_n \subseteq \mathcal{F}_n \subseteq \mathcal{F}\} \geq 1 - \gamma/2$, and

$L_\mu(f, f_*) \le r_0/n$ for all $f \in \mathcal{F} \setminus \mathcal{F}(r_0)$, to get

$$\begin{aligned}
&\mathbb{P}\Big\{ L_\mu(f_n, f_*) > b + \frac{B_*}{r} + \frac{r_0}{n} + \frac{\ln(4/\gamma)}{t} \Big\} \\
&\quad \le \frac{\gamma}{2} + \mathbb{P}\Big\{ L_\mu(f_n, f_*)\, \mathbb{I}\{\hat{\mathcal{F}}_n \subseteq \mathcal{F}_n \subseteq \mathcal{F}(r_0)\} > b + \frac{B_*}{r} + \frac{r_0}{n} + \frac{\ln(4/\gamma)}{t} \Big\},
\end{aligned} \tag{3.5}$$

for any $b, t > 0$. Next, recall the definition of the empirical excess risk $L_n(f, f_*) \doteq R_n(f) - R_n(f_*)$, and notice that if $\hat{\mathcal{F}}_n \subseteq \mathcal{F}_n \subseteq \mathcal{F}(r_0)$ holds, we can use the (α, β)-ERM($\mathcal{F}_n$) property (3.1) with any nonnegative penalty function $\beta \ge 0$ to decompose the excess risk as

$$\begin{aligned}
L_\mu(f_n, f_*) &= L_\mu(f_n, f_*) - \frac{1}{r} L_n(f_n, f_*) + \frac{1}{r} L_n(f_n, f_*) \\
&\le \frac{1}{r} \sup_{f \in \mathcal{F}_n} \Big\{ r L_\mu(f, f_*) - L_n(f, f_*) \Big\} + \frac{1}{r} \inf_{f \in \mathcal{F}_n} L_n(f, f_*) + \beta(f) + \alpha \\
&\le \frac{1}{r} \sup_{f \in \mathcal{F}(r_0)} \Gamma_r(f, \mathcal{D}_n) + \frac{1}{r} \Big(\inf_{f \in \hat{\mathcal{F}}_n} L_n(f, f_*) + \beta(f) + \alpha \Big),
\end{aligned}$$

where $\Gamma_r(f, \mathcal{D}_n) \doteq r L_\mu(f, f_*) - L_n(f, f_*)$. Combining this with (3.5), using the union bound with $\mathbb{P}\big\{ \inf_{f \in \hat{\mathcal{F}}_n} L_n(f, f_*) + \beta(f) + \alpha > B_* \big\} \le \gamma/4$, and Markov's inequality, we obtain

$$\begin{aligned}
&\mathbb{P}\Big\{ L_\mu(f_n, f_*) > b + \frac{B_*}{r} + \frac{r_0}{n} + \frac{\ln(4/\gamma)}{t} \Big\} \\
&\quad \le \frac{3\gamma}{4} + \mathbb{P}\Big\{ \frac{1}{r} \sup_{f \in \mathcal{F}(r_0)} \Gamma_r(f, \mathcal{D}_n) > b + \frac{r_0}{n} + \frac{\ln(4/\gamma)}{t} \Big\} \\
&\quad \le \frac{3\gamma}{4} + \frac{\gamma}{4} \mathbb{E}\Big[e^{(t/r) \sup_{f \in \mathcal{F}(r_0)} \Gamma_r(f, \mathcal{D}_n)} \Big] e^{-tb}.
\end{aligned} \tag{3.6}$$

Then notice that we get a bound on the excess risk $L_\mu(f_n, f_*)$ with probability at least $1 - \gamma$ for any b and t which satisfies $\mathbb{C}_t\big[e^{(1/r) \sup_{f \in \mathcal{F}(r_0)} \Gamma_r(f, \mathcal{D}_n)} \big] \le b$. To find such values b and t, we use the following concentration inequality:

Lemma 3.6. *Let $(\mathcal{P}, \psi)$ be a separable metric space,[4] $\mathcal{W}$ be a random variable on some set $\mathbb{W}$, and $\Gamma : \mathcal{P} \times \mathbb{W} \to \mathbb{R}$ be a function. Furthermore, define the function $\Lambda(p, w) \doteq \Gamma(p, w) - \mathbb{E}[\Gamma(p, \mathcal{W})]$ for all $p \in \mathcal{P}$, $w \in \mathbb{W}$, and suppose there exist $\tau : \mathbb{W} \to [0, \infty)$, $S \ge 0$ and $\theta > 0$ satisfying the following conditions for all $p, q \in \mathcal{P}$:*

[4] In a separable metric space, we have a countable dense bases. So the suprema over $\mathcal{P}$ could be redefined over a countable set, keeping the resulting random variable measurable.

(a) $\Lambda(p,\mathcal{W}) - \Lambda(q,\mathcal{W}) \le \psi(p,q)\tau(\mathcal{W})$ *a.s.,*

(b) $\Lambda(p,\mathcal{W}) - \Lambda(q,\mathcal{W})$ *is centered sub-Gaussian with* $S\psi(p,q)$*,*

(c) $\mathbb{E}\big[\exp\big(\Gamma(p,\mathcal{W})/\theta\big)\big] \le 1.$

Then, for all $0 < \delta \le \epsilon$ *and* $t \in (0, 1/(2\theta)]$*,*

$$\mathbb{C}_t\left[\sup_{p\in\mathcal{P}} \Gamma(p,W)\right] \le \theta\mathcal{H}_\psi(\epsilon,\mathcal{P}) + 16S\int_\delta^\epsilon \max\left\{\sqrt{\mathcal{H}_\psi(z,\mathcal{P})},\, 8tSz^2/\delta\right\} dz + 8\delta\,\mathbb{C}_{2t}[\tau(W)]\,.$$

Furthermore, the result holds without Condition (b) *(that is* $S = \infty$*) with* $\epsilon = \delta$ *defining* $\infty \cdot 0 = 0$*.*

The proof of Lemma 3.6 is postponed to Section 3.3.4.

Recall that $R_\mu(f) - R_\mu(f_*) = L_\mu(f,f_*) = \mathbb{E}[\mathcal{Z}(f,f_*)]$. Using this, Γ_r can be rewritten for any f as $\Gamma_r(f,\mathcal{D}_n) = r\mathbb{E}\big[\mathcal{Z}(f,f_*)\big] - \frac{1}{n}\sum_{i=1}^n \mathcal{Z}_i(f,f_*)$. Then, define $\Lambda(f,\mathcal{D}_n) \doteq \Gamma_r(f,\mathcal{D}_n) - \mathbb{E}[\Gamma_r(f,\mathcal{D}_n)]$, and observe that for all $f,g \in \mathcal{F}(r_0)$, we have $\Lambda(f,\mathcal{D}_n) = \mathbb{E}\big[\mathcal{Z}(f,f_*)\big] - \frac{1}{n}\sum_{i=1}^n \mathcal{Z}_i(f,f_*)$, and

$$\Lambda(f,\mathcal{D}_n) - \Lambda(g,\mathcal{D}_n) = \mathbb{E}\big[\mathcal{Z}(f,g)\big] - \frac{1}{n}\sum_{i=1}^n \mathcal{Z}_i(f,g)\,. \tag{3.7}$$

Hence, Condition (C2) implies that Λ is the sum of n independent and centered $(S\psi(f,g))$-sub-Gaussian random variables. Then by Lemma A.1f, Λ is also centered sub-Gaussian with $S\psi(p,q)/\sqrt{n}$, and so Λ satisfies Lemma 3.6b for $\mathcal{P} \doteq \mathcal{F}(r_0)$ and $\mathcal{W} \doteq \mathcal{D}_n$ with $S/\sqrt{n}$.

Next, using Condition (C1), we can upper bound (3.7) by $\tau(\mathcal{D}_n)\,\psi(f,g)$, where $\tau(\mathcal{D}_n) \doteq \frac{1}{n}\sum_{i=1}^n \big\{\mathbb{E}[G(\boldsymbol{\mathcal{X}}_i,\mathcal{Y}_i)] + G(\boldsymbol{\mathcal{X}}_i,\mathcal{Y}_i)\big\}$. Then Lemma 3.6a also holds for Λ.

Finally, using the i.i.d. property of the sample $\mathcal{D}_n$ and Condition (C3), we have for any $f \in \mathcal{F}(r_0)$ that

$$\mathbb{E}\Big[e^{\Gamma_r(f,\mathcal{D}_n)/(\theta/n)}\Big] = \prod_{i=1}^n \mathbb{E}\left[e^{(r/\theta)\mathbb{E}[\mathcal{Z}(f,f_*)] - (1/\theta)\mathcal{Z}_i(f,f_*)}\right] \le 1\,, \tag{3.8}$$

so Γ_r satisfies Lemma 3.6c with θ/n.

Hence, all the requirements of Lemma 3.6 hold, and with $t \doteq n/(2\theta)$ we obtain

$$\mathbb{C}_t\left[\sup_{f\in\mathcal{F}(r_0)} \Gamma_r(f,\mathcal{D}_n)\right] \le \frac{\theta\,\mathcal{H}_\psi(\epsilon,\mathcal{F})}{n} + 16S\int_\delta^\epsilon \max\left\{\sqrt{\frac{\mathcal{H}_\psi(z,\mathcal{F})}{n}}, \frac{4S}{\theta\delta}\,z^2\right\} dz + 16\,\delta\,\mathbb{C}_{1/\theta}\big[G(\boldsymbol{\mathcal{X}},\mathcal{Y})\big]\,,$$

where we also used $\mathcal{H}_\psi(z,\mathcal{F}(r_0)) \le \mathcal{H}_\psi(z,\mathcal{F})$ and

$$\mathbb{C}_{n/\theta}\big[\tau(\mathcal{D}_n)\big] = \mathbb{E}\big[G(\boldsymbol{\mathcal{X}},\mathcal{Y})\big] + \mathbb{C}_{n/\theta}\Big[\sum_{i=1}^n G(\boldsymbol{\mathcal{X}}_i,\mathcal{Y}_i)\Big] \le 2\,\mathbb{C}_{1/\theta}\big[G(\boldsymbol{\mathcal{X}},\mathcal{Y})\big]\,,$$

due to $\mathbb{E}[G(\boldsymbol{\mathcal{X}},\mathcal{Y})] \le \mathbb{C}_{1/\theta}[G(\boldsymbol{\mathcal{X}},\mathcal{Y})]$, Lemma B.2 and the i.i.d. property of the sample $\mathcal{D}_n$. Combining this with (3.6), we get the claim of Theorem 3.2.

3.3.4 Suprema of empirical processes

In this section, we prove Lemma 3.6, which we used in Section 3.3.3 as the main tool to prove Theorem 3.2. For this, we start with finite class lemmas, then adapt the classical chaining argument (for example, Pollard, 1990, Section 3; van de Geer, 2000, Chapter 3; Boucheron et al., 2012, Section 13.1) to our setting, and finally put these together to prove Lemma 3.6.

First, consider the well-known inequality about the maximum of finitely many sub-Gaussian random variables (for example, Cesa-Bianchi and Lugosi, 1999, Lemma 7; Boucheron et al., 2012, Theorem 2.5), adapted to the cumulant generating function.

Lemma 3.7. *Let $\mathcal{P}$ be a finite, nonempty set (that is $1 \le |\mathcal{P}| < \infty$), $\sigma \in [0,\infty)$ and $\mathcal{W}_p$ be centered σ-sub-Gaussian random variables for all $p \in \mathcal{P}$. Then $\mathbb{C}_t\big[\max_{p\in\mathcal{P}} \mathcal{W}_p\big] \le \max\big\{\sigma\sqrt{2\ln|\mathcal{P}|}, t\sigma^2\big\}$ for all $t > 0$.*

Proof. Let $s \doteq \sqrt{2\ln|\mathcal{P}|}/\sigma$. Then bounding $\max_{p\in\mathcal{P}} x_p$ by $\ln\sum_{p\in\mathcal{P}} e^{x_p}$, and using the sub-Gaussian condition on $\mathcal{W}_p$, we have

$$\begin{aligned}\mathbb{C}_s\big[\max_{p\in\mathcal{P}} \mathcal{W}_p\big] &= \frac{1}{s}\ln\mathbb{E}\Big[\exp\max_{p\in\mathcal{P}} s\mathcal{W}_p\Big] \\ &\le \frac{1}{s}\ln\sum_{p\in\mathcal{P}}\mathbb{E}\big[e^{s\mathcal{W}_p}\big] \le \frac{\ln|\mathcal{P}|}{s} + \frac{s\sigma^2}{2} = \sigma\sqrt{2\ln|\mathcal{P}|}\,.\end{aligned}$$

Then the claim follows for $t \le s$ by monotonicity, $\mathbb{C}_t[\cdot] \le \mathbb{C}_s[\cdot]$. Otherwise, for $t > s$, we have $t^2\sigma^2/2 > \ln|\mathcal{P}|$, hence by a similar derivation we obtain $\mathbb{C}_t\big[\max_{p\in\mathcal{P}} \mathcal{W}_p\big] \le (1/t)\ln|\mathcal{P}| + t\sigma^2/2 \le t\sigma^2$. ■

When a moment condition, similar to Condition (C3), is satisfied for $\mathcal{W}_p$, Lemma 3.7 can be strengthened by the following result (for further explanation, see the discussion after the lemma).

Lemma 3.8. *Let $\mathcal{P}$ be a finite, nonempty set (that is $1 \le |\mathcal{P}| < \infty$), and $\mathcal{W}_p$ be random variables such that $\max_{p\in\mathcal{P}} \mathbb{E}\big[\exp(\mathcal{W}_p/\theta)\big] \le 1$ holds for some $\theta > 0$. Then $\mathbb{C}_t\big[\max_{p\in\mathcal{P}} \mathcal{W}_p\big] \le \theta\ln|\mathcal{P}|$ for all $t \in (0, 1/\theta]$.*

Proof. Bound $\max_{p\in\mathcal{P}} x_p$ by $\ln\sum_{p\in\mathcal{P}} e^{x_p}$, and use the moment condition on $\mathcal{W}_p$, to obtain

$$\mathbb{C}_{1/\theta}\Big[\max_{p\in\mathcal{P}} \mathcal{W}_p\Big] = \theta\ln\mathbb{E}\Big[\exp\max_{p\in\mathcal{P}} \mathcal{W}_p/\theta\Big] \le \theta\ln\sum_{p\in\mathcal{P}} \mathbb{E}\big[e^{\mathcal{W}_p/\theta}\big] \le \theta\ln(|\mathcal{P}|)\,.$$

Then the claim follows for $t \le 1/\theta$ by the monotonicity of $s \mapsto \mathbb{C}_s[\mathcal{W}]$. ■

To see that Lemma 3.8 is indeed stronger than Lemma 3.7 for our purposes, notice that they scale differently in their parameters θ and σ when applied to averages of independent random variables. If $\mathcal{W}_p^{(1)}, \ldots, \mathcal{W}_p^{(n)}$ are $n \in \mathbb{N}$ independent centered σ-sub-Gaussian random variables, their average, $\frac{1}{n}\sum_{i=1}^n \mathcal{W}_p^{(i)}$ is a $(\sigma/\sqrt{n})$-sub-Gaussian random variable (Lemma A.1f). On the other hand, it is straightforward to show (as we did by (3.8)) that if $\mathcal{W}_p^{(1)}, \ldots, \mathcal{W}_p^{(n)}$ are n independent (not necessarily centered) random variables with $\mathbb{E}\big[\exp(\mathcal{W}_p^{(i)}/\theta)\big] \le 1$, then their average satisfies the moment condition with θ/n. This speed-up, from $\frac{\sigma}{\sqrt{n}}\ln|\mathcal{P}|$ to $\frac{\theta}{n}\ln|\mathcal{P}|$, will allow us to derive better bounds when the moment condition (C3) holds.

We now extend Lemma 3.7 to infinite classes by adapting the chaining argument to the cumulant-generating function. The proof goes along the development of Lemma 3.4 of Pollard (1990), replacing the packing sets by internal covering numbers (for better numerical constants) and the sample continuity condition by uniform Lipschitzness (for truncating the integral at δ). The result is also similar to Proposition 3 of Cesa-Bianchi and Lugosi (1999), which

works for the sub-Gaussian case, and uses external covering numbers with a slightly different chaining argument.

Lemma 3.9. *Let $(\mathcal{P}, \psi)$ be a separable metric space, $\mathcal{W}$ be a random variable on some set $\mathbb{W}$, and $\Lambda : \mathcal{P} \times \mathbb{W} \to \mathbb{R}$ be a function. Furthermore, suppose there exist $\tau : \mathbb{W} \to [0, \infty)$, $\beta \geq 0$, $p_0 \in \mathcal{P}$ and $S \geq 0$ for which the following conditions hold for all $p, q \in \mathcal{P}$:*

(a) $\Lambda(p, \mathcal{W}) - \Lambda(q, \mathcal{W}) \leq \psi(p, q)\tau(\mathcal{W})$ *a.s.,*

(b) $\Lambda(p, \mathcal{W}) - \Lambda(q, \mathcal{W})$ *is centered sub-Gaussian with* $S\psi(p, q)$*,*

(c) $\beta \geq \sup_{p \in \mathcal{P}} \psi(p, p_0)$*.*

Then, for all $\delta \in (0, \beta/2]$ and $t > 0$,

$$\mathbb{C}_t\left[\sup_{p \in \mathcal{P}} \Lambda(p, \mathcal{W})\right] \leq 4S \int_{\delta}^{\beta/2} \max\left\{\sqrt{2\,\mathcal{H}_{\psi}(z, \mathcal{P})},\, 7tSz^2/\delta\right\} dz + 4\delta\, \mathbb{C}_{2t}[\tau(\mathcal{W})] + \mathbb{C}_{(\beta/\delta)t}\left[\Lambda(p_0, \mathcal{W})\right].$$

Additionally, the result holds without Condition (b) (that is when $S = \infty$) with $\delta = \beta/2$ defining $\infty \cdot 0 = 0$.

Proof. If there exists $z \in (\delta, \beta/2]$ such that $\mathcal{N}_{\psi}(z, \mathcal{P}) = \infty$, then the integral is infinite and so the claim is trivial. The claim is also trivial for $\beta = 0$ or $S = 0$. So assume that $0 < \beta$, $S \in (0, \infty)$ and $\mathcal{N}_{\psi}(z, \mathcal{P}) < \infty$ for all $z \in (\delta, \beta/2]$.

Let $m \in \mathbb{N}$ be such that $2\delta \leq \beta/2^m < 4\delta$. Now let $\mathcal{P}_0 = \{p_0\}$, $\epsilon_0 = \beta$, $\epsilon_k = \beta/2^k$ and $\mathcal{P}_k$ be an ϵ_k-cover of $\mathcal{P}$ under ψ having minimal cardinality for all $k \in \{1, \ldots, m\}$. Notice that $\mathcal{P}_0$ is an ϵ_0-cover by Condition (c). Furthermore, let $\gamma_k = 2t\beta/\epsilon_{m-k} = 2t2^{m-k}$ and $q_k(p) \in \operatorname{argmin}_{q \in \mathcal{P}_k} \psi(p, q)$ be the closest element to $p \in \mathcal{P}$ in $\mathcal{P}_k$ for all $k = 0, \ldots, m$.

Fix some $k \in \{0, \ldots, m-1\}$ and $p \in \mathcal{P}_{k+1}$. When $k = 0$, we have $\psi(p, q_k(p)) = \psi(p, p_0) \leq \beta = \epsilon_0$, while for $k > 0$, the definition of $\mathcal{P}_k$ implies that $\psi(p, q_k(p)) \leq \epsilon_k$. So by Condition (b), $\Lambda(p, \mathcal{W}) - \Lambda(q_k(p), \mathcal{W})$ is a centered $\epsilon_k S$-sub-Gaussian random variable. Combining this with Lemma 3.7, using $\mathbb{C}_{\gamma}[\mathcal{Z}_1 + \mathcal{Z}_2] \leq \mathbb{C}_{2\gamma}[\mathcal{Z}_1] + \mathbb{C}_{2\gamma}[\mathcal{Z}_2]$, which holds for any $\gamma > 0$ and random

variables $\mathcal{Z}_1, \mathcal{Z}_2$, we can chain maximal inequalities for all $k = 0, \ldots, m-1$ as

$$\begin{aligned}
&\mathbb{C}_{\gamma_{k+1}}\left[\max_{p\in\mathcal{P}_{k+1}} \Lambda(p,\mathcal{W})\right] \\
&\quad= \mathbb{C}_{\gamma_{k+1}}\left[\max_{p\in\mathcal{P}_{k+1}} \Big\{\Lambda(q_k(p),\mathcal{W}) + \Lambda(p,\mathcal{W}) - \Lambda(q_k(p),W)\Big\}\right] \\
&\quad\le \mathbb{C}_{\gamma_k}\left[\max_{p\in\mathcal{P}_k} \Lambda(p,\mathcal{W})\right] + \mathbb{C}_{\gamma_k}\left[\max_{p\in\mathcal{P}_{k+1}} \Big\{\Lambda(p,\mathcal{W}) - \Lambda(q_k(p),\mathcal{W})\Big\}\right] \qquad (3.9)\\
&\quad\le \mathbb{C}_{\gamma_k}\left[\max_{p\in\mathcal{P}_k} \Lambda(p,\mathcal{W})\right] + \max\left\{\epsilon_k S\sqrt{2\ln|\mathcal{P}_{k+1}|},\ \gamma_k\epsilon_k^2 S^2\right\} \\
&\quad= \mathbb{C}_{\gamma_k}\left[\max_{p\in\mathcal{P}_k} \Lambda(p,\mathcal{W})\right] + \epsilon_k S\max\left\{\sqrt{2\ln\mathcal{N}_\psi(\epsilon_{k+1},\mathcal{P})},\ \gamma_k\epsilon_k S\right\}.
\end{aligned}$$

Using Condition (a) and $\gamma_m = 2t$, we further have

$$\begin{aligned}
&\mathbb{C}_t\left[\sup_{p\in\mathcal{P}} \Lambda(p,\mathcal{W})\right] \\
&\quad= \mathbb{C}_t\left[\sup_{p\in\mathcal{P}} \Big\{\Lambda\big(q_m(p),W\big) + \Lambda(p,\mathcal{W}) - \Lambda\big(q_m(p),\mathcal{W}\big)\Big\}\right] \\
&\quad\le \mathbb{C}_{2t}\left[\max_{p\in\mathcal{P}_m} \Lambda(p,\mathcal{W})\right] + \mathbb{C}_{2t}\left[\sup_{p\in\mathcal{P}} \Big\{\Lambda(p,\mathcal{W}) - \Lambda\big(q_m(p),\mathcal{W}\big)\Big\}\right] \qquad (3.10)\\
&\quad\le \mathbb{C}_{\gamma_m}\left[\max_{p\in\mathcal{P}_m} \Lambda(p,\mathcal{W})\right] + \left(\sup_{p\in\mathcal{P}} \psi\big(p,q_m(p)\big)\right)\mathbb{C}_{2t}\big[\tau(\mathcal{W})\big].
\end{aligned}$$

By $\psi(p, q_m(p)) \le \epsilon_m < 4\delta$, the second term can be bounded by $4\delta\,\mathbb{C}_{2t}[\tau(\mathcal{W})]$. To bound the first term, we use (3.9) repeatedly with $k = m-1, m-2, \ldots, 0$, $\gamma_0 = 2t\beta/\epsilon_m \le (\beta/\delta)t$ and $\gamma_k\epsilon_k = 8t\epsilon_{k+1}^2/\epsilon_m \le (4t/\delta)\epsilon_{k+1}^2$ to get

$$\begin{aligned}
\mathbb{C}_t\left[\sup_{p\in\mathcal{P}} \Lambda(p,\mathcal{W})\right] < S\sum_{k=0}^{m-1}\epsilon_k \max\left\{\sqrt{2\ln\mathcal{N}_\psi(\epsilon_{k+1},\mathcal{P})},\ (4tS/\delta)\epsilon_{k+1}^2\right\} \\
+ 4\delta\,\mathbb{C}_{2t}[\tau(\mathcal{W})] + \mathbb{C}_{(\beta/\delta)t}[\Lambda(p_0,\mathcal{W})].
\end{aligned}$$

Now notice that $(\beta^3/2)\,2^{-3(k+1)} = (12/7)\int_{\beta\,2^{-(k+2)}}^{\beta\,2^{-(k+1)}} z^2\,dz$, and the nondecreasing property of covering numbers implies

$$\begin{aligned}
&\epsilon_k \max\left\{\sqrt{2\ln\mathcal{N}_\psi\big(\epsilon_{k+1},\mathcal{P}\big)},\ (4tS/\delta)\epsilon_{k+1}^2\right\} \\
&\qquad= 4\,\frac{\beta}{2^{k+2}}\max\left\{\sqrt{2\ln\mathcal{N}_\psi\big(\beta/2^{k+1},\mathcal{P}\big)},\ (4tS/\delta)\big(\beta/2^{k+1}\big)^2\right\} \\
&\qquad\le 4\int_{\beta/2^{k+2}}^{\beta/2^{k+1}} \max\left\{\sqrt{2\ln\mathcal{N}_\psi\big(z,\mathcal{P}\big)},\ (7tS/\delta)z^2\right\} dz,
\end{aligned}$$

for all $k = 0, \dots, m-1$. Combining this with $\delta \le \beta/2^{m+1}$ proves the claim for all $\delta \in (0, \beta/2]$ and $S \in (0, \infty)$.

Finally notice that for $\delta = \beta/2$ (that is $m = 0$), we use only (3.10) and ignore Condition (b) altogether, hence justifying the $0 \cdot \infty = 0$ convention for the $S = \infty$ case. ■

Now we extend the improved finite class lemma (Lemma 3.8) to infinite classes and prove Lemma 3.6. The idea behind the proof is to apply Lemma 3.8 in the first step of the chain and the previously developed chaining technique (Lemma 3.9) to the remainder.

Proof of Lemma 3.6. Fix $0 < \delta \le \epsilon$. When $\mathcal{N}_\psi(z, \mathcal{P}) = \infty$ for some $z \in (\delta, \epsilon]$, the claim is trivial, so we can assume that $\mathcal{N}_\psi(z, \mathcal{P}) < \infty$ for all $z \in (\delta, \epsilon]$.

Let $\mathcal{P}_\epsilon$ be an ϵ-cover of $\mathcal{P}$ under ψ with minimal cardinality and define $q_p \in \operatorname{argmin}_{q \in \mathcal{P}_\epsilon} \psi(p, q)$, the closest element to $p \in \mathcal{P}$ in $\mathcal{P}_\epsilon$. Due to Jensen's inequality and Condition (c), $\mathbb{E}[\Gamma(p, \mathcal{W})] \le 0$ holds for all $p \in \mathcal{P}$. Define[5] $q_p^* \in \operatorname{argmax}_{q \in \mathcal{P}: \psi(q, q_p) \le \epsilon} \mathbb{E}[\Gamma(q, \mathcal{W})]$. Then, for all $p \in \mathcal{P}$, due to $\psi(q_p, p) \le \epsilon$, $\mathbb{E}[\Gamma(p, \mathcal{W})] \le \mathbb{E}[\Gamma(q_p^*, \mathcal{W})]$. Further, $\psi(p, q_p^*) \le \psi(p, q_p) + \psi(q_p, q_p^*) \le 2\epsilon$ so $\mathcal{P}_\epsilon^* \doteq \{q_p^* : q_p \in \mathcal{P}_\epsilon\}$ is a 2ϵ-cover of $\mathcal{P}$ under ψ with $|\mathcal{P}_\epsilon^*| = |\mathcal{P}_\epsilon| = \mathcal{N}_\psi(\epsilon, \mathcal{P})$.

Now, for the first step of the chain, consider the following decomposition,

$$
\begin{aligned}
&\sup_{p \in \mathcal{P}} \Gamma(p, \mathcal{W}) \\
&\quad = \sup_{p \in \mathcal{P}} \Big\{ \Gamma(q_p^*, \mathcal{W}) + \Gamma(p, \mathcal{W}) - \Gamma(q_p^*, \mathcal{W}) \Big\} \\
&\quad \le \max_{q \in \mathcal{P}_\epsilon^*} \Gamma(q, \mathcal{W}) + \sup_{p \in \mathcal{P}} \Big\{ \Lambda(p, \mathcal{W}) - \Lambda(q_p^*, \mathcal{W}) + \mathbb{E}\Big[\Gamma(p, \mathcal{W}) - \Gamma(q_p^*, \mathcal{W}) \Big] \Big\} \\
&\quad \le \max_{q \in \mathcal{P}_\epsilon^*} \Gamma(q, \mathcal{W}) + \sup_{p \in \mathcal{P}} \Big\{ \Lambda(p, \mathcal{W}) - \Lambda\big(q_p^*, \mathcal{W}\big) \Big\} . \qquad (3.11)
\end{aligned}
$$

Then by Lemma 3.8 and Condition (c), we obtain for any $2t \le 1/\theta$ that

$$
\mathbb{C}_{2t}\left[\max_{q \in \mathcal{P}_\epsilon^*} \Gamma(q, \mathcal{W}) \right] \le \theta \ln |\mathcal{P}_\epsilon^*| = \theta \ln \mathcal{N}_\psi(\epsilon, \mathcal{P}) . \qquad (3.12)
$$

The rest of the proof is about to upper bound the scaled cumulant of the supremal term on the right side of (3.11), $\mathbb{C}_{2t}\big[\sup_{p \in \mathcal{P}} \Lambda(p, \mathcal{W}) - \Lambda(q_p^*, \mathcal{W}) \big]$, using the chaining result of Lemma 3.9.

[5] If such q_p^* element does not exist, one can choose another element which is arbitrary close to the supremum and shrink the gap to zero at the end of the analysis.

Let $\mathcal{K} \doteq \{(p, q_p^*) : p \in \mathcal{P}\} \subset \mathcal{P} \times \mathcal{P}_\epsilon^*$ and choose $p_0 \in \operatorname{argmax}_{p \in \mathcal{P}_\epsilon^*} \mathbb{E}[\Gamma(p, \mathcal{W})]$ so that $p_0 = q_{p_0}^*$ (since $\psi(q_{p_0}, p_0) \le \epsilon$), implying $(p_0, p_0) \in \mathcal{K}$. Further, define

$$\tilde{\Lambda}\big((p_1, q_1^*), w\big) \doteq \Lambda(p_1, w) - \Lambda(q_1^*, w)\,,$$
$$\tilde{\psi}\big((p_1, q_1^*), (p_2, q_2^*)\big) \doteq \min\big\{\psi(p_1, p_2) + \psi(q_1^*, q_2^*),\, 4\epsilon\big\}\,,$$

for all $(p_1, q_1^*), (p_2, q_2^*) \in \mathcal{P} \times \mathcal{P}_\epsilon^*$, and $w \in \mathbb{W}$. Now notice that $(\mathcal{P} \times \mathcal{P}_\epsilon^*, \tilde{\psi})$ is a metric space[6], and by $(p_0, p_0) \in \mathcal{K}$, for any $(p, q_p^*) \in \mathcal{K}$ we have that

$$\tilde{\Lambda}\big((p_0, p_0), \mathcal{W}\big) = 0 \text{ a.s.}\,, \quad \tilde{\psi}\big((p_0, p_0), (p, q_p^*)\big) \le 4\epsilon\,, \tag{3.13}$$

hence, $\tilde{\psi}$ and (p_0, p_0) satisfies Lemma 3.9c with $\beta = 4\epsilon$. Since $\psi(p, q_p^*) \le 2\epsilon$ holds for all $(p, q_p^*) \in \mathcal{K}$, Condition (b) implies for all $(p, q_p^*), (h, q_h^*) \in \mathcal{K}$ that

$$\begin{aligned}
\tilde{\Lambda}\big((p, q_p^*), \mathcal{W}\big) - \tilde{\Lambda}\big((h, q_h^*), \mathcal{W}\big) &= \Lambda(f, \mathcal{W}) - \Lambda(h, \mathcal{W}) + \Lambda(q_h^*, \mathcal{W}) - \Lambda(q_p^*, \mathcal{W}) \\
&\quad \text{is sub-Gaussian with } \big(\psi(p, h) + \psi(q_h^*, q_p^*)\big)S\,, \\
\tilde{\Lambda}\big((p, q_p^*), \mathcal{W}\big) - \tilde{\Lambda}\big((h, q_h^*), \mathcal{W}\big) &= \Lambda(p, \mathcal{W}) - \Lambda(q_p^*, \mathcal{W}) + \Lambda(q_h^*, \mathcal{W}) - \Lambda(h, \mathcal{W}) \\
&\quad \text{is sub-Gaussian with } \big(\psi(p, q_p^*) + \psi(q_h^*, h)\big)S \le 4\epsilon S\,,
\end{aligned}$$

hence, $\tilde{\Lambda}$ and $\tilde{\psi}$ satisfies Lemma 3.9b with S. Similarly, Condition (a) implies that

$$\begin{aligned}
\tilde{\Lambda}\big((p, q_p^*), \mathcal{W}\big) - \tilde{\Lambda}\big((h, q_h^*), \mathcal{W}\big) &= \Lambda(p, \mathcal{W}) - \Lambda(h, \mathcal{W}) + \Lambda(q_h^*, \mathcal{W}) - \Lambda(q_p^*, \mathcal{W}) \\
&\le \big(\psi(p, h) + \psi(q_h^*, q_p^*)\big)\tau(\mathcal{W}) \text{ a.s.}\,, \\
\tilde{\Lambda}\big((p, q_p^*), \mathcal{W}\big) - \tilde{\Lambda}\big((h, q_h^*), \mathcal{W}\big) &= \Lambda(p, \mathcal{W}) - \Lambda(q_p^*, \mathcal{W}) + \Lambda(q_h^*, \mathcal{W}) - \Lambda(h, \mathcal{W}) \\
&\le \big(\psi(p, q_p^*) + \psi(q_h^*, h)\big)\tau(\mathcal{W}) \le 4\epsilon\,\tau(\mathcal{W}) \text{ a.s.}\,,
\end{aligned}$$

so $\tilde{\Lambda}$ and $\tilde{\psi}$ satisfies Lemma 3.9a with τ.

Then the requirements of Lemma 3.9 hold (using $\mathcal{P} \leftarrow \mathcal{K}$, $\Lambda \leftarrow \tilde{\Lambda}$, $\psi \leftarrow \tilde{\psi}$, $p_0 \leftarrow (p_0, p_0)$, $\beta \leftarrow 4\epsilon$, $\delta \leftarrow 2\delta$), and by (3.13), $\mathbb{C}_{(\epsilon/\delta)t}[\tilde{\Lambda}((p_0, p_0), \mathcal{W})] = 0$ for any $t > 0$, so we get

$$\begin{aligned}
\mathbb{C}_{2t}\left[\sup_{p \in \mathcal{P}} \left\{\Lambda(p, \mathcal{W}) - \Lambda\big(q_p^*, \mathcal{W}\big)\right\}\right] &= \mathbb{C}_{2t}\left[\sup_{\kappa \in \mathcal{K}} \tilde{\Lambda}(\kappa, \mathcal{W})\right] \\
&\le 4S \int_{2\delta}^{2\epsilon} \max\left\{\sqrt{2 \ln \mathcal{N}_{\tilde{\psi}}(z, \mathcal{K})},\, 4tSz^2/\delta\right\} dz + 8\delta\, \mathbb{C}_{4t}[\tau(\mathcal{W})]\,.
\end{aligned} \tag{3.14}$$

[6] To prove the triangle inequality, use $\min\{a+b, c\} \le \min\{a, c\} + \min\{b, c\}$ for $a, b, c \ge 0$.

It remains to bound the entropy of $(\mathcal{K}, \tilde{\psi})$. For any $z \in (2\delta, 2\epsilon]$, let $\mathcal{P}_z$ be a z-cover of $\mathcal{P}$ under ψ with minimal cardinality and define $\mathcal{K}_z \doteq \mathcal{P}_z \times \mathcal{P}_\epsilon^*$. Then $\mathcal{K}_z$ is an external z-cover of $\mathcal{K}$ in the metric space $(\mathcal{P} \times \mathcal{P}_\epsilon^*, \tilde{\psi})$, which means that $\mathcal{K}_z$ might not be a subset of $\mathcal{K}$, but for any $\kappa \in \mathcal{K}$ there exists $\hat{\kappa} \in \mathcal{K}_z$ for which $\tilde{\psi}(\kappa, \hat{\kappa}) \leq z$. Then, as $|\mathcal{K}_{z/2}| = |\mathcal{P}_{z/2}| \cdot |\mathcal{P}_\epsilon^*| \leq \mathcal{N}_\psi(z/2, \mathcal{P})^2$, using the relation between internal and external covering numbers (Dudley, 1999, Theorem 1.2.1), we get

$$\sqrt{2 \ln \mathcal{N}_{\tilde{\psi}}(z, \mathcal{K})} \leq \sqrt{2 \ln |\mathcal{K}_{z/2}|} \leq 2 \sqrt{\ln \mathcal{N}_\psi(z/2, \mathcal{P})}\,. \tag{3.15}$$

Then, applying $\mathbb{C}_t[\cdot]$ to (3.11), Lemma B.1d, and plugging in (3.12),(3.14) and (3.15), we get the claim of Lemma 3.6. ■

Chapter 4

Linear least squares regression

Here we provide an analysis for the so-called *linear least squares regression* setting, which uses the squared loss ($\ell = \ell_{\mathrm{sq}}$) and considers ERM estimators over affine hypothesis classes for sub-Gaussian regression problems defined as

$$\mathbb{M}^{B,\sigma,d}_{\mathrm{subgs}}(\mathcal{F}_*) \doteq \big\{\mu \,\big|\, (\boldsymbol{\mathcal{X}}, \mathcal{Y}) \sim \mu,\, \boldsymbol{\mathcal{X}} \in \mathbb{R}^d,\, \mathcal{Y} \in \mathbb{R},\\ \|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|_{\Psi_2} \le B,\, \|\mathcal{Y} - f_{\mu,\mathcal{F}_*}(\boldsymbol{\mathcal{X}})\|_{\Psi_2} \le \sigma\big\}\,,$$

with some sub-Gaussian parameters $B, \sigma > 0$, and an affine reference class $\mathcal{F}_* \subseteq \mathcal{F}_{\mathrm{aff}} \doteq \{\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} + b,\, \boldsymbol{x} \in \mathbb{R}^d\}$.

In this chapter, we use *least squares estimators* (LSEs), that is ERM estimators (3.1) using the squared loss, over some hypothesis class within affine functions $\mathcal{F}_n \subseteq \mathcal{F}_{\mathrm{aff}}$. These LSEs are compared to the best estimate in the constrained class $\mathcal{F}^{L,p}_{\mathrm{aff}} \doteq \{\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} + b : \|\boldsymbol{a}\|_p \le L\} \subset \mathcal{F}_{\mathrm{aff}}$ with some $L > 0$ and $p \in \{1, 2\}$, that is we set $\mathcal{F}_* \doteq \mathcal{F}^{L,p}_{\mathrm{aff}}$ and so $f_* = f_{\mu,\mathcal{F}^{L,p}_{\mathrm{aff}}}$.

As pointed out by Shamir (2015, Section 1), simultaneous scaling of the bounds B and L by scaling the random variable $\boldsymbol{\mathcal{X}}$ as $c\boldsymbol{\mathcal{X}}$ in μ and the weights of affine estimators $\boldsymbol{a}$ as $\boldsymbol{a}/c$ using some $c > 0$ does not change the problem, so we can assume without loss of generality that $B \doteq 1$.

We point out that the linear regression setting is a special case of convex regression (as affine estimates are convex) and includes nonlinear regression scenarios using finite feature expansions. To see this, consider an estimation problem of some d' dimensional regression function over some domain $\mathbb{X}' \in \mathbb{R}^{d'}$ using data $\mathcal{D}'_n \doteq \{(\boldsymbol{\mathcal{X}}'_i, \mathcal{Y}_i) : i = 1, \ldots, n\} \subseteq (\mathbb{X}' \times \mathbb{Y})^n$. The data set $\mathcal{D}'_n$ can be transformed by a feature map Φ as $\mathcal{D}_n \doteq \{(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) : \boldsymbol{\mathcal{X}}_i = \Phi(\boldsymbol{\mathcal{X}}'_i),\, i = 1, \ldots, n\}$,

where the components of $\Phi : \mathbb{R}^{d'} \to \mathbb{R}^d$ are given as $\Phi(\boldsymbol{x}') = [\phi_1(\boldsymbol{x}') \ \dots \ \phi_d(\boldsymbol{x}')]^\top$ with some $\phi_j : \mathbb{R}^{d'} \to \mathbb{R}$ function for all $j = 1, \dots, d$, which reduces the nonlinear estimation problem to linear regression. However, to make the analysis of this chapter work for these cases, the feature map Φ has to be independent of the sample $\mathcal{D}'_n$, which guarantees that the random elements $(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i)$ of $\mathcal{D}_n$ remain independent. Because of this, linear regression is often referred as *parametric regression* emphasizing that the model parameter Φ is fixed and not adaptive to the sample.

Notice that the estimates $f'_n(\boldsymbol{x}') \doteq \Phi(\boldsymbol{x}')^\top \boldsymbol{a}$ over the original domain $\mathbb{X}'$ remain convex if ϕ_j is convex for all $j = 1, \dots, d$ and the slope parameters $\boldsymbol{a} = [a_1 \ \dots \ a_d]^\top \in \mathbb{R}^d$ of the linear estimates $f_n(\boldsymbol{x}) \doteq \boldsymbol{x}^\top \boldsymbol{a} = f'_n(\boldsymbol{x}')$ defined over the feature space $\mathbb{X} \doteq \{\boldsymbol{x} = \Phi(\boldsymbol{x}') : \boldsymbol{x}' \in \mathbb{X}'\}$ are restricted to be nonnegative on coordinates $j \in \{1, \dots, d\}$, where the feature map ϕ_j is nonlinear. As such constraint (similar to $\boldsymbol{a} \geq \boldsymbol{0}$) would barely decrease the entropy of the class $\mathcal{F}_{\text{aff}}^{L,p}$, the bounds of this chapter are valid for *convex parametric regression* settings as well.

Finally, it is known that some regularization is needed in order to ensure the finiteness of the excess risk for linear regression settings (Huang and Szepesvári, 2014, Example 3.5). We address this by considering two regularization schemes, lasso (Section 4.2.1) and ridge regression (Section 4.2.2), both are being widely used in practice.

4.1 Lower bound on the minimax rate

First, we present two examples to derive a lower bound on the minimax rate for sub-Gaussian linear regression problems $\mathbb{M}_{\text{subgs}}^{1,\sigma,d}(\mathcal{F}_{\text{aff}}^{L,p})$. Both examples, presented in the following sections, describe realizable settings when the regression function lies within the hypothesis class $\mathcal{F}_{\text{aff}}^{L,p}$. The combination of these examples, (4.1) and (4.2), provides the following result:

Theorem 4.1. *For all* $n \geq 57 \cdot 2^{4/d} d^3 \sigma^2 / L^2$,

$$\mathcal{R}_n\left(\mathbb{M}_{subgs}^{1,\sigma,d}(\mathcal{F}_{aff}^{L,p}), \ell_{sq}, \mathcal{F}_{aff}^{L,p}\right) \geq \frac{\max\{L^2/4, (2/9)d\sigma^2\}}{n}\,.$$

Theorem 4.1 is comparable to Theorem 1 of Shamir (2015) which is valid for σ-bounded noise settings (so slightly different than the unbounded Gaussian example in Section 4.1.1), expected excess risk (so being weaker than our high-probability lower bound), and any sample size n, but requires an extra condition as $L \geq 2\sigma$.

4.1.1 Gaussian example

Let $\boldsymbol{\mathcal{X}} \sim P_{\mathbb{X}}$ be a d-dimensional, centered random variable with independent Gaussian coordinates having variance $\hat{B}^2 \doteq 1/(4d) > 0$, so we have $\mathbb{E}\big[\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^\top\big] = \hat{B}^2 I_d$ and $\|\boldsymbol{\mathcal{X}}\|_{\Psi_2} \leq 1$. Consider the set of Gaussian problems $\mathbb{M}^{\sigma}_{\mathrm{gs}}(\mathcal{F}_L, P_{\mathbb{X}})\big)$ as for Theorem 3.1 with $\mathcal{F}^{L,p}_{\mathrm{lin}} \doteq \{\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} : \|\boldsymbol{a}\|_p \leq L\} \subset \mathcal{F}^{L,p}_{\mathrm{aff}}$ and $f_* \in \mathcal{F}^{L,p}_{\mathrm{lin}}$. Then, notice that $\mathbb{M}^{\sigma}_{\mathrm{gs}}(\mathcal{F}^{L,p}_{\mathrm{lin}}, P_{\mathbb{X}}) \subset \mathbb{M}^{1,\sigma,d}_{\mathrm{subgs}}(\mathcal{F}^{L,p}_{\mathrm{aff}})$.

For any $f, g \in \mathcal{F}_{\mathrm{aff}}$, represented by $f(\boldsymbol{x}) \doteq \boldsymbol{a}_f^\top \boldsymbol{x}$ and $g(\boldsymbol{x}) \doteq \boldsymbol{a}_g^\top \boldsymbol{x}$, we have

$$\begin{aligned}\|f - g\|_{P_{\mathbb{X}}} &= \mathbb{E}\big[|f(\boldsymbol{\mathcal{X}}) - g(\boldsymbol{\mathcal{X}})|^2\big]^{1/2} \\ &= \mathbb{E}\big[(\boldsymbol{a}_f - \boldsymbol{a}_g)^\top \boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^\top (\boldsymbol{a}_f - \boldsymbol{a}_g)\big]^{1/2} = \hat{B}\,\|\boldsymbol{a}_f - \boldsymbol{a}_g\|\,,\end{aligned}$$

which implies that $\mathcal{H}_{P_{\mathbb{X}}}(\epsilon, \mathcal{F}^{L,p}_{\mathrm{lin}}) = \mathcal{H}_2\big(\epsilon/\hat{B}, \mathcal{P}\big)$ with $\mathcal{P} \doteq \{\boldsymbol{a} \in \mathbb{R}^d : \|\boldsymbol{a}\|_p \leq L\}$. Similarly, $\|f\|_{P_{\mathbb{X}}} \leq \epsilon_*$ holds if and only if $\|\boldsymbol{a}_f\| \leq \epsilon_*/\hat{B}$. Hence, by Lemma C.1 and $\|\cdot\| \leq \|\cdot\|_p$, we get $\mathcal{H}^*_{P_{\mathbb{X}}}(\epsilon, \epsilon_*, \mathcal{F}^{L,p}_{\mathrm{aff}}) = \mathcal{H}^*_{\|\cdot\|}(\epsilon/\hat{B}, \epsilon_*/\hat{B}, \mathcal{P}) \leq d \ln(3\epsilon_*/\epsilon)$ for all $\epsilon \in (0, 3\epsilon_*]$ and $\epsilon_* > 0$. Further, by Lemma C.1 and $\|\cdot\|_p \leq \sqrt{d}\,\|\cdot\|$, we have $d \ln(d^{-1/2}\hat{B}L/\epsilon) \leq \mathcal{H}_{\|\cdot\|}(\epsilon/\hat{B}, \mathcal{P}) = \mathcal{H}_{P_{\mathbb{X}}}(\epsilon, \mathcal{F}^{L,p}_{\mathrm{aff}})$ for all $\epsilon > 0$. Then by Theorem 3.1 (with $c_0 \leftarrow \infty$, $c_1 \leftarrow d^{-1/2}\hat{B}L$, $c_2 \leftarrow 3$), we get for all $n \geq 57 \cdot 2^{4/d} d^3\, \sigma^4/L^2$ that

$$\mathcal{R}_n\Big(\mathbb{M}^{\sigma}_{\mathrm{gs}}(\mathcal{F}^{L,p}_{\mathrm{aff}}, P_{\mathbb{X}}), \ell_{\mathrm{sq}}, \mathcal{F}^{L,p}_{\mathrm{aff}}\Big) \geq \frac{2d\,\sigma^2}{9n}\,. \tag{4.1}$$

4.1.2 Bernoulli example

Let $\boldsymbol{u} \in \mathbb{R}^d$ be an arbitrary unit vector such that $\|\boldsymbol{u}\|_1 = \|\boldsymbol{u}\| = 1$ holds, and consider the following regression problem class:

$$\begin{aligned}\mathbb{M}^{L}_{\mathrm{bni}} \doteq \Big\{\mu \,\Big|\, &(\boldsymbol{\mathcal{X}}, \mathcal{Y}) \sim \mu,\ f_*(\boldsymbol{x}) = a_* \boldsymbol{u}^\top \boldsymbol{x},\ a_* \in \{-L, L\}, \\ &\mathcal{Y} = f_*(\boldsymbol{\mathcal{X}}),\ \boldsymbol{x}_0 = \boldsymbol{0},\ \boldsymbol{x}_1 = \boldsymbol{u}/\sqrt{2}, \\ &\mathbb{P}\{\boldsymbol{\mathcal{X}} = \boldsymbol{x}_0\} = 1 - r,\ \mathbb{P}\{\boldsymbol{\mathcal{X}} = \boldsymbol{x}_1\} = r,\ r \in (0, 1)\Big\}\,.\end{aligned}$$

As these problems are noiseless, $\|a_* \boldsymbol{u}\| = a_* \|\boldsymbol{u}\|_p = L$ by the choice of $\boldsymbol{u}$, $\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|^2 \leq \|\boldsymbol{\mathcal{X}}\|^2 \leq 1/2$ a.s. implies $\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|_{\Psi_2} \leq 1$, and $f_* \in \mathcal{F}_{\text{aff}}^{L,p}$, we have $\mathbb{M}_{\text{bni}}^{L} \subset \mathbb{M}_{\text{subgs}}^{1,\sigma,d}(\mathcal{F}_{\text{aff}}^{L,p})$.

Clearly, as the problems in $\mathbb{M}_{\text{bni}}^{L}$ are noiseless, an optimal estimator can fit f_* perfectly when the sample contains both points $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$. However, having only $\boldsymbol{x}_0$ present in the sample would make any estimate $f_n \in \mathcal{F}_{\text{aff}}^{L,p}$ choosing the value $f_n(\boldsymbol{x}_1)$ without any hint. So for this case, by the symmetry of $\{-L, L\}$, the best estimate regarding the minimax error is $f_n = 0$, as shown on Figure 4.1.

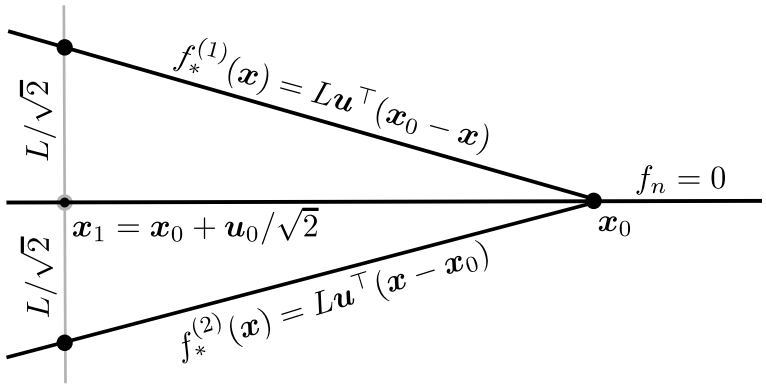

Figure 4.1: Worst case Bernoulli example of estimating a linear function. Without seeing the value of f_* at $\boldsymbol{x}_1$, no estimate f_n can decide between the cases $f_* = f_*^{(1)}$ and $f_* = f_*^{(2)}$. Then the best choice regarding the minimax error is $f_n(\boldsymbol{x}_1) = f_n(\boldsymbol{x}_0) = 0$ and so $f_n = 0$.

Hence, we can lower bound the excess risk for this example as

$$\sup_{\mu \in \mathbb{M}_{\text{bni}}^{L}} L_\mu(f_n, f_*) \geq \sup_{\substack{a_* \in \{-L, L\} \\ r \in (0,1)}} r \left|(\boldsymbol{0} - a_* \boldsymbol{u})^\top \boldsymbol{u}\right|^2 \mathbb{I}\{E_n\} = \sup_{r \in (0,1)} r \, \frac{L^2}{2} \, \mathbb{I}\{E_n\} \,,$$

where E_n denotes the event $E_n \doteq \left\{\boldsymbol{\mathcal{X}}_1 = \cdots = \boldsymbol{\mathcal{X}}_n = \boldsymbol{x}_0\right\}$. Then, by choosing $r \doteq 1 - 2^{-1/n}$ to satisfy $\mathbb{P}\{E_n\} = (1-r)^n = 1/2$, and using $r \geq 1/(2n)$ due to Lemma F.1, we get

$$\mathcal{R}_n\big(\mathbb{M}_{\text{bni}}^{L}, \ell_{\text{sq}}, \mathcal{F}_{\text{aff}}^{L,p}\big) \geq \frac{L^2}{4n} \,. \tag{4.2}$$

Finally, we point out that ERM upper bounds scaling linearly with the kurtosis bound $\sup_{f \in \mathcal{F}} \mathbb{K}_0[\mathcal{W}_f]$ (including Theorem A of Lecué and Mendelson

2013, and Theorem 7 of Liang et al. 2015) do not provide a near-minimax rate for the class $\mathbb{M}_{\text{bni}}^{L}$. To see this, use Lemma F.1 to get

$$\sup_{f \in \mathcal{F}_{\text{aff}}^{L,p}} \mathbb{K}_0[\mathcal{W}_f] \geq \mathbb{K}_0\big[0 - a_* \boldsymbol{u}^\top \boldsymbol{\mathcal{X}}\big] = \frac{r\,(a_*/\sqrt{2})^4}{\big(r\,(a_*/\sqrt{2})^2\big)^2} = \frac{1}{r} \geq \frac{n}{\ln 2}\,.$$

Hence, no result scaling polynomially as a function of $\sup_{f \in \mathcal{F}} \mathbb{K}_0[\mathcal{W}_f]$ can deliver a near-minimax rate for these problems. Fortunately, the logarithmic dependence on the kurtosis bound of Lemma 3.3 will allow us to derive near-minimax ERM upper bounds using Theorem 3.2 for the entire sub-Gaussian problem class $\mathbb{M}_{\text{subgs}}^{1,\sigma,d}(\mathcal{F}_{\text{aff}}^{L,p})$.

4.2 Near-minimax upper bounds for LSEs

In this section we derive upper bounds on the excess risk for (α, β)-ERM$(\mathcal{F}_{\text{aff}}^{L,p})$ estimators (3.1) on sub-Gaussian regression problems $\mathbb{M}_{\text{subgs}}^{1,\sigma,d}(\mathcal{F}_{\text{aff}}^{L,p})$ with squared loss ($\ell = \ell_{\text{sq}}$) using the uniformly L-Lipschitz affine class $\mathcal{F}_* = \mathcal{F}_{\text{aff}}^{L,p}$ and the reference function $f_* = f_{\mu,\mathcal{F}_{\text{aff}}^{L,p}}$. The result (Lemma 4.2) is applied to two practical scenarios, ERM with explicit Lipschitz constraint (lasso) in Section 4.2.1, and $\|\cdot\|^2$-penalized ERM$(\mathcal{F}_{\text{aff}}^{L,p})$ estimation (ridge regression) in Section 4.2.2.

For the analysis, we only consider penalty functions $\beta : \mathcal{F}_{\text{aff}} \to \mathbb{R}_{\geq 0}$, which are independent of the estimate's bias term, so satisfy $\frac{\partial}{\partial b}\beta(\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} + b) = 0$. Now introduce the following linear function classes:

$$\mathcal{F}_{\text{aff}}^{L,p,\mu}(t) \doteq \big\{\boldsymbol{x} \mapsto \boldsymbol{a}^\top(\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + b : \|\boldsymbol{a}\|_p \leq L,\, b - \mathbb{E}\mathcal{Y} \in [-t, t]\big\}\,,$$
$$\mathcal{F}_{\text{aff}}^{L,p,n}(t) \doteq \big\{\boldsymbol{x} \mapsto \boldsymbol{a}^\top(\boldsymbol{x} - \bar{\boldsymbol{\mathcal{X}}}) + b : \|\boldsymbol{a}\|_p \leq L,\, b - \bar{\mathcal{Y}} \in [-t, t]\big\}\,,$$

where $t \in \mathbb{R}_{\geq 0} \cup \{\infty\}$, $\bar{\boldsymbol{\mathcal{X}}} \doteq \frac{1}{n}\sum_{i=1}^n \boldsymbol{\mathcal{X}}_i$ and $\bar{\mathcal{Y}} \doteq \frac{1}{n}\sum_{i=1}^n \mathcal{Y}_i$. For convenience, we also use $\mathcal{F}_{\text{aff}}^{L,p,n} \doteq \mathcal{F}_{\text{aff}}^{L,p,n}(0)$. Observe that $\mathcal{F}_{\text{aff}}^{L,p,n}(t)$ is a data-dependent approximation to $\mathcal{F}_{\text{aff}}^{L,p,\mu}(t)$, and using the bias independence of the penalty term β, we have

$$\mathbb{E}[\mathcal{Y} - \boldsymbol{a}^\top \boldsymbol{\mathcal{X}}] = \operatorname*{argmin}_{b \in \mathbb{R}} \mathbb{E}\Big[|\boldsymbol{a}^\top \boldsymbol{\mathcal{X}} + b - \mathcal{Y}|^2 + \beta(\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} + b)\Big]\,,$$

so $f_* = f_{\mu,\mathcal{F}_{\text{aff}}^{L,p}} = f_{\mu,\mathcal{F}_{\text{aff}}^{L,p,\mu}(t)}$ holds for any $t \geq 0$. For a similar reason, an (α, β)-ERM$(\mathcal{F}_{\text{aff}}^{L,p,n}(t))$ estimator is also (α, β)-ERM$(\mathcal{F}_{\text{aff}}^{L,p,n})$ for any $t \geq 0$.

Because distribution μ is unknown, estimators cannot be represented by the class $\mathcal{F}_{\text{aff}}^{L,p,\mu}(t)$, just by its data-dependent approximation $\mathcal{F}_{\text{aff}}^{L,p,n}(t)$. However, as the quantities $\mathbb{E}\boldsymbol{\mathcal{X}}$ and $\mathbb{E}\mathcal{Y}$ are "well-approximated" by $\bar{\boldsymbol{\mathcal{X}}}$ and $\bar{\mathcal{Y}}$ for sub-Gaussian random variables $\boldsymbol{\mathcal{X}}$ and $\mathcal{Y}$, the function classes $\mathcal{F}_{\text{aff}}^{L,p,\mu}(t)$ and $\mathcal{F}_{\text{aff}}^{L,p,n}(t)$ are "close" with high-probability. More precisely, it is possible to show (see the proof of Lemma 4.2) that $\mathbb{P}\{\hat{\mathcal{F}}_n \subseteq \mathcal{F}_n \subseteq \mathcal{F}\} \geq 1-\gamma/2$ holds for function classes $\hat{\mathcal{F}}_n \doteq \mathcal{F}_{\text{aff}}^{L,p,\mu}(0)$, $\mathcal{F}_n \doteq \mathcal{F}_{\text{aff}}^{L,p,n}(t_n)$, and $\mathcal{F} \doteq \mathcal{F}_{\text{aff}}^{L,p,\mu}(2t_n)$, where $t_n \doteq \Theta\big(\max\{L,\sigma\}\sqrt{\ln(1/\gamma)/n}\big)$.

Hence, our general regression result (Theorem 3.2) is applicable for the function sets $\hat{\mathcal{F}}_n$, $\mathcal{F}_n$, $\mathcal{F}$, and provides an upper bound for the (α,β)-ERM$(\mathcal{F}_n)$ class which is satisfied by (α,β)-ERM$(\mathcal{F}_{\text{aff}}^{L,p,n})$ estimates. For this, we still have to show that Conditions (C1) and (C3) hold,[1] for which we provide the details by the following result:

Lemma 4.2. *Consider any sub-Gaussian problem $\mu \in \mathbb{M}_{subgs}^{1,\sigma,d}(\mathcal{F}_{aff}^{L,p})$ with the squared loss ($\ell = \ell_{sq}$), $f_* = f_{\mu,\mathcal{F}_{aff}^{L,p}}$, and an (α,β)-ERM$(\mathcal{F}_{aff}^{L,p,n})$ estimate f_n with penalty term $\beta : \mathcal{F}_{aff} \to \mathbb{R}_{\geq 0}$ satisfying $\frac{\partial}{\partial b}\beta(\boldsymbol{x} \mapsto \boldsymbol{a}^\top\boldsymbol{x}+b) = 0$ and $\mathbb{P}\big\{\beta(f_*)+\alpha > B_*\big\} \leq \gamma/4$. Then for all $\gamma > 0$ and $n \geq d\ln(1/\gamma)$, we have with probability at least $1-\gamma$ that*

$$L_\mu(f_n, f_*) = O\bigg(d\theta\,\frac{\ln\big(n/(d\gamma)\big)}{n} + B_*\bigg),$$

where $\theta = \Omega\big(Q_n \max\{L,\sigma\}^2\big)$ and $Q_n = O(\ln(n))$ as defined for Lemma 3.3.

Proof. We prove the claim by applying Theorem 3.2 for the function classes $\hat{\mathcal{F}}_n$, $\mathcal{F}_n$, and $\mathcal{F}$ as defined above, and using our previous observations that $f_* \in \hat{\mathcal{F}}_n$ and f_n is also (α,β)-ERM$(\mathcal{F}_n)$.

To prove the required high-probability relationship of $\hat{\mathcal{F}}_n$, $\mathcal{F}_n$, and $\mathcal{F}$, write the optimal estimator as $f_*(\boldsymbol{x}) = \boldsymbol{a}_*^\top(\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \mathbb{E}\mathcal{Y}$ with some $\boldsymbol{a}_* \in \mathbb{R}^d$ satisfying $\|\boldsymbol{a}_*\|_p \leq L$. Next, apply Hölder's inequality with $\|\cdot\|_q \leq \|\cdot\|$ for $q \doteq p/(1-p) \in \{2,\infty\}$ to obtain

$$\begin{aligned} |\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| &= \big|\bar{\mathcal{Y}} - f_*(\bar{\boldsymbol{\mathcal{X}}}) + \boldsymbol{a}_*^\top(\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}})\big| \\ &\leq |\bar{\mathcal{Y}} - f_*(\bar{\boldsymbol{\mathcal{X}}})| + L\|\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|\,. \end{aligned} \tag{4.3}$$

[1] Condition (C2) will be ignored by setting $\epsilon \doteq \delta$ and $S \doteq \infty$.

Furthermore, by (4.3), Lemma A.2d and Lemma A.3, we also have

$$\big\|L\|\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + |\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}|\big\|_{\Psi_2} \le 2L\|\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|_{\Psi_2} + \|\bar{\mathcal{Y}} - f_*(\bar{\boldsymbol{\mathcal{X}}})\|_{\Psi_2} \le t_0\,, \tag{4.4}$$

with $t_0 \doteq 3\max\{L,\sigma\}\sqrt{8d/n}$. Then, by the definition of $\hat{\mathcal{F}}_n, \mathcal{F}_n, \mathcal{F}$ and Lemma A.2a with (4.4), we get for any $\gamma > 0$, and $t_n \doteq t_0\sqrt{\ln(4/\gamma)}$ that

$$\begin{aligned}\mathbb{P}\big\{\hat{\mathcal{F}}_n \subseteq \mathcal{F}_n \subseteq \mathcal{F}\big\} &\ge 1 - \mathbb{P}\big\{L\|\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + |\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| > t_n\big\}\\ &\ge 1 - 2e^{-t_n^2/t_0^2} = 1 - \gamma/2\,.\end{aligned}$$

Write $f, f_* \in \mathcal{F}$ as $f(\boldsymbol{x}) \doteq \boldsymbol{a}^\top(\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + b$ and $f_*(\boldsymbol{x}) \doteq \boldsymbol{a}_*^\top(\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \hat{b}$. Observe that by Hölder's inequality and $\|\cdot\|_q \le \|\cdot\|$, we have

$$|\mathcal{W}_f| = |f(\boldsymbol{\mathcal{X}}) - f_*(\boldsymbol{\mathcal{X}})| = |(\boldsymbol{a} - \boldsymbol{a}_*)^\top(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}})| \le 2L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|\,.$$

To verify Condition (C3), pick $f \in \mathcal{F}$ arbitrarily, and use $(a+b)^2 \le 2a^2 + 2b^2$ with Jensen's inequality, to show that $\mathcal{W}_f$ is sub-Gaussian, that is

$$\mathbb{E}\big[e^{\mathcal{W}_f^2/B_n^2}\big] \le \mathbb{E}\big[e^{8L^2\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|^2/B_n^2}\big]\, e^{8t_n^2/B_n^2} \le 2\,,$$

with $B_n \doteq 4\max\{L, t_n\}$, where t_n is set as given above for the definition of $\mathcal{F}_n$. Further, as $f_* \in \hat{\mathcal{F}}_n \subseteq \mathcal{F}$ and the function set $\mathcal{F}$ is convex, the requirements of Lemma 3.5 are satisfied, and we get Condition (C3) with $r_0 = 6\max\{B_n, \sigma\}^2$, $r = 1/2$, and any $\theta \ge 240 Q_n \max\big\{B_n, \sigma\big\}^2$, where $Q_n \in [\ln 2, \ln n]$ as defined for Lemma 3.3.

Next, to show Condition (C1), write $g \in \mathcal{F}$ as $g(\boldsymbol{x}) \doteq \hat{\boldsymbol{a}}^\top(\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \hat{b}$. Using $|\mathcal{Y} - \mathbb{E}\mathcal{Y}| \le L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + |\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})|$ similar to (4.3), and Hölder's inequality with $\|\cdot\|_q \le \|\cdot\|$ again, we obtain

$$\begin{aligned}\mathcal{Z}(f,g) &= \big(f(\boldsymbol{\mathcal{X}}) - g(\boldsymbol{\mathcal{X}})\big)\big(f(\boldsymbol{\mathcal{X}}) + g(\boldsymbol{\mathcal{X}}) - 2\mathcal{Y}\big)\\ &= \big((\boldsymbol{a} + \hat{\boldsymbol{a}})^\top(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}) + b + \hat{b} - 2\mathcal{Y}\big)\big((\boldsymbol{a} - \hat{\boldsymbol{a}})^\top(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}) + b - \hat{b}\big)\\ &\le \Big(2L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + 4t_n + 2|\mathcal{Y} - \mathbb{E}\mathcal{Y}|\Big)^2\psi(f,g)\\ &\le \underbrace{\Big(3L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + 4t_n + 2|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})|\Big)^2}_{\doteq G(\boldsymbol{\mathcal{X}},\mathcal{Y})}\psi(f,g)\,,\end{aligned}$$

where $\psi(f,g) \doteq \sqrt{\|\boldsymbol{a} - \hat{\boldsymbol{a}}\|^2/(4L^2) + |b - \hat{b}|^2/(4t_n)^2}$ is a metric on $\mathcal{F}$. As the radius of $\mathcal{F}$ under ψ is bounded by $5/4$, that is $\sup_{f\in\mathcal{F}}\psi(f,0) \le 5/4$,

we have by Lemma C.1 that $\mathcal{H}_\psi(\epsilon, \mathcal{F}) \le (d+1)\ln\big(4/\epsilon\big)$ for all $\epsilon \in (0,4]$. Furthermore, as $\|G(\boldsymbol{\mathcal{X}}, \mathcal{Y})\|_{\Psi_2} \le 11 t_n$, Lemma B.4 implies $\mathbb{C}_{\frac{1}{\theta}}\big[G(\boldsymbol{\mathcal{X}}, \mathcal{Y})\big] \le \theta$ for any $\theta \ge (11 t_n)^2$.

Finally, we can apply Theorem 3.2 with $\epsilon \doteq \delta$ ignoring Condition (C2) with $S = \infty$, choosing $\delta \doteq d/n$, to get with probability at least $1-\gamma$ that

$$\begin{aligned} L_\mu(f_n, f_*) &\le 2\bigg(\frac{\theta \mathcal{H}_\psi(\epsilon, \mathcal{P})}{n} + 16\epsilon\, \mathbb{C}_{\frac{1}{\theta}}[G(\boldsymbol{\mathcal{X}}, \mathcal{Y})] + B_*\bigg) + \frac{r_0 + 4\theta \ln(4/\gamma)}{n} \\ &\le \frac{2\theta}{n}\Big((d+1)\ln(4n/d) + 16d + 1 + 2\ln(4/\gamma)\Big) + 2B_*\,, \end{aligned}$$

which proves the claim with $\theta = \Omega(Q_n\, t_n^2) = \Theta\big(Q_n \max\{L,\sigma\}^2\big)$ thanks to $n \ge d\ln(1/\gamma)$. ■

4.2.1 The lasso

Here we specialize Lemma 4.2 to train an affine LSE with $\|\cdot\|_p$-bounded slope without using any penalty term ($\beta \doteq 0$). For $p = 1$, this technique is called *lasso* (Tibshirani, 1996, 2011). The result is the following:

Theorem 4.3. *Consider any sub-Gaussian problem $\mu \in \mathbb{M}^{1,\sigma,d}_{subgs}(\mathcal{F}^{L,p}_{aff})$ with the squared loss ($\ell = \ell_{sq}$), $\mathcal{F}_* = \mathcal{F}^{L,p}_{aff}$, $f_* = f_{\mu,\mathcal{F}^{L,p}_{aff}}$, and an α-ERM($\mathcal{F}^{L,p,n}_{aff}$) estimate f_n with any α having $\mathbb{P}\big\{\alpha = \Omega\big(d\max\{L,\sigma\}^2 \ln(n/(d\gamma))/n\big)\big\} \le \gamma/4$.*[2] *Then for all $\gamma > 0$ and $n \ge d\ln(1/\gamma)$, we have with probability at least $1-\gamma$ that*

$$L_\mu(f_n, f_*) = O\bigg(dQ_n \max\{L,\sigma\}^2 \frac{\ln\big(n/(d\gamma)\big)}{n}\bigg),$$

where $Q_n = O(\ln(n))$ as defined for Lemma 3.3.

Proof. The claim follows directly from Lemma 4.2 by using the zero penalty function $\beta = 0$ to obtain $\mathbb{P}\{\beta(f_*) + \alpha > B_*\} = \mathbb{P}\{\alpha > B_*\} \le \gamma/4$ for some $B_* = \Theta\big(d\max\{L,\sigma\}^2 \ln(n/(d\gamma))/n\big)$.

■

Notice that Theorem 4.3 provides the $O(\ln(n)/n)$ rate for any regression problem in $\mathbb{M}^{1,\sigma,d}_{\text{subgs}}(\mathcal{F}^{L,p}_{\text{aff}})$ regardless of the magnitude of the kurtosis bound

[2] For example, $\alpha = \frac{1}{n}\big(\frac{1}{n}\sum_{i=1}^n |\mathcal{Y}_i - \bar{\mathcal{Y}}|^2\big)$ works.

$\sup_{f \in \mathcal{F}_{\text{aff}}^{L,p}} \mathbb{K}_0[\mathcal{W}_f]$, which is not true for Theorem A of Lecué and Mendelson (2013) applied to the linear class $\mathcal{F}_{\text{aff}}^{L,p}$ (see Section 4.1.2 for an example).

Furthermore, for $p = 2$, the bound of Theorem 4.3 is comparable to the conjecture of Shamir (2015) stating that ERM estimates achieve optimal expected excess risk up to logarithmic factors. Here, our bound is slightly weaker than this by scaling with $d \max\{L, \sigma\}^2$ instead of $\max\{L^2, d\sigma^2\}$.

4.2.2 Ridge regression

Now we consider replacing the Lipschitz constraint to the quadratic penalty $\beta_\lambda(\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} + b) \doteq \lambda \|\boldsymbol{a}\|^2$, and consider the *ridge regression* (Hoerl and Kennard, 1970) estimate f_n^λ satisfying the $(0, \beta_\lambda)$-ERM$(\mathcal{F}_{\text{aff}})$ property (3.1). This can be computed in closed-form as $f_n^\lambda(\boldsymbol{x}) \doteq \boldsymbol{a}_n^\top(\boldsymbol{x} - \bar{\boldsymbol{\mathcal{X}}}) + \bar{\mathcal{Y}}$ with

$$\boldsymbol{a}_n \doteq \left(\lambda I_d + \frac{1}{n}\sum_{i=1}^n (\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mathcal{X}}})(\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mathcal{X}}})^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mathcal{X}}})(\mathcal{Y}_i - \bar{\mathcal{Y}})\right), \quad (4.5)$$

where I_d denotes the $d \times d$ identity matrix.

We only analyze estimate f_n^λ for problems $\mu \in \mathbb{M}_{\text{subgs}}^{1,\sigma,d}(\mathcal{F}_{\text{aff}})$ for which the reference function $f_* = f_{\mu,\mathcal{F}_{\text{aff}}}$, written as $f_*(\boldsymbol{x}) \doteq \boldsymbol{a}_*^\top(\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \mathbb{E}\mathcal{Y}$, satisfies $\|\boldsymbol{a}_*\| \leq L$ for some $L > 0$. For such settings, we use $\lambda \doteq \frac{\lambda_0}{4nL^2}\sum_{i=1}^n |\mathcal{Y}_i - \bar{\mathcal{Y}}|^2$ with some $\lambda_0 \in (0, 1]$, so Lemma F.2 with (4.5) bounds the Lipschitz factor of the ridge regression estimate f_n^λ as $\|\boldsymbol{a}_n\| \leq \sqrt{\frac{1}{4n}\sum_{i=1}^n |\mathcal{Y}_i - \bar{\mathcal{Y}}|^2/\lambda} = L/\sqrt{\lambda_0}$. Then we apply Lemma 4.2 with the common Lispchitz bound on f_n^λ and f_* given as $\max\{\|\boldsymbol{a}_n\|, \|\boldsymbol{a}_*\|\} \leq L/\sqrt{\lambda_0}$, and obtain the following result:

Theorem 4.4. *Consider any distribution $\mu \in \mathbb{M}_{subgs}^{1,\sigma,d}(\mathcal{F}_{aff})$ such that $\|\boldsymbol{a}_*\| \leq L$, the squared loss ($\ell = \ell_{sq}$), $\mathcal{F}_* = \mathcal{F}_{aff}$, $f_* = f_{\mu,\mathcal{F}_{aff}}$, and a $(0, \beta_\lambda)$-ERM$(\mathcal{F}_{aff})$ estimate f_n with $\lambda = \frac{\lambda_0}{4nL^2}\sum_{i=1}^n |\mathcal{Y}_i - \bar{\mathcal{Y}}|^2$ and some $\lambda_0 \in (0, 1]$. Then for all $\gamma > 0$ and $n \geq d\ln(1/\gamma)$, we have with probability at least $1 - \gamma$ that*

$$\begin{aligned} L_\mu(f_n, f_*) &= O\left(\left(\frac{d\ln(n)}{n}\max\left\{L^2/\lambda_0, \sigma^2\right\} + \lambda_0 \max\{L, \sigma\}^2\right)\ln\left(n/(d\gamma)\right)\right) \\ &= O\left(\ln(n)\max\{L, \sigma\}^2 \ln\left(n/(d\gamma)\right)\sqrt{d/n}\right), \end{aligned}$$

where $\lambda_0 = \sqrt{d/n}$ for the second result.

Proof. In order to apply Lemma 4.2, we first need to upper bound $\beta_\lambda(f_*)$. For this, set $B_* \doteq \lambda_0 \kappa \ln(8/\gamma)$ with any $\kappa \geq 16 \max\{L, \sigma\}^2$. Then, by using $\beta_\lambda(f_*) \leq \lambda L^2$, Markov's inequality, $(a+b)^2 \leq 2a^2 + 2b^2$, Jensen's inequality twice, $|\mathcal{Y} - \mathbb{E}\mathcal{Y}|^2 \leq 2|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})|^2 + 2L^2 \|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|^2$ similar to (4.3), and the Cauchy-Schwartz inequality, we get

$$\begin{aligned}
\mathbb{P}\big\{\beta_\lambda(f_*) > B_*\big\} &\leq \mathbb{P}\Big\{\frac{1}{n\kappa}\sum_{i=1}^n |\mathcal{Y}_i - \bar{\mathcal{Y}}|^2 > \ln(8/\gamma)\Big\} \\
&\leq \frac{\gamma}{8}\mathbb{E}\Big[e^{\frac{1}{n\kappa}\sum_{i=1}^n |\mathcal{Y}_i - \bar{\mathcal{Y}}|^2}\Big] \leq \frac{\gamma}{8}\mathbb{E}\Big[e^{\frac{2}{n\kappa}\sum_{i=1}^n \{|\mathcal{Y}_i - \mathbb{E}\mathcal{Y}|^2 + |\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}|^2\}}\Big] \\
&\leq \frac{\gamma}{8}\mathbb{E}\Big[e^{\frac{4}{\kappa}|\mathcal{Y} - \mathbb{E}\mathcal{Y}|^2}\Big] \leq \frac{\gamma}{8}\mathbb{E}\Big[e^{\frac{16}{\kappa}|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})|^2}\Big]^{\frac{1}{2}}\mathbb{E}\Big[e^{\frac{16L^2}{\kappa}\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|^2}\Big]^{\frac{1}{2}} \leq \frac{\gamma}{4}\,.
\end{aligned}$$

As β_λ does not depend on the bias term of affine estimators we use Lemma 4.2 with Lipschitz bound $\max\{\|\boldsymbol{a}_*\|, \|\boldsymbol{a}_n\|\} \leq L/\sqrt{\lambda_0}$, error term $\alpha = 0$, and $\mathbb{P}\{\beta_\lambda(f_*) > B_*\} \leq \gamma/4$ to get with probability at least $1 - \gamma$ that

$$L_\mu(f_n, f_*) = O\bigg(d\theta\,\frac{\ln\big(n/(d\gamma)\big)}{n} + B_*\bigg)\,,$$

where $\theta = \Omega\big(Q_n \max\big\{\max\{L\lambda_0^{-1/2}\}, \sigma\big\}^2\big)$. Finally, we get the claim by $B_* = \Theta\big(\lambda_0 \max\{L, \sigma\}^2 \ln(1/\gamma)\big)$. ∎

With $\lambda_0 = \sqrt{d/n}$, Theorem 4.4 provides a suboptimal $O\big(\ln(n)/\sqrt{n}\big)$ rate for ridge regression. Notice that if we had $\|\boldsymbol{a}_n\| \leq L$ with high-probability, we could improve the rate of Theorem 4.4 to $O(1/n)$ up to logarithmic factors. However, when the feature covariance matrix is well-conditioned, we can use Theorem 4.4 to prove such a Lipschitz bound and improve the rate of f_n^λ.

Formally, suppose that the covariance matrix of the features $\boldsymbol{\mathcal{X}}$ is positive definite $\mathbb{E}\big[(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}})(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}})^\top\big] \succ \eta_\mu I_d$ with smallest eigenvalue $\eta_\mu > 0$. Then, we can use Theorem 4.4 with $\lambda_0 = \frac{d\ln(n)}{n}\ln\big(\frac{n}{d\gamma}\big)$ to get the excess risk bound $L_\mu(f_n^\lambda, f_*) = O(L^2)$ with high-probability for a sufficiently large n satisfying $\frac{d\ln(n)}{n}\max\{L, \sigma\}^2 \ln^2\big(\frac{n}{d\gamma}\big) = O(L^2)$.[3] This improves the Lipschitz bound to $\|\boldsymbol{a}_n\| = O\big((1 + \eta_\mu^{-1/2})L\big)$ with probability at least $1 - \gamma/2$, which does not diverge for $\lambda_0 \to 0$ as $n \to \infty$, so can be used to improve the excess risk bound. The details are provided by the following result:

[3] When n is not large enough, the desired rate follows immediately.

Theorem 4.5. *Consider any distribution $\mu \in \mathbb{M}^{1,\sigma,d}_{subgs}(\mathcal{F}_{aff})$ such that $\|\boldsymbol{a}_*\| \le L$ and $\mathbb{E}\big[(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}})(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}})^\top\big] \succ \eta_\mu I_d$, the squared loss $\ell = \ell_{sq}$, $\mathcal{F}_* = \mathcal{F}_{aff}$, $f_* = f_{\mu,\mathcal{F}_{aff}}$. Take a $(0,\beta_\lambda)$-ERM($\mathcal{F}_{aff}$) estimate f_n with $\lambda = \frac{\lambda_0}{4nL^2}\sum_{i=1}^n |\mathcal{Y}_i - \bar{\mathcal{Y}}|^2$ and set $\lambda_0 = \frac{d\ln(n)}{n}\ln\left(\frac{n}{d\gamma}\right)$ for some $\gamma > 0$. Then for all $n \ge d\ln(2/\gamma)$ such that $\lambda_0 \le 1$, we have with probability at least $1-\gamma$ that*

$$L_\mu(f_n, f_*) = O\Big(d\ln(n)\max\big\{(1+\eta_\mu^{-1/2})L, \sigma\big\}^2 \frac{\ln\big(n/(d\gamma)\big)}{n}\Big).$$

Proof. First, notice that if $\frac{d\ln(n)}{n}\max\{L,\sigma\}^2\ln^2\left(\frac{n}{d\gamma}\right) = \Omega(L^2)$ is satisfied, then Theorem 4.4 immediately provides the claim with $\lambda_0 = d\ln(n)\ln(n/(d\gamma))/n$. Otherwise, we have $\frac{d\ln(n)}{n}\max\{L,\sigma\}^2\ln^2\left(\frac{n}{d\gamma}\right) = O(L^2)$.

Next, use (4.4) with Lemma A.2a to get with probability at least $1-\gamma/2$ that

$$\big|\boldsymbol{a}_*^\top(\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}\big|^2 = O\big(d\max\{L,\sigma\}^2\ln(1/\gamma)/n\big) = O(L^2).$$

Then, either $\mathbb{E}\big[|(\boldsymbol{a}_n - \boldsymbol{a}_*)^\top(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})|^2\big]^{\frac{1}{2}} \le \big|\boldsymbol{a}_*^\top(\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}\big|/4 = O(L)$ holds, or if not, we have

$$\begin{aligned}\mathbb{E}\big[\mathcal{W}^2_{f_n^\lambda}\big] &= \mathbb{E}\big[|(\boldsymbol{a}_n - \boldsymbol{a}_*)^\top(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}}) + \boldsymbol{a}_*^\top(\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}|^2\big] \\ &\ge \mathbb{E}\big[|(\boldsymbol{a}_n - \boldsymbol{a}_*)^\top(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})|^2\big]/2\,.\end{aligned} \tag{4.6}$$

Now use the Bernstein condition (3.3) which is satisfied for the squared loss $\ell = \ell_{\mathrm{sq}}$ and the convex hypothesis class $\mathcal{F}_{\mathrm{aff}}$ with $C = 2$, Theorem 4.4 with $\lambda_0 = d\ln(n)\ln\big(n/(d\gamma)\big)/n$ and $\frac{d\ln(n)}{n}\max\{L,\sigma\}^2\ln^2\left(\frac{n}{d\gamma}\right) = O(L^2)$, and (4.6) to get with probability at least $1-\gamma/2$ that

$$O(L^2) = CL_\mu(f_n^\lambda, f_*) \ge \mathbb{E}\big[\mathcal{W}^2_{f_n^\lambda}\big] \ge \mathbb{E}\big[|(\boldsymbol{a}_n - \boldsymbol{a}_*)^\top(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})|^2\big]/2\,.$$

Hence, in either way, we have $\mathbb{E}\big[|(\boldsymbol{a}_n - \boldsymbol{a}_*)^\top(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})|^2\big] = O(L^2)$ with probability at least $1-\gamma/2$.

Next, observe that

$$\mathbb{E}\big[(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})^\top\big] = \mathbb{E}\big[(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}})(\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}})^\top\big] + (\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}})(\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}})^\top,$$

where the expectations are taken with respect to $\boldsymbol{\mathcal{X}}$ only. So by the positive definiteness of the covariance matrix and the triangle inequality, we have

$$\mathbb{E}\big[|(\boldsymbol{a}_n - \boldsymbol{a}_*)^\top(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})|^2\big] \ge \eta_\mu\|\boldsymbol{a}_n - \boldsymbol{a}_*\|^2 \ge \eta_\mu\big|\,\|\boldsymbol{a}_n\| - \|\boldsymbol{a}_*\|\,\big|^2,$$

implying that $\|\boldsymbol{a}_n\| = O\big((1+\eta_\mu^{-1/2})L\big)$ with probability at least $1-\gamma/2$.

Then, we have $L_\mu(f_n^\lambda, f_*) \le L_\mu(f_n^\lambda, f_*)\mathbb{I}\big\{f_n^\lambda \in \mathcal{F}_{\mathrm{aff}}^{L,2,n}\big\}$ with probability at least $1-\gamma/2$ for Lipschitz bound $L \doteq O\big((1+\eta_\mu^{-1/2})L\big)$, so we get the claim by applying Lemma 4.2 with $\gamma \leftarrow \gamma/2$ and bounding the regularization term β_λ with probability at least $1-\gamma/8$ by $\beta_\lambda(f_*) = \Theta\big(\lambda_0 L^2 \max\{L,\sigma\}^2 \ln(1/\gamma)\big)$ as shown in the proof of Theorem 4.4. ■

Finally, we note that Theorems 4.4 and 4.5 are comparable to the work of Hsu et al. (2014, Remarks 4 and 12). Our results provide the same rates in d and n, and cover the more general sub-Gaussian setting instead of using the boundedness assumption.

Chapter 5

Convex nonparametric least squares regression

In this chapter we apply the general regression analysis of Chapter 3 to convex nonparametric least squares estimation settings. Here we discuss regression problems (Section 3.1) with the squared loss $\ell = \ell_{\mathrm{sq}}$, and set the reference class as $\mathcal{F}_* \doteq \{\mathbb{X} \to \mathbb{R}\}$, so every reference function f_* is also a regression function. Furthermore, we assume that the domain $\mathbb{X} \subseteq \mathbb{R}^d$ is convex, and there exists a convex regression function f_*, that is $f_* \in \mathcal{F}_{\mathrm{cx}} \doteq \{f : \mathbb{X} \to \mathbb{R} \,|\, f \text{ is convex}\}$.

Similar to the linear case (Chapter 4), we need a bound on the slope of f_* and the estimates for the derivation of excess risk bounds (see the discussion in Section 5.1 for more details). To formalize such a slope bound, denote the *subdifferential (subgradient set)* of a convex function $f : \mathbb{X} \to \mathbb{R}$ at $\boldsymbol{x} \in \mathbb{X}$ by

$$\partial f(\boldsymbol{x}) \doteq \left\{\boldsymbol{s} \in \mathbb{R}^d \,|\, \forall \boldsymbol{z} \in \mathbb{X} : f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{s}^\top (\boldsymbol{z} - \boldsymbol{x})\right\}.$$

Then define the smallest *subgradient* of f at $\boldsymbol{x}$ with respect to the Euclidean norm $\|\cdot\|$ by $\nabla_* f(\boldsymbol{x}) \doteq \operatorname{argmin}_{\boldsymbol{s} \in \partial f(\boldsymbol{x})} \|\boldsymbol{s}\|$. As $\partial f(\boldsymbol{x})$ is a closed, convex set (Rockafellar, 1972, Page 215), and $\|\cdot\|$ is strictly convex, $\nabla_* f(\boldsymbol{x})$ is well-defined if f is subdifferentiable at $\boldsymbol{x}$, that is $\partial f(\boldsymbol{x}) \neq \emptyset$, because in this case the minimizer in the definition of $\nabla_* f(\boldsymbol{x})$ exists and is unique. Otherwise, when $\partial f(\boldsymbol{x}) = \emptyset$, we set $\nabla_* f(\boldsymbol{x}) \doteq \infty \cdot \mathbf{1}$, so we can compactly write the set of subdifferentiable convex functions with bounded slope ($L \in \mathbb{R}_{>0} \cup \{\infty\}$) as

$$\mathcal{F}_{\mathrm{cx}}^L \doteq \left\{f \in \mathcal{F}_{\mathrm{cx}} \,\middle|\, \sup_{\boldsymbol{x} \in \mathbb{X}} \|\nabla_* f(\boldsymbol{x})\| < L\right\}.$$

The goal of this chapter is to derive sample efficient nonparametric estimators for the class of convex sub-Gaussian regression problems given as

$$\hat{\mathbb{M}}^{B,\sigma,d}_{\text{subgs}}(\mathcal{F}^L_{\text{cx}}) \doteq \Big\{ \mu \,\Big|\, (\boldsymbol{\mathcal{X}}, \mathcal{Y}) \sim \mu,\, \boldsymbol{\mathcal{X}} \in \mathbb{R}^d,\, \mathcal{Y} \in \mathbb{R},\, \|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|_{\Psi_2} \le B, \\ f_* \in \operatorname*{argmin}_{f \in \{\mathbb{X} \to \mathbb{R}\}} R_\mu(f) \subset \mathcal{F}^L_{\text{cx}},\, \mathbb{E}\big[e^{|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})|^2/\sigma^2} \big| \boldsymbol{\mathcal{X}}\big] \le 2 \Big\}.$$

Notice that for all problems in $\hat{\mathbb{M}}^{B,\sigma,d}_{\text{subgs}}(\mathcal{F}^L_{\text{cx}})$, f_* is an L-Lipschitz, convex regression function satisfying $f_* \in \mathcal{F}^L_{\text{cx}}$. For these estimation tasks, we use *max-affine functions*, formed by the maximum of finitely many hyperplanes, which are able to achieve a near-minimax rate.

5.1 Max-affine functions for nonparametric estimation

It is well-known that one can find convex α-ERM($\mathcal{F}_{\text{cx}}$) estimators (3.1) among max-affine functions. The reason is that any α-ERM($\mathcal{F}_{\text{cx}}$) estimate f_n can be linearized above the data points $\boldsymbol{\mathcal{X}}_1, \dots, \boldsymbol{\mathcal{X}}_n$ and approximated from below by taking the maximum of these linearizations (first-order Taylor approximation) as $f_n(\boldsymbol{x}) \ge \max_{i=1,\dots,n} \boldsymbol{a}_i^\top (\boldsymbol{x} - \boldsymbol{\mathcal{X}}_i) + f_n(\boldsymbol{\mathcal{X}}_i)$ with subgradients $\boldsymbol{a}_i \in \partial f_n(\boldsymbol{\mathcal{X}}_i)$. Because such a max-affine approximation attains the same values as f_n at the points $\boldsymbol{\mathcal{X}}_1, \dots, \boldsymbol{\mathcal{X}}_n$, it also attains the same empirical risk, so it is an α-ERM($\mathcal{F}_{\text{cx}}$) estimate as well.

Hence, one can solve the infinite dimensional α-ERM($\mathcal{F}_{\text{cx}}$) optimization problem by searching through the finite dimensional space of max-affine representations $f_n(\boldsymbol{x}) = \max_{k=1,\dots,n} \boldsymbol{a}_k^\top (\boldsymbol{x} - \boldsymbol{\mathcal{X}}_i) + b_k$, and computing the solution by the following quadratic program (for example, Holloway, 1979; Boyd and Vandenberghe, 2004, Section 6.5.5; Kuosmanen, 2008):

$$\min_{\substack{\boldsymbol{a}_1,\dots,\boldsymbol{a}_n,\\ b_1,\dots,b_n}} \sum_{i=1}^n (\mathcal{Y}_i - b_i)^2 \quad \text{subject to} \\ b_i \ge \boldsymbol{a}_k^\top (\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{X}}_k) + b_k\,, \quad i,k = 1,\dots,n\,. \tag{5.1}$$

In (5.1), the objective value is the empirical risk $R_n(f_n)$ and the constraints enforce that the i-th hyperplane placed over $\boldsymbol{\mathcal{X}}_i$ is also active at $\boldsymbol{\mathcal{X}}_i$ by providing the value of f_n at $\boldsymbol{\mathcal{X}}_i$, that is $f_n(\boldsymbol{\mathcal{X}}_i) = \max_{k=1,\dots,K} \boldsymbol{a}_k^\top (\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{X}}_k) + b_k = b_i$.

Although max-affine ERM($\mathcal{F}_{\text{cx}}$) estimators are consistent (Seijo and Sen, 2011; Lim and Glynn, 2012), their slope can be arbitrarily large near the boundary of the sample, and so their excess risk might become arbitrarily large too. Moreover, Balázs et al. (2015, Section 4.3) provided an example as shown on Figure 5.1 for which any ERM($\mathcal{F}_{\text{cx}}$) estimate has an infinite expected excess risk $\mathbb{E}[L_\mu(f_n, f_*)]$.

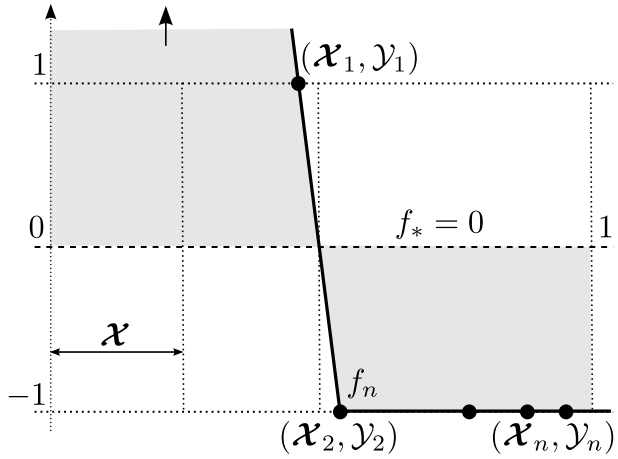

Figure 5.1: Worst case example of overfitting by unregularized max-affine estimators. As $\boldsymbol{\mathcal{X}}_1$ gets close to $\boldsymbol{\mathcal{X}}_2$, the slope of the estimate f_n between $(\boldsymbol{\mathcal{X}}_1, \mathcal{Y}_1)$ and $(\boldsymbol{\mathcal{X}}_2, \mathcal{Y}_2)$ grows to infinity, and so does the distance between f_n and f_* as indicated by the gray area.

To discuss Figure 5.1, take the unit domain $\mathbb{X} \doteq [0, 1]$, and let $\boldsymbol{\mathcal{X}} \in \mathbb{X}$, $\mathcal{Y} \in \{-1, +1\}$ be independent uniform random variables (so $f_* = 0$). Further, let $n \geq 2$, and define the event

$$E_n \doteq \big\{\boldsymbol{\mathcal{X}}_1 \in [\tfrac{1}{4}, \tfrac{1}{2}],\, \boldsymbol{\mathcal{X}}_2 \in [\tfrac{1}{2}, \tfrac{3}{4}],\, \boldsymbol{\mathcal{X}}_3, \ldots, \boldsymbol{\mathcal{X}}_n \geq \tfrac{3}{4},\, \boldsymbol{\mathcal{X}} \leq \tfrac{1}{4},\\ \mathcal{Y}_1 = +1,\, \mathcal{Y}_2 = \ldots = \mathcal{Y}_n = -1\big\}.$$

Then $\mathbb{P}\{E_n\} = (1/4)^{n+1}(1/2)^n > 0$, and using the observation that any ERM($\mathcal{F}_{\text{cx}}$) estimate f_n minimizing the excess risk on $[0, \frac{1}{4}]$ is linear in $[0, \boldsymbol{\mathcal{X}}_2]$, we can lower bound the expected excess risk as

$$\begin{aligned}\mathbb{E}\big[L_\mu(f_n, f_*)\big] &\geq \mathbb{E}\left[\left(\frac{2\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2}{\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2}\right)^2 \,\middle|\, E_n\right]\mathbb{P}\{E_n\}\\ &\geq \mathbb{E}\left[\frac{1}{4(\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2)^2} \,\middle|\, E_n\right]\mathbb{P}\{E_n\} = \infty\,.\end{aligned}$$

Also notice that a finite high-probability excess risk bound cannot exist either, otherwise (3.2) would imply finite expected excess risk as well.

To avoid infinite excess risk, one can regularize by bounding the slope with some $L \in \mathbb{R}_{>0}$, and adapting (5.1) to compute an α-ERM$(\mathcal{F}_{\text{cx}}^L)$ estimate. For a bounded domain $\mathbb{X} \subset \mathbb{R}^d$, Lim (2014) showed that such an α-ERM$(\mathcal{F}_{\text{cx}}^L)$ estimate f_n satisfies

$$L_\mu(f_n, f_*) = \begin{cases} O_{\mathbb{P}}\big(n^{-4/(d+4)}\big) & \text{if } d < 4, \\ O_{\mathbb{P}}\big(\ln(n)\, n^{-1/2}\big) & \text{if } d = 4, \\ O_{\mathbb{P}}\big(n^{-2/d}\big) & \text{if } d > 4, \end{cases} \tag{5.2}$$

where the stochastic growth notation $\mathcal{W}_n = O_{\mathbb{P}}(r_n)$ holds for some random variables $\mathcal{W}_n$, and constants $r_n > 0$, $n \in \mathbb{N}$, if there exists $c > 0$ such that $\limsup_{n\to\infty} \mathbb{P}\{|\mathcal{W}_n| \leq c\, r_n\} = 1$. Similar bounds have been proved for the expected excess risk $\mathbb{E}[L_\mu(f_n, f_*)]$ by Balázs et al. (2015, Section 4.2), which are $\ln(n)$ factor weaker than (5.2), but their dependence on d is also shown to scale only polynomially in d. However, the rates given by (5.2) for $d > 4$ are weaker than the minimax rate of the problem class $\hat{\mathbb{M}}_{\text{subgs}}^{B,\sigma,d}(\mathcal{F}_{\text{cx}}^L)$ which is of order $n^{-4/(d+4)}$ as shown below in Section 5.2.

The phase transition in the bound (5.2) at $d = 4$ happens because the class of uniformly L-Lipschitz, convex functions $\mathcal{F}_{\text{cx}}^L$ becomes large enough to make the entropy integral in Theorem 3.2 diverge for $d > 4$ with a polynomial rate. This phenomenon is well-known for nonparametric settings (for example regression over Sobolev spaces, see van de Geer, 2000, Page 188), and might be avoided by sieved ERM estimation (van de Geer, 2000, Section 10.3) when the hypothesis class $\mathcal{F}_{\text{cx}}^L$ is further restricted to balance the estimation and approximation error terms, $\theta\mathcal{H}_\psi(\epsilon, \mathcal{F})/n$ and B_* in Theorem 3.2, respectively, by choosing ϵ and $\hat{\mathcal{F}}_n$ appropriately.

In particular, Balázs et al. (2015, Section 4.4) has shown for the bounded case that estimators restricting $\mathcal{F}_{\text{cx}}^L$ to max-affine functions with at most $\lceil n^{d/(d+4)} \rceil$ hyperplanes (instead of n as used for (5.1)) achieve a near-minimax rate for any $d \in \mathbb{N}$. After discussing the minimax lower bound in Section 5.2 and the approximation properties of max-affine representations in Section 5.3, we study the theoretical properties of these sieved max-affine ERM estimators in Section 5.4 by extending their analysis for sub-Gaussian convex problems $\hat{\mathbb{M}}_{\text{subgs}}^{B,\sigma,d}(\mathcal{F}_{\text{cx}}^L)$ from the bounded case.

5.2 Lower bound on the minimax rate

To derive a lower bound on the minimax rate for convex nonparametric regression, we consider Gaussian problems as required for Theorem 3.1. For this we need an asymptotic bound on the entropy of $\mathcal{F}_{\mathrm{cx}}^L$, for which Guntuboyina and Sen (2013) proved[1] that $\mathcal{H}_{P_{\mathbb{X}_0}}(\epsilon, \mathcal{F}_{\mathrm{cx}}^L) = \Theta(\epsilon^{-d/2})$ for a sufficiently small $\epsilon > 0$ if $\mathbb{X}$ is bounded and $P_{\mathbb{X}_0}$ is the uniform distribution on $\mathbb{X}$. Combining this with Theorem 3.1b, we obtain a lower bound on the minimax rate (Balázs et al., 2015, Theorem 4.1) as presented by the following result:

Theorem 5.1. *For a sufficiently large n,*

$$\mathcal{R}_n\Big(\mathbb{M}_{gs}^{\sigma}(\mathcal{F}_{cx}^L, P_{\mathbb{X}_0}), \ell_{sq}, \{\mathbb{X} \to \mathbb{R}\}\Big) = \Omega\big(n^{-4/(d+4)}\big)\,.$$

As $\mathbb{M}_{\mathrm{gs}}^{\sigma}(\mathcal{F}_{\mathrm{cx}}^L, P_{\mathbb{X}_0}) \subset \hat{\mathbb{M}}_{\mathrm{subgs}}^{B,\sigma,d}(\mathcal{F}_{\mathrm{cx}})$, the lower bound of Theorem 5.1 holds for convex sub-Gaussian regression problems as well. The rest of the chapter is devoted to show that there exist sieved max-affine estimators which achieve this minimax rate up to logarithmic factors, so implying that the lower bound on the minimax rate of Theorem 5.1 is tight.

Finally, we mention that the entropy of the class $\mathcal{F}_{\mathrm{cx}}$ ($\mathcal{F}_{\mathrm{cx}}^L$ without the Lipschitz bound) over the unit ball domain $\mathbb{X}$ is only $\mathcal{H}_{\psi}(\epsilon, \mathcal{F}_{\mathrm{cx}}) = \Theta(\epsilon^{1-d})$ (Gao and Wellner, 2015, Corollary 1.4 and Theorem 1.5). Then, the minimax risk is also weaker in this case. In fact Han and Wellner (2016, Theorems 2.3 and 2.4) showed that $\mathcal{R}_n\big(\mathbb{M}_{\mathrm{gs}}^{\sigma}(\mathcal{F}_{\mathrm{cx}}, P_{\mathbb{X}_0}), \ell_{\mathrm{sq}}, \{\mathbb{X} \to \mathbb{R}\}\big) = \Theta\big(n^{-2/(d+1)}\big)$ up to logarithmic factors. Hence, the Lipschitz restriction is necessary to keep the minimax rate independent from the shape of the domain boundary.

5.3 Max-affine approximations

To construct sample efficient sieved ERM estimators with max-affine representations, we need to understand their approximation accuracy to the function

[1] Guntuboyina and Sen (2013) proved the lower bound without the Lipschitz bound, which is a larger function class. However, in the proof of their Theorem 3.3, they construct a packing subset by functions having a 2-bounded Lipschitz constant with respect to $\|\cdot\|$. These functions could be rescaled appropriately to deliver our statement. For the upper bound, simply consider the sup-norm result, Theorem 3.2 in their paper.

class $\mathcal{F}_{\text{cx}}^L$. Hence, we now study the approximation properties of the following max-affine class with functions represented by the maximum of at most $K \in \mathbb{N}$ affine functions,

$$\mathcal{F}_{\text{ma}}^{K,L}(\boldsymbol{x}_0) \doteq \left\{ h : \mathbb{X} \to \mathbb{R} \,\middle|\, h(\boldsymbol{x}) = \max_{k=1,\dots,K} \boldsymbol{a}_k^\top (\boldsymbol{x} - \boldsymbol{x}_0) + b_k,\ \|\boldsymbol{a}_k\| \le L \right\},$$

where $\boldsymbol{x}_0 \in \mathbb{X}$ is some reference point. To emphasize the "owner function" of the parameters $\boldsymbol{a}_k, b_k$, we sometimes write $\boldsymbol{a}_k(h), b_k(h)$ to refer the quantities satisfying $h(\boldsymbol{x}) = \max_{k=1,\dots,K} \boldsymbol{a}_k(h)^\top (\boldsymbol{x} - \boldsymbol{x}_0) + b_k(h)$.

Then consider the following result (Lemma 5.2), which describes the approximation accuracy of $\mathcal{F}_{\text{ma}}^{K,L}(\boldsymbol{x}_0)$ to $\mathcal{F}_{\text{cx}}^L$ over a bounded domain. This lemma is a slight modification of Lemma 4.1 in Balázs et al. (2015), which transfers the convex set estimation result of Bronshteyn and Ivanov (1975) to convex functions over a bounded domain. Then main trick of the proof is to construct a cover over a subset of $\mathbb{R}^d$, which covers both $\mathbb{X} \cap \mathcal{B}_2(\boldsymbol{x}_0, R)$ and the gradient space $\{\nabla_* f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X} \cap \mathcal{B}_2(\boldsymbol{x}_0, R)\}$ of a convex function $f \in \mathcal{F}_{\text{cx}}^L$. This way we can double the approximation rate for the convex case, providing $O(K^{-2/d})$ accuracy instead of the weaker $O(K^{-1/d})$ which is the approximation rate of piecewise linear functions to Lipschitz continuous ones (Cooper, 1995).

Lemma 5.2. *For each $f \in \mathcal{F}_{cx}^L$ there exists $h_f \in \mathcal{F}_{ma}^{K,L}(\boldsymbol{x}_0)$ such that*

$$\sup_{\boldsymbol{x} \in \mathbb{X} \cap \mathcal{B}_2(\boldsymbol{x}_0, R)} \left| f(\boldsymbol{x}) - h_f(\boldsymbol{x}) \right| \le 36 L R K^{-2/d}.$$

Furthermore, $0 \le f(\boldsymbol{x}) - h_f(\boldsymbol{x}) \le 2L(R + \|\boldsymbol{x} - \boldsymbol{x}_0\|)$ holds for all $\boldsymbol{x} \in \mathbb{X}$, and $b_k(h_f) - f(\boldsymbol{x}_0) \in [-2LR, 0]$ for all $k = 1, \dots, K$.

Proof. Fix $f \in \mathcal{F}_{\text{cx}}^L$ arbitrarily and recall that $\|\nabla_* f(\boldsymbol{x})\| \le L$ for all $\boldsymbol{x} \in \mathbb{X}$. Let $t > 0$ to be chosen later, $\mathbb{X}_R \doteq \mathbb{X} \cap \mathcal{B}_2(\boldsymbol{x}_0, R)$ and define the mapping $\nu(\boldsymbol{x}) \doteq \boldsymbol{x} + t \nabla_* f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{X}_R$. Notice that by the convexity of f, ν is an injective function (that is $\nu(\boldsymbol{x}) = \nu(\boldsymbol{z}) \iff \boldsymbol{x} = \boldsymbol{z}$ for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{X}_R$) because

$$\begin{aligned}
f(\boldsymbol{z}) &\ge \nabla_* f(\boldsymbol{x})^\top (\boldsymbol{z} - \boldsymbol{x}) + f(\boldsymbol{x}) && \text{(by convexity of } f\text{)} \\
&= \|\boldsymbol{z} - \boldsymbol{x}\|^2 / t + \nabla_* f(\boldsymbol{z})^\top (\boldsymbol{z} - \boldsymbol{x}) + f(\boldsymbol{x}) && \text{(by } \nu(\boldsymbol{x}) = \nu(\boldsymbol{z})\text{)} \\
&\ge \|\boldsymbol{z} - \boldsymbol{x}\|^2 / t + f(\boldsymbol{z}), && \text{(by convexity of } f\text{)}
\end{aligned}$$

which holds only if $\boldsymbol{x} = \boldsymbol{z}$.

Now define $R_t \doteq R + tL$, $\mathcal{K} \doteq \{\nu(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}_R\} \subseteq \mathcal{B}_2(\boldsymbol{x}_0, R_t)$ and for some $\epsilon > 0$ to be chosen later, let $\mathcal{K}_\epsilon \subseteq \mathcal{K}$ be a $\sqrt{\epsilon}$-cover of $\mathcal{K}$ under $\|\cdot\|$. Furthermore, let $\mathbb{X}_\epsilon \doteq \{\nu^{-1}(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{K}_\epsilon\}$, where ν^{-1} denotes the inverse of ν (which is invertible on $\mathcal{K}$ because it is injective on $\mathbb{X}_R$). Then by Lemma C.1, $|\mathbb{X}_\epsilon| = |\mathcal{K}_\epsilon| = \mathcal{N}_{\|\cdot\|}(\sqrt{\epsilon}, \mathcal{K}) \leq (9R_t^2/\epsilon)^{d/2}$ for all $\epsilon \in (0, 9R_t^2]$.

Next we show that $\mathcal{K}_\epsilon$ induces two $\sqrt{\epsilon}$-covers, one on $\mathbb{X}_R$ and one on the scaled gradient space $\{t\nabla_* f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\}$. For this, the convexity of f implies

$$\begin{aligned}\big(\nabla_* f(\boldsymbol{x}) - \nabla_* f(\boldsymbol{z})\big)^\top (\boldsymbol{x} - \boldsymbol{z}) &= \nabla_* f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{z}) + \nabla_* f(\boldsymbol{z})^\top (\boldsymbol{z} - \boldsymbol{x}) \\ &\geq f(\boldsymbol{x}) - f(\boldsymbol{z}) + f(\boldsymbol{z}) - f(\boldsymbol{x}) = 0\,,\end{aligned} \tag{5.3}$$

for any $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{X}$. Let $\hat{\boldsymbol{x}} \doteq \operatorname{argmin}_{\boldsymbol{z} \in \mathbb{X}_\epsilon} \|\boldsymbol{x} - \boldsymbol{z}\|$ for any $\boldsymbol{x} \in \mathbb{X}$ (if multiple minima exist, fix one arbitrarily). Notice that if $\boldsymbol{x} \in \mathbb{X}_R$, we also have $\|\nu(\boldsymbol{x}) - \nu(\hat{\boldsymbol{x}})\| \leq \sqrt{\epsilon}$ due to the construction of $\mathcal{K}_\epsilon$. Hence, by (5.3), we obtain

$$\begin{aligned}&\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + t^2 \|\nabla_* f(\boldsymbol{x}) - \nabla_* f(\hat{\boldsymbol{x}})\|^2 \\ &\quad\leq \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + 2t\big(\nabla_* f(\boldsymbol{x}) - \nabla_* f(\hat{\boldsymbol{x}})\big)^\top (\boldsymbol{x} - \hat{\boldsymbol{x}}) + t^2 \|\nabla_* f(\boldsymbol{x}) - \nabla_* f(\hat{\boldsymbol{x}})\|^2 \\ &\quad= \|\nu(\boldsymbol{x}) - \nu(\hat{\boldsymbol{x}})\|^2 \leq \epsilon\,,\end{aligned}$$

for any $\boldsymbol{x} \in \mathbb{X}_R$. So we have $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leq \sqrt{\epsilon}$ and $t\|\nabla_* f(\boldsymbol{x}) - \nabla_* f(\hat{\boldsymbol{x}})\| \leq \sqrt{\epsilon}$.

Now we construct a max-affine approximation to f. Choose $\epsilon \in (0, 9R_t^2]$ to satisfy $K = (9R_t^2/\epsilon)^{d/2} \geq |\mathbb{X}_\epsilon|$ and a set $\mathbb{X}_K \doteq \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_K\}$ such that $\mathbb{X}_\epsilon \subseteq \mathbb{X}_K$. Notice that $\hat{\boldsymbol{x}} \in \mathbb{X}_K$ for all $\boldsymbol{x} \in \mathbb{X}_R$. Then consider the max-affine function $h_f(\boldsymbol{x}) \doteq \max_{k=1,\ldots,K} f(\hat{\boldsymbol{x}}_k) + \nabla_* f(\hat{\boldsymbol{x}}_k)^\top (\boldsymbol{x} - \hat{\boldsymbol{x}}_k)$. Using the convexity of f, we have for all $\boldsymbol{x} \in \mathbb{X}$ that

$$\begin{aligned}&0 \leq f(\boldsymbol{x}) - h_f(\boldsymbol{x}) \leq f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}) - \nabla_* f(\hat{\boldsymbol{x}})^\top (\boldsymbol{x} - \hat{\boldsymbol{x}}) \\ &\quad\leq \nabla_* f(\boldsymbol{x})^\top (\boldsymbol{x} - \hat{\boldsymbol{x}}) - \nabla_* f(\hat{\boldsymbol{x}})^\top (\boldsymbol{x} - \hat{\boldsymbol{x}}) = \big(\nabla_* f(\boldsymbol{x}) - \nabla_* f(\hat{\boldsymbol{x}})\big)^\top (\boldsymbol{x} - \hat{\boldsymbol{x}})\,,\end{aligned} \tag{5.4}$$

which implies $f(\boldsymbol{x}) - h_f(\boldsymbol{x}) \leq 2L(\|\boldsymbol{x} - \boldsymbol{x}_0\| + R)$. Additionally, if $\boldsymbol{x} \in \mathbb{X}_R$, use the Cauchy-Schwartz inequality for (5.4) to get

$$0 \leq f(\boldsymbol{x}) - h_f(\boldsymbol{x}) \leq \frac{1}{t}\Big(t\|\nabla_* f(\boldsymbol{x}) - \nabla_* f(\hat{\boldsymbol{x}})\|\,\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|\Big) \leq \epsilon/t\,.$$

Then, rearranging $K = (9R_t^2/\epsilon)^{d/2}$, we obtain the bound in the claim over $\mathbb{X}_R$ by $\epsilon = 9R_t^2 K^{-2/d}$ and $t \doteq R/L$ as $\sup_{\boldsymbol{x} \in \mathbb{X}_R} |f(\boldsymbol{x}) - h_f(\boldsymbol{x})| \leq \epsilon/t = 36RLK^{-2/d}$.

Finally, let $b_k \doteq f(\hat{\boldsymbol{x}}_k) + \nabla_* f(\hat{\boldsymbol{x}}_k)^\top (\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_k)$ and $\boldsymbol{a}_k \doteq \nabla_* f(\hat{\boldsymbol{x}}_k)$, so we have $h_f(\boldsymbol{x}) = \max_{k=1,\dots,K} \boldsymbol{a}_k^\top (\boldsymbol{x} - \boldsymbol{x}_0) + b_k$. Clearly, $\|\boldsymbol{a}_k\| \leq L$ so $h_f \in \mathcal{F}_{\mathrm{ma}}^{K,L}(\boldsymbol{x}_0)$. Furthermore, by the convexity of f, $b_k \leq f(\boldsymbol{x}_0)$ and

$$\begin{aligned} b_k &= f(\hat{\boldsymbol{x}}_k) - f(\boldsymbol{x}_0) + f(\boldsymbol{x}_0) + \nabla_* f(\hat{\boldsymbol{x}}_k)^\top (\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_k) \\ &\geq f(\boldsymbol{x}_0) + \big(\nabla_* f(\boldsymbol{x}_0) - \nabla_* f(\hat{\boldsymbol{x}}_k)\big)^\top (\hat{\boldsymbol{x}}_k - \boldsymbol{x}_0) \geq f(\boldsymbol{x}_0) - 2LR\,, \end{aligned}$$

which proves the claim for $b_k(h)$. ■

5.4 Near-minimax upper bounds for max-affine LSEs

Here we provide the main convex regression result of this thesis (Theorem 5.6), that is an excess risk upper bound for α-ERM$\big(\mathcal{F}_{\mathrm{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})\big)$ max-affine estimators and sub-Gaussian convex regression settings $\mu \in \hat{\mathbb{M}}_{\mathrm{subgs}}^{B,\sigma,d}(\mathcal{F}_{\mathrm{cx}}^{L})$ with squared loss ($\ell = \ell_{\mathrm{sq}}$). We show that by using $K = \lceil n^{d/(d+4)} \rceil$ hyperplanes these max-affine estimators achieve a minimax rate $n^{-4/(d+4)}$ up to logarithmic factors.

To derive the excess risk bound, we use Theorem 3.2. However, as the entropy of the hypothesis class $\mathcal{F}_{\mathrm{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})$ is infinite due to the unbounded magnitude of its functions, Theorem 3.2 cannot be applied directly. To handle this, first we show that max-affine ERM estimators in $\mathcal{F}_{\mathrm{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})$ lie in some data-dependent set $\mathcal{F}_n$ with a bounded magnitude such that $\mathcal{F}_n$ approximates any regression function $f_* \in \mathcal{F}_{\mathrm{cx}}^{L}$ with enough accuracy. Then, we can find a data-independent envelope $\mathcal{F}$ of $\mathcal{F}_n$ such that $\mathcal{F}_n \subseteq \mathcal{F}$ holds with high-probability, and class $\mathcal{F}$ has an appropriately bounded entropy.

Let $f_n(\boldsymbol{x}) = \max_{k=1,\dots,K} \boldsymbol{a}_k^\top (\boldsymbol{x} - \bar{\boldsymbol{\mathcal{X}}}) + b_k$ be an α-ERM$(\mathcal{F}_n)$ estimate, and observe that its offset terms, $b_k \in \mathbb{R}$, $k = 1, \dots, K$, are unconstrained by the definition of $\mathcal{F}_{\mathrm{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})$. To derive a bound on the offset terms b_k, we use the α-ERM$(\mathcal{F}_n)$ property (3.1), and drop every hyperplane out of the representation of f_n, which are not active at any point $\boldsymbol{\mathcal{X}}_1, \dots, \boldsymbol{\mathcal{X}}_n$. Formally, we assume that for all $k \in \{1, \dots, K\}$ there exists $i \in \{1, \dots, n\}$ such that $f_n(\boldsymbol{\mathcal{X}}_i) = \boldsymbol{a}_k^\top (\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mathcal{X}}}) + b_k$. This assumption is not restrictive as dropping the inactive hyperplanes from the representation of f_n does not affect the empirical

risk $R_n(f_n)$. We also use this assumption in the rest of the section without further notice. Then, Lemma 5.3 proves a bound on the b_k terms.

Lemma 5.3. *If f_n is an α-ERM$\big(\mathcal{F}_{ma}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})\big)$ estimate (3.1) for $\ell = \ell_{sq}$ with any $\alpha \leq 2(L\mathcal{X}_{\max})^2$, then f_n satisfies the α-ERM($\mathcal{F}_n$) property as well with*

$$\mathcal{F}_n \doteq \left\{ f \in \mathcal{F}_{ma}^{K,L}(\bar{\boldsymbol{\mathcal{X}}}) : |b_k(f) - \bar{\mathcal{Y}}| \leq 5L\mathcal{X}_{\max} + \sqrt{2R_n(f_*)},\ k = 1, \dots, K \right\},$$

where $\mathcal{X}_{\max} \doteq \max_{i=1,\dots,n} \|\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mathcal{X}}}\|$.

Proof. Let $f_n(\boldsymbol{x}) = \max_{k=1,\dots,K} \boldsymbol{a}_k^\top(\boldsymbol{x} - \bar{\boldsymbol{\mathcal{X}}}) + b_k$ be an α-ERM$\big(\mathcal{F}_{\text{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})\big)$ estimate and $l \in \{1, \dots, K\}$. Take some $j \in \{1, \dots, n\}$ for which the l-th hyperplane is active on $\boldsymbol{\mathcal{X}}_j$, that is $f_n(\boldsymbol{\mathcal{X}}_j) = \boldsymbol{a}_l^\top(\boldsymbol{\mathcal{X}}_j - \bar{\boldsymbol{\mathcal{X}}}) + b_l$. Using this, the triangle inequality with the Lipschitz property of f_n, and the Cauchy-Schwartz inequality, we get

$$\begin{aligned}
|b_l - \bar{\mathcal{Y}}| &= |f_n(\boldsymbol{\mathcal{X}}_j) - \bar{\mathcal{Y}} + \boldsymbol{a}_l^\top(\boldsymbol{\mathcal{X}}_j - \bar{\boldsymbol{\mathcal{X}}})| \\
&\leq |f_n(\boldsymbol{\mathcal{X}}_j) - f_n(\bar{\boldsymbol{\mathcal{X}}}) + f_n(\bar{\boldsymbol{\mathcal{X}}}) - \bar{\mathcal{Y}}| + L\mathcal{X}_{\max} \\
&\leq |f_n(\bar{\boldsymbol{\mathcal{X}}}) - \bar{\mathcal{Y}}| + 2L\mathcal{X}_{\max} \\
&= \left| f_n(\bar{\boldsymbol{\mathcal{X}}}) - \frac{1}{n}\sum_{i=1}^n f_n(\boldsymbol{\mathcal{X}}_i) + \frac{1}{n}\sum_{i=1}^n f_n(\boldsymbol{\mathcal{X}}_i) - \bar{\mathcal{Y}} \right| + 2L\mathcal{X}_{\max} \qquad (5.5) \\
&\leq \left| \frac{1}{n}\sum_{i=1}^n f_n(\boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i \right| + 3L\mathcal{X}_{\max} \\
&\leq \sqrt{R_n(f_n)} + 3L\mathcal{X}_{\max}\,.
\end{aligned}$$

Observe that any constant function including $f_0(\boldsymbol{x}) \doteq f_*(\bar{\boldsymbol{\mathcal{X}}})$, $\boldsymbol{x} \in \mathbb{X}$, is in $\mathcal{F}_{\text{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})$. Hence, using the α-ERM$\big(\mathcal{F}_{\text{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})\big)$ property (3.1) of f_n, $(a+b)^2 \leq 2a^2 + 2b^2$, the Lipschitz property of f_* with $\alpha \leq 2(L\mathcal{X}_{\max})^2$, we get

$$\begin{aligned}
R_n(f_n) \leq R_n(f_0) + \alpha &= \frac{1}{n}\sum_{i=1}^n \left| \mathcal{Y}_i - f_*(\boldsymbol{\mathcal{X}}_i) + f_*(\boldsymbol{\mathcal{X}}_i) - f_*(\bar{\boldsymbol{\mathcal{X}}}) \right|^2 + \alpha \\
&\leq 2R_n(f_*) + \frac{2}{n}\sum_{i=1}^n \left| f_*(\boldsymbol{\mathcal{X}}_i) - f_*(\bar{\boldsymbol{\mathcal{X}}}) \right|^2 + \alpha \leq 2R_n(f_*) + 4(L\mathcal{X}_{\max})^2\,,
\end{aligned}$$

which implies the claim by $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. ∎

Next, we show that $\mathcal{F}_n$ preserves the universal approximation properties of max-affine functions $\mathcal{F}_{\text{ma}}^{K,L}(\bar{\boldsymbol{\mathcal{X}}})$ to uniformly Lipschitz, convex maps $\mathcal{F}_{\text{cx}}^{L}$

as shown by Lemma 5.2. For the proof, we take a subset of $\mathcal{F}_n$ which still preserves approximation quality, but by being independent of $\mathcal{Y}_1, \dots, \mathcal{Y}_n$, it also let us exploit the properties of the squared loss for the sub-Gaussian setting (Lemma F.3). The details are given by the following result:

Lemma 5.4. *For any* $\mu \in \hat{\mathbb{M}}^{B,\sigma,d}_{subgs}(\mathcal{F}^L_{cx})$, $\gamma > 0$ *and* $\alpha \leq 2(L\mathcal{X}_{\max})^2 K^{-4/d}$,

$$\mathbb{P}\Big\{ \inf_{f \in \mathcal{F}_n} L_n(f, f_*) + \alpha > B_* \Big\} \leq \frac{\gamma}{4}\,,$$

where $B_* \doteq \max\big\{8(36^2+1)(LB)^2K^{-4/d},\, 2\sigma^2/n\big\} \ln(8n/\gamma)$.

Proof. First, consider the class

$$\hat{\mathcal{F}}_n \doteq \Big\{ f \in \mathcal{F}^{K,L}_{\text{ma}}(\bar{\boldsymbol{\mathcal{X}}}) : |b_k(f) - f_*(\bar{\boldsymbol{\mathcal{X}}})| \leq 2L\mathcal{X}_{\max},\, k = 1, \dots, K \Big\}\,,$$

and notice that $\hat{\mathcal{F}}_n \subseteq \mathcal{F}_n$, because by Jensen's inequality, the triangle inequality with the Lipschitzness of f_*, and the Cauchy-Schwartz inequality, we have

$$\begin{aligned} |f_*(\bar{\boldsymbol{\mathcal{X}}}) - \bar{\mathcal{Y}}| &\leq \frac{1}{n}\sum_{i=1}^n \big| f_*(\bar{\boldsymbol{\mathcal{X}}}) - f_*(\boldsymbol{\mathcal{X}}_i) + f_*(\boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i \big| \\ &\leq L\mathcal{X}_{\max} + \frac{1}{n}\sum_{i=1}^n |f_*(\boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i| \leq L\mathcal{X}_{\max} + \sqrt{R_n(f_*)}\,. \end{aligned}$$

Next, let $h_* \in \mathcal{F}^{K,L}_{\text{ma}}(\bar{\boldsymbol{\mathcal{X}}})$ be the max-affine approximation to $f_* \in \mathcal{F}^L_{\text{cx}}$ as given by Lemma 5.2 with $\boldsymbol{x}_0 = \bar{\boldsymbol{\mathcal{X}}}$ and $R = \mathcal{X}_{\max}$, which also provides $b_k(h_*) - f_*(\bar{\boldsymbol{\mathcal{X}}}) \in [-2L\mathcal{X}_{\max}, 0]$ for all $k = 1, \dots, K$, so we have $h_* \in \hat{\mathcal{F}}_n$. Additionally, the approximation bound for h_*, provided by Lemma 5.2, and the bound on α implies

$$\inf_{f \in \hat{\mathcal{F}}_n} \frac{2}{n}\sum_{i=1}^n |f(\boldsymbol{\mathcal{X}}_i) - f_*(\boldsymbol{\mathcal{X}}_i)|^2 + \alpha \leq 2(36^2+1)(L\mathcal{X}_{\max})^2 K^{-4/d}\,.$$

Further, $\mathcal{X}_{\max} \leq \max_{i=1,\dots,n} \|\boldsymbol{\mathcal{X}}_i - \mathbb{E}\boldsymbol{\mathcal{X}}\| + \|\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| \leq 2\max_{i=1,\dots,n} \|\boldsymbol{\mathcal{X}}_i - \mathbb{E}\boldsymbol{\mathcal{X}}\|$ due to Jensen's inequality, so we also have

$$\mathbb{E}\big[e^{\mathcal{X}^2_{\max}/(2B)^2}\big] = \mathbb{E}\Big[\max_{i=1,\dots,n} e^{\|\boldsymbol{\mathcal{X}}_i - \mathbb{E}\boldsymbol{\mathcal{X}}\|^2/B^2}\Big] \leq 2n\,. \tag{5.6}$$

Hence, by using Lemma F.3 with $T = \alpha$, $R = 8(36^2+1)(LB)^2K^{-4/d}$ and $c = 2n$, we get the claim for B_*. ∎

In order to provide a data-independent, high-probability envelope of $\mathcal{F}_n$, we replace its sub-Gaussian quantities $\bar{\boldsymbol{\mathcal{X}}}$, $\bar{\mathcal{Y}}$ by their limit values $\mathbb{E}\boldsymbol{\mathcal{X}}$, $\mathbb{E}\mathcal{Y}$, respectively, and the bound on the offset terms by an appropriate data-independent value $t > 0$. Formally, we consider the envelope $\mathcal{F}_n \subseteq \mathcal{F}(t)$ with

$$\mathcal{F}(t) \doteq \left\{ f \in \mathcal{F}_{\text{ma}}^{K,L}(\mathbb{E}\boldsymbol{\mathcal{X}}) : |b_k(f) - \mathbb{E}\mathcal{Y}| \leq t,\, k = 1, \dots, K \right\}.$$

The details, providing the value of t with respect to the sample size n and the probabilistic guarantee γ are given by Lemma 5.5.

Lemma 5.5. *Let $\mu \in \hat{\mathbb{M}}_{subgs}^{B,\sigma,d}(\mathcal{F}_{cx}^{L})$, $\gamma > 0$ and $\alpha \leq (L\mathcal{X}_{\max})^2$. Then we have $\mathbb{P}\{\mathcal{F}_n \subseteq \mathcal{F}(t)\} \geq 1 - \gamma/2$ with $t \doteq 4\max\{11LB, 2\sigma\}\sqrt{\ln(4\sqrt{n}/\gamma)}$.*

Proof. Let $\hat{\mathcal{X}}_{\max} \doteq \max_{i=1,\dots,n} \|\boldsymbol{\mathcal{X}}_i - \mathbb{E}\boldsymbol{\mathcal{X}}\|$, and notice that $\mathcal{X}_{\max} \leq 2\hat{\mathcal{X}}_{\max}$ due to Jensen's inequality implying $\|\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| \leq \hat{\mathcal{X}}_{\max}$. Then, we also have

$$\begin{aligned}\mathbb{P}\{\mathcal{F}_n \subseteq \mathcal{F}(t)\} &\geq \mathbb{P}\left\{L\|\bar{\boldsymbol{\mathcal{X}}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + |\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| + 5L\mathcal{X}_{\max} + \sqrt{2R_n(f_*)} \leq t\right\} \\ &\geq 1 - \mathbb{P}\left\{11L\hat{\mathcal{X}}_{\max} + |\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| + \sqrt{2R_n(f_*)} > t\right\}. \end{aligned} \tag{5.7}$$

Next, we bound the sub-Gaussian parameter of $|\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| + \sqrt{2R_n(f_*)}$. For this, first decompose $|\mathcal{Y} - \mathbb{E}\mathcal{Y}|$ using the triangle inequality with the Lipschitz property of f_* and $\mathbb{E}[\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|] \leq B$ by Lemma A.2b for $k = 1$, to get

$$\begin{aligned}|\mathcal{Y} - \mathbb{E}\mathcal{Y}| &\leq |\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})| + |f_*(\boldsymbol{\mathcal{X}}) - f_*(\mathbb{E}\boldsymbol{\mathcal{X}})| + |f_*(\mathbb{E}\boldsymbol{\mathcal{X}}) - \mathbb{E}[f_*(\boldsymbol{\mathcal{X}})]| \\ &\leq |\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})| + L\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + LB. \end{aligned} \tag{5.8}$$

Further, Jensen's inequality implies $\|\sqrt{R_n(f_*)}\|_{\Psi_2} \leq \|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})\|_{\Psi_2} \leq \sigma$ and $\left\|\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}\right\|_{\Psi_2} \leq \|\mathcal{Y} - \mathbb{E}\mathcal{Y}\|_{\Psi_2}$. Then, use $\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|_{\Psi_2} \leq B$ with Lemma A.2d to obtain $\||\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| + \sqrt{2R_n(f_*)}\|_{\Psi_2} \leq 2(LB + \sigma)$.

Finally, using this sub-Gaussian property we bound the probability of (5.7). For this, we also use $(a+b)^2 \leq 2a^2 + 2b^2$, Markov's inequality, the Cauchy-Schwartz inequality, and (5.6), to obtain

$$\begin{aligned}&\mathbb{P}\left\{11L\hat{\mathcal{X}}_{\max} + |\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| + \sqrt{2R_n(f_*)} > t\right\} \\ &\qquad \leq \mathbb{P}\left\{(11L\hat{\mathcal{X}}_{\max})^2 + (|\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| + \sqrt{2R_n(f_*)})^2 > t^2/2\right\} \\ &\qquad \leq \mathbb{E}\left[e^{2(11L\hat{\mathcal{X}}_{\max})^2/m^2}\right]^{1/2} \mathbb{E}\left[e^{2\left(|\bar{\mathcal{Y}} - \mathbb{E}\mathcal{Y}| + \sqrt{2R_n(f_*)}\right)^2/m^2}\right]^{1/2} e^{-t^2/(2m^2)} \\ &\qquad \leq 2\sqrt{n}\, e^{-t^2/(2m^2)} = \gamma/2, \end{aligned}$$

where $m \doteq \sqrt{2}\max\{22LB, 2(LB + \sigma)\}$ and $t = m\sqrt{2\ln(4\sqrt{n}/\gamma)}$. ∎

Then, we are about to apply our general regression result Theorem 3.2 as specialized for the squared loss by Lemma 3.5 to prove an upper bound on the excess risk of max-affine estimators for sub-Gaussian convex regression settings. For this, we ignore Condition (C2) by setting $\delta = \epsilon$, and provide Conditions (C1) and (C3) using the sub-Gaussian property of the quantities $\mathcal{W}_f \doteq f(\boldsymbol{\mathcal{X}}) - f_*(\boldsymbol{\mathcal{X}})$, $f \in \mathcal{F}(t)$ which is implied by Lipschitzness and the sub-Gaussian distribution of $\boldsymbol{\mathcal{X}}$. The details are given by Theorem 5.6.

Theorem 5.6. *Consider the squared loss ($\ell = \ell_{sq}$), and a sub-Gaussian convex regression problem $\mu \in \hat{\mathbb{M}}^{B,\sigma,d}_{subgs}(\mathcal{F}^L_{cx})$. Further, let f_n be an α-ERM$\big(\mathcal{F}^{K,L}_{ma}(\overline{\boldsymbol{\mathcal{X}}})\big)$ estimate with $K \doteq \lceil n^{d/(d+4)} \rceil$ and $\alpha \leq (L\mathcal{X}_{\max})^2 K^{-4/d}$. Then, for any $\gamma > 0$, we have with probability at least $1-\gamma$ that*

$$L_\mu(f_n, f_*) = O\Big(\max\{LB, \sigma\}^2 \, n^{-4/(d+4)} \ln(n) \ln(n/\gamma) \Big) .$$

Proof. Let $\mathcal{F} \doteq \mathcal{F}(t)$ be the high-probability envelope of $\mathcal{F}_n$ as given for Lemma 5.5, and fix $f \in \mathcal{F}$ arbitrarily. First, we prove that $\mathcal{W}_f$ is sub-Gaussian.

Notice, that for the squared loss ($\ell = \ell_{\mathrm{sq}}$) and the regression function f_*, we have $\mathbb{E}[\mathcal{Y}] = \mathbb{E}[f_*(\boldsymbol{\mathcal{X}})]$. Hence, using the Lipschitz property of f_* with $\mathbb{E}[\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\|] \leq B$, we get $|\mathbb{E}[\mathcal{Y}] - f_*(\mathbb{E}\boldsymbol{\mathcal{X}})| = |\mathbb{E}[f_*(\boldsymbol{\mathcal{X}}) - f_*(\mathbb{E}\boldsymbol{\mathcal{X}})]| \leq LB$. Then, using that $f \in \mathcal{F}$ satisfies $|f(\boldsymbol{\mathcal{X}}) - \mathbb{E}\mathcal{Y}| \leq L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + t$, we obtain

$$\begin{aligned} |\mathcal{W}_f| &= \big| f(\boldsymbol{\mathcal{X}}) - \mathbb{E}[\mathcal{Y}] + \mathbb{E}[\mathcal{Y}] - f_*(\mathbb{E}\boldsymbol{\mathcal{X}}) + f_*(\mathbb{E}\boldsymbol{\mathcal{X}}) - f_*(\boldsymbol{\mathcal{X}}) \big| \\ &\leq t + LB + 2L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| \ , \end{aligned}$$

which implies that $\|\mathcal{W}_f\|_{\Psi_2} \leq t + 3LB$.

Next, we prove Condition (C1). For this, represent any $f, \hat{f} \in \mathcal{F}$ as $f(\boldsymbol{x}) \doteq \max_{k=1,\dots,K} \boldsymbol{a}_k^\top (\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + b_k$, $\hat{f}(\boldsymbol{x}) \doteq \max_{k=1,\dots,K} \hat{\boldsymbol{a}}_k^\top (\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) + \hat{b}_k$. Then for any $\boldsymbol{x} \in \mathbb{X}$, we have

$$\begin{aligned} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| &\leq \max_{k=1,\dots,K} \big| (\boldsymbol{a}_k - \hat{\boldsymbol{a}}_k)^\top (\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}) \big| + |b_k - \hat{b}_k| \\ &\leq \big(2L\,\|\boldsymbol{x} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + 2t \big)\, \psi(f, \hat{f}) \, , \end{aligned}$$

with metric $\psi(f, \hat{f}) \doteq \max_{k=1,\dots,K} \big\{ \frac{\|\boldsymbol{a}_k - \hat{\boldsymbol{a}}_k\|}{2L} + \frac{|b_k - \hat{b}_k|}{2t} \big\}$. Then, we get

$$\begin{aligned} |\mathcal{Z}(f, \hat{f})| &= \big| \big(f(\boldsymbol{\mathcal{X}}) + \hat{f}(\boldsymbol{\mathcal{X}}) - 2\mathbb{E}[\mathcal{Y}] - 2(\mathcal{Y} - \mathbb{E}\mathcal{Y}) \big) \big(f(\boldsymbol{\mathcal{X}}) - \hat{f}(\boldsymbol{\mathcal{X}}) \big) \big| \\ &\leq \Big(2L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + 2t + 2|\mathcal{Y} - \mathbb{E}\mathcal{Y}| \Big) \big| f(\boldsymbol{\mathcal{X}}) - \hat{f}(\boldsymbol{\mathcal{X}}) \big| \\ &\leq G(\boldsymbol{\mathcal{X}}, \mathcal{Y})\, \psi(f, \hat{f}) \, , \end{aligned}$$

with $G(\boldsymbol{\mathcal{X}}, \mathcal{Y}) \doteq 4\big(L\,\|\boldsymbol{\mathcal{X}} - \mathbb{E}\boldsymbol{\mathcal{X}}\| + t + |\mathcal{Y} - \mathbb{E}\mathcal{Y}|\big)^2$, which is Condition (C1).

To prove Condition (C3), we use Lemma 3.5 for the squared loss ℓ_{sq} and the regression function f_*. To use the lemma, we have $\|\mathcal{W}_f\|_{\Psi_2} \le t + 3LB$, $\|\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}})\|_{\Psi_2} \le \sigma$, $r_0 = 3LB$ implying $Q_n \le \ln(n)$, so we obtain Condition (C3) for any $\theta \ge 240 Q_n \max\{t + 3LB, \sigma\}^2$.

Further, we have to bound the entropy $\mathcal{H}_\psi(\epsilon, \mathcal{F})$, and $\mathbb{C}_{1/\theta}[G(\boldsymbol{\mathcal{X}}, \mathcal{Y})]$. For the entropy, notice that the radius of $\big\{\frac{\boldsymbol{a}_k}{2L} : \|\boldsymbol{a}_k\| \le L\big\}$ with respect to $\|\cdot\|_{\Psi_2}$ is bounded by $1/2$, and the same is true for $\big\{\frac{b_k - \mathbb{E}\mathcal{Y}}{2t} : |b_k - \mathbb{E}\mathcal{Y}| \le t\big\}$ using the metric $|\cdot|$. Hence, we can take $\epsilon/2$-covers of these sets for all $k = 1, \dots, K$ and use Lemma C.1 to show $\mathcal{H}_\psi(\epsilon, \mathcal{F}) \le K(d+1)\ln(3/\epsilon)$ for all $\epsilon \in (0, 3]$. Next, to bound $\mathbb{C}_{1/\theta}[G(\boldsymbol{\mathcal{X}}, \mathcal{Y})]$, observe that $\|\mathcal{Y} - \mathbb{E}\mathcal{Y}\|_{\Psi_2} \le 2LB + \sigma$ holds by (5.8). Then, $\|\sqrt{G(\boldsymbol{\mathcal{X}}, \mathcal{Y})}\|_{\Psi_2} \le 2(3LB + t + \sigma)$, which implies that $\mathbb{C}_{1/\theta}[G(\boldsymbol{\mathcal{X}}, \mathcal{Y})] \le \theta$ for any $\theta \ge 4(3LB + t + \sigma)^2$.

Finally, we are ready to apply Theorem 3.2 and putting the pieces together. For this, we set $\delta \doteq \epsilon$ (so ignoring Condition (C2) with $S \doteq \infty$), $r \doteq 1/2$, $K \doteq \lceil n^{d/(d+4)} \rceil$, $\epsilon \doteq n^{-4/(d+4)} = \Theta(K/n) = \Theta(K^{-4/d})$, and use $r_0 = 3LB$, $\theta = O(t^2) = O\big(\max\{LB, \sigma\}^2 \ln(n/\gamma)\big)$, Lemma 5.4 with $B_* = O\big(\theta\, K^{-4/d}\big)$, Lemma 5.5, and the bounds $\mathcal{H}_\psi(\epsilon, \mathcal{F}) = O\big(dK \ln(1/\epsilon)\big)$, $\mathbb{C}_{1/\theta}[G(\boldsymbol{\mathcal{X}}, \mathcal{Y})] \le \theta$, to get with probability at least $1 - \gamma$ that

$$
\begin{aligned}
L_\mu(f_n, f_*) &\le 2\left(\frac{\theta \mathcal{H}_\psi(\epsilon, \mathcal{F})}{n} + 16\epsilon\, \mathbb{C}_{\frac{1}{\theta}}[G(\boldsymbol{\mathcal{X}}, \mathcal{Y})] + B_*\right) + \frac{r_0 + 4\theta \ln(4/\gamma)}{n} \\
&= O\left(\frac{\theta d K \ln(1/\epsilon)}{n} + \theta K^{-4/d} + \frac{LB + \theta \ln(1/\gamma)}{n}\right),
\end{aligned}
$$

which proves the claim with $d \ln(1/\epsilon) = \frac{4d}{d+4} \ln(n) = O\big(\ln(n)\big)$. ■

Considering the minimax lower bound of Theorem 5.1, Theorem 5.6 proves a near-minimax rate for α-ERM$\big(\mathcal{F}_{\text{ma}}^{K,L}(\overline{\boldsymbol{\mathcal{X}}})\big)$ max-affine estimates on the class of convex sub-Gaussian regression problems. This result extends Theorem 4.2 of Balázs et al. (2015) by replacing the bounded domain $\mathbb{X}$ with a sub-Gaussian one, dropping the uniform boundedness constraint on the hypothesis class and the regression function, and providing a probabilistic guarantee instead of the weaker expected value result (3.2).

Chapter 6

Computation of max-affine estimators

In this chapter, we discuss the computational properties of multivariate max-affine estimators represented by at most $K \in \mathbb{N}$ hyperplanes as $f_n(\boldsymbol{x}) = \max_{k=1,\dots,K} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k$, $\boldsymbol{x} \in \mathbb{R}^d$ where the parameters $\{K, (\boldsymbol{a}_k, b_k) : k = 1, \dots, K\}$ are trained on the dataset $\mathcal{D}_n = \{(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) : i = 1, \dots, n\}$.

To discuss the training algorithms, we will use four examples, referred to as *discussion problems*. Two of them have smooth quadratic regression functions $(f_*^{\mathrm{fq}}, f_*^{\mathrm{hq}})$, while $f_*^{\mathrm{lfq}}, f_*^{\mathrm{lhq}}$ are their linearized variants:

$$
\begin{aligned}
f_*^{\mathrm{fq}}(\boldsymbol{x}) &\doteq (1/2)\boldsymbol{x}^\top H_* \boldsymbol{x} + \boldsymbol{f}_*^\top \boldsymbol{x} + c_* \,, \\
f_*^{\mathrm{hq}}(\boldsymbol{x}) &\doteq (1/2)\boldsymbol{x}_+^\top H_* \boldsymbol{x}_+ \,, \quad \boldsymbol{x}_+ \doteq \max\{\boldsymbol{0}, \boldsymbol{x}\} \,, \\
f_*^{\mathrm{lfq}}(\boldsymbol{x}) &\doteq \max_{k=1,\dots,K_*} \nabla f_*^{\mathrm{fq}}(\boldsymbol{x}_k^*)^\top \boldsymbol{x} + f_*^{\mathrm{fq}}(\boldsymbol{x}_k^*) \,, \\
f_*^{\mathrm{lhq}}(\boldsymbol{x}) &\doteq \max_{k=1,\dots,K_*} \nabla f_*^{\mathrm{hq}}(\boldsymbol{x}_k^*)^\top \boldsymbol{x} + f_*^{\mathrm{hq}}(\boldsymbol{x}_k^*) \,.
\end{aligned}
\tag{6.1}
$$

Here $\boldsymbol{x} \in \mathbb{R}^d$, $d \in \mathbb{N}$, $0 \preceq H_* \in \mathbb{R}^{d \times d}$, $\boldsymbol{f}_* \in \mathbb{R}^d$, $c_* \in \mathbb{R}$, and $K_* \in \mathbb{N}$, $\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_{K_*}^* \in \mathbb{R}^d$ are all fixed.[1] For all problems, the data points are drawn from the uniform distribution $\mathcal{U}(\mathbb{X})$ over a bounded convex domain $\mathbb{X} \subseteq \mathbb{R}^d$ and the noise model is simply Gaussian, that is $\mathcal{Y} - f_*(\boldsymbol{\mathcal{X}}) \sim \mathcal{N}(0, \sigma^2)$.[2] We also rotate each problem instance by a random orthogonal matrix (Stewart, 1980, Section 3), that is replace f_* by $\boldsymbol{x} \mapsto f_*(Q^\top \boldsymbol{x})$ on each run, which

[1] We use $h_{k,l}^* \doteq \mathbb{I}\{k = l\} + (k + l)^{-1}$, $f_k^* \doteq k/d$, $c_* \doteq -d$ and $H_* = [h_{k,l}^*]_{k,l=1,\dots,d}$, $\boldsymbol{f}_* = [f_1^*, \dots, f_k^*]^\top$. Further, we define $K_* \doteq d + 3$ as $\boldsymbol{x}_k^* \doteq \boldsymbol{e}_k - (k/d)\boldsymbol{1}$ for all $k = 1, \dots, d$, and $\boldsymbol{x}_{d+1}^* \doteq \boldsymbol{0}$, $\boldsymbol{x}_{d+2}^* \doteq -\boldsymbol{1}$, $\boldsymbol{x}_{d+3}^* \doteq \boldsymbol{1}$, where $\boldsymbol{e}_k$ is the k-th standard basis in $\mathbb{R}^d$.

[2] We set $\mathbb{X} \doteq [-2, 2]^d$ and $\sigma \doteq 1$.

increases the variance of the results for algorithms that are not rotationally invariant.

We emphasize that the problems (6.1) were selected to illustrate some major issues of max-affine estimators. The smooth quadratic target f_*^{fq} is "difficult" to represent "accurately" by a max-affine function, while its truncated version f_*^{hq} has a large flat plateau maximizing the chance of overfitting the noise at the boundary, similar to the case shown on Figure 5.1. Finally, a "good" max-affine estimator should be able to improve the rate for the linearized versions f_*^{fq}, f_*^{hq} when the regression function can be represented by only a few hyperplanes.

For all of our numerical experiments, the hardware has a Dual-Core AMD Opteron(tm) Processor 250 (2.4GHz, 1KB L1 Cache, 1MB L2 Cache) with 8GB RAM, which scored around 2 GFLOPS on the benchmark of Moler (1994) using randomly generated dense matrices of size 15000×15000. The software uses MATLAB (The MathWorks, Inc., 2010), sometimes C (Stallman et al., 2007), and the MOSEK Optimization Toolbox (MOSEK ApS, 2015).

6.1 Max-affine LSEs

Recall from Section 5.1 that one can find max-affine LSEs with at most n hyperplanes by reformulating the non-convex max-affine optimization problem as a convex quadratically constrained quadratic program (QCQP),

$$\min_{\substack{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n,\\ b_1,\ldots,b_n}} \sum_{i=1}^{n} \Big(\max_{j=1,\ldots,n} \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j - \mathcal{Y}_i \Big)^2 \tag{6.2}$$

$$\text{such that for all } j = 1,\ldots,n : \|\boldsymbol{a}_j\| \le L\,,$$

$$= \min_{\substack{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n,\\ b_1,\ldots,b_n}} \sum_{i=1}^{n} (\boldsymbol{a}_i^\top \boldsymbol{\mathcal{X}}_i + b_i - \mathcal{Y}_i)^2 \quad \text{such that for all } i,j = 1,\ldots,n : \tag{6.3}$$

$$\boldsymbol{a}_i^\top \boldsymbol{\mathcal{X}}_i + b_i \ge \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j\,, \quad \|\boldsymbol{a}_i\| \le L\,.$$

The resulting estimate $f_n(\boldsymbol{x}) = \max_{i=1,\ldots,n} \boldsymbol{a}_i^\top \boldsymbol{x} + b_i$ enjoys the worst case excess risk upper bound (5.2), and (6.3) can be solved in $O(d^2 n^5)$ time using interior

point methods (Boyd and Vandenberghe, 2004, Section 11.5).[3] Together, the risk guarantee and the polynomial computation time makes this estimator attractive. However, even if the computational cost is only a worst-case upper bound on dense QCQP problems (so it is unlikely to be tight here), problem (6.3) still gets too large for today's QCQP solvers even for modest sample sizes.[4] This motivated research on alternative solution techniques such as cutting plane methods (Section 6.1.2), and alternating direction method of multipliers (ADMM) algorithms (Section 6.1.3).

6.1.1 Partitioned max-affine LSEs

Before discussing solution methods, we generalize problem (6.3) to find hyperplanes over an arbitrary fixed partition of the data points (Balázs et al., 2015, Section 4.4). This problem is also convex and will be useful later for partitioning max-affine estimators (Section 6.2).

Fix a partition $P \doteq \{\mathcal{C}_k : k = 1, \dots, K\}$ of the index set $\{1, \dots, n\}$, that is $\cup_{k=1}^{K} \mathcal{C}_k = \{1, \dots, n\}$ and $\mathcal{C}_k \cap \mathcal{C}_j = \emptyset$ if and only if $k \neq j$ for all $k, j = 1, \dots, K$. We say that a max-affine function $f(\boldsymbol{x}) = \max_{k=1,\dots,K} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k$ *induces* P if $\boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k \geq \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j$ holds for all $k, j = 1, \dots, K$ and $i \in \mathcal{C}_k$. Then consider the following QCQP problem of finding max-affine LSEs inducing P:

$$\begin{aligned} \min_{\substack{\boldsymbol{a}_1,\dots,\boldsymbol{a}_K,\\ b_1,\dots,b_K}} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} (v_i - \mathcal{Y}_i)^2 \quad & \text{such that for all } k, j = 1, \dots, K,\ i \in \mathcal{C}_k : \\ & v_i = \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k \geq \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j, \quad \|\boldsymbol{a}_k\| \leq L\,. \end{aligned} \tag{6.4}$$

Notice that if $K = n$ and $P = P_{1:n} \doteq \big\{\{1\}, \dots, \{n\}\big\}$, then the two problems (6.3) and (6.4) are identical, so the latter problem is indeed more general. Hence, it is enough to discuss solution methods for (6.4) in the following sections and run the algorithms with $P = P_{1:n}$ for solving (6.3).

As shown by Theorem 5.6, there exists a near-optimal max-affine estimator with at most $\lceil n^{d/(d+4)} \rceil$ hyperplanes. If we knew a partition P induced by such

[3] An upper bound on the number of arithmetic operations needed to solve a convex QCQP problem, $\min_{\boldsymbol{x} \in \mathbb{R}^N} \frac{1}{2}\boldsymbol{x}^\top H \boldsymbol{x} + \boldsymbol{f}^\top \boldsymbol{x}$ subject to $\frac{1}{2}\boldsymbol{x}^\top Q_j \boldsymbol{x} + \boldsymbol{g}_j^\top \boldsymbol{x} \leq c_j$ for $j = 1, \dots, M$ with $0 \preceq H, Q_1, \dots, Q_M$, is $O(N^2 M^{1.5})$.

[4] For our hardware/software setup (page 62), direct solution of (6.3) took the QCQP solver 34–74 times more time for $d = 8$ and $n = 500$ on problems (6.1) than for alternative techniques, and run out of the 8GB memory for $n = 1000$.

a near-optimal estimator, we could compute the estimator itself by solving problem (6.4). Unfortunately, finding such a partition is far from trivial, and according to our knowledge, there is no algorithm today that could do this efficiently. Still, since every max-affine function induces a partition, (6.4) can be useful for improving partitioning max-affine estimators (Section 6.2) by providing a tool to efficiently find the best possible fit for their induced partitions, justifying the study of this problem.

6.1.2 Cutting plane methods

The key observation for solving (6.4) using cutting plane (CP) techniques (Lee et al., 2013) is that the hyperplane constraints $\boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k \geq \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j$ are often redundant in the sense that by solving

$$\begin{aligned} \min_{\substack{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_K,\\ b_1,\ldots,b_K}} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} (v_i - \mathcal{Y}_i)^2 \quad & \text{such that for all } (k,i,j) \in \mathcal{I}: \\ & v_i = \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k \geq \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j, \quad \|\boldsymbol{a}_k\| \leq L\,, \end{aligned} \tag{6.5}$$

we often obtain the same solution as for (6.4) with a well-chosen constraint set $\mathcal{I} \subset \mathcal{I}_* \doteq \{(k,i,j) : k,j = 1,\ldots,K,\ k \neq j,\ i \in \mathcal{C}_k\}$ having $|\mathcal{I}| \ll n(K-1)$.

In particular for the univariate case ($d = 1$), one can order the cells (non-overlapping intervals) by some permutation $l_1,\ldots,l_K$ of $1,\ldots,K$, and verify the hyperplane constraints $\boldsymbol{a}_{l_k}^\top \boldsymbol{\mathcal{X}}_i + b_{l_k} \geq \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j$, $i \in \mathcal{C}_{l_k}$ only for the neighbors $j \in \{l_k - 1, l_k + 1\}$, so using a constraint set of size $|\mathcal{I}| = 2(K-1)$. However, ordering the cells is not possible for the multivariate case ($d \geq 2$), and there can be hyperplanes with arbitrary number of neighbors. Still, in many cases it might be enough to check the hyperplane constraints for fewer than $n-1$ other points in $\{\boldsymbol{\mathcal{X}}_i : i \notin \mathcal{C}_k\}$ for each cell $k = 1,\ldots,K$. For example, consider estimating a truncated Euclidean cone over a 2 dimensional domain $\mathbb{X}$ as shown by Figure 6.1, where the base hyperplane can have arbitrary many neighbors, but the radial hyperplanes have only three.

Unfortunately, finding a set $\mathcal{I}$ with the fewest elements that makes (6.5) equivalent to (6.4) is far from trivial, hence we resort to heuristic methods. For this, we consider the CP scheme of Lee et al. (2013) adapted to (6.5) by

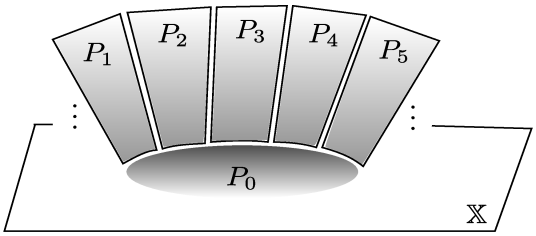

Figure 6.1: For the estimation of a truncated Euclidean cone over a 2 dimensional domain $\mathbb{X}$, the base hyperplane (P_0) can have arbitrary many neighbors, while the radial ones (P_i, $i > 0$) have exactly three.

Balázs et al. (2015). The scheme shown as Algorithm 6.1, increases the size of the constraint set $\mathcal{I}$ by at most K in every iteration. In order to select the

1. **input:** training set $\mathcal{D}_n$, Lipschitz bound L, partition P
2. $\mathcal{I} \leftarrow \emptyset$
3. **repeat**
4. solve (6.5) using $\mathcal{I}$, obtain $\boldsymbol{a}_1, \dots, \boldsymbol{a}_K$ and $b_1, \dots, b_K$
5. **for** $k = 1, \dots, K$ **do**
6. $\mathcal{I}_k \leftarrow \left\{(i,j) : i \in \mathcal{C}_k,\ j = 1, \dots, K : \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k < \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j\right\}$
7. when $\mathcal{I}_k \neq \emptyset$, choose $(i,j) \in \mathcal{I}_k$ and put (k,i,j) into $\mathcal{I}$
8. **end for**
9. **until** exists $(k,i,j) \in \mathcal{I}_*$ such that $\boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k < \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j$
10. **output:** hyperplane slopes $\boldsymbol{a}_1, \dots, \boldsymbol{a}_K$, and heights $b_1, \dots, b_K$

Algorithm 6.1: Cutting plane (CP) algorithm training a max-affine LSE with K hyperplanes over a fixed partition P.

constraint (k,i,j) in step 7 of Algorithm 6.1, we consider two heuristics. The first one choosing the most violated constraint (6.6), while the second divides the violation severity value by the distance of the violated location to the cell center related to the violating constraint (6.7). Formally, for the first heuristic

$$(i,j) \in \operatorname*{argmax}_{(i',j') \in \mathcal{I}_k} (\boldsymbol{a}_{j'} - \boldsymbol{a}_k)^\top \boldsymbol{\mathcal{X}}_{i'} + b_{j'} - b_k \,, \tag{6.6}$$

while for the second

$$(i,j) \in \operatorname*{argmax}_{(i',j') \in \mathcal{I}_k} \left((\boldsymbol{a}_{j'} - \boldsymbol{a}_k)^\top \boldsymbol{\mathcal{X}}_{i'} + b_{j'} - b_k\right) \|\boldsymbol{\mathcal{X}}_{i'} - \bar{\boldsymbol{\mathcal{X}}}_{j'}\|^{-1} \,, \tag{6.7}$$

where $\bar{\boldsymbol{\mathcal{X}}}_j \doteq \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} \boldsymbol{\mathcal{X}}_i$ is the center of the j-th cell. The idea behind the second method (6.7) is that choosing closer points is more likely to fix many others being further away, similar as neighbors did for the 1-dimensional case.

We compare these heuristics on the four discussion problems (6.1) when $P = P_{1:n}$ and a problem specific tight Lipschitz upper bound $L = L_*$. The results are summarized by Figure 6.2, which shows that (6.7) constructs a smaller constraint index set $\mathcal{I}$ and performs fewer number of iterations, which makes it slightly more scalable with the sample size n. The result also shows

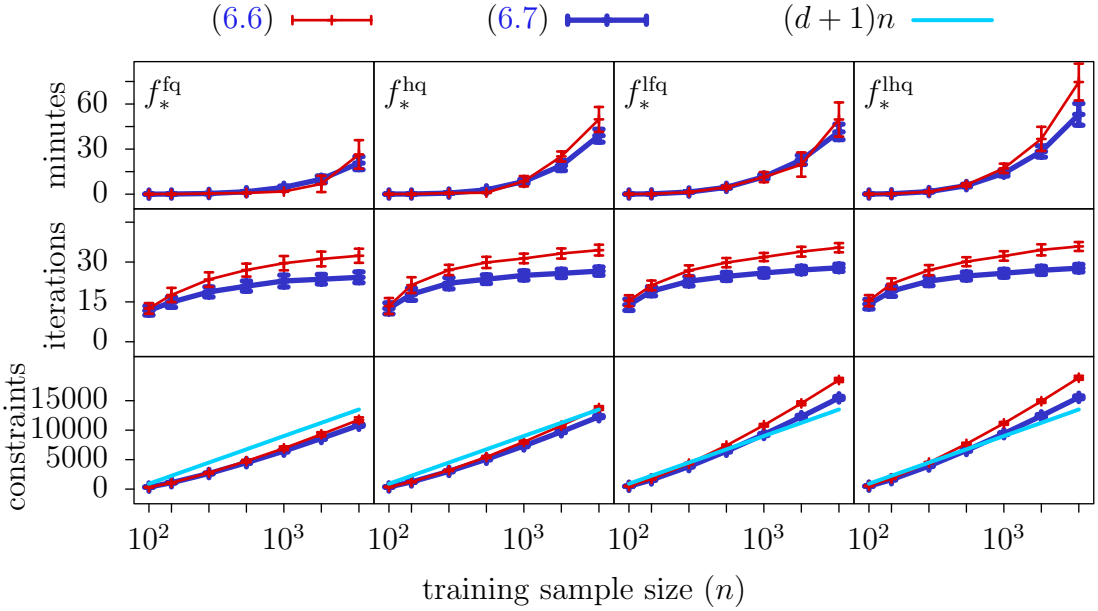

Figure 6.2: Empirical comparison of constraint selection heuristics for the max-affine LSE computed by cutting plane methods on the discussion problems $f_*^{\mathrm{fq}}, f_*^{\mathrm{hq}}, f_*^{\mathrm{lfq}}, f_*^{\mathrm{lhq}}$. The dimension is $d = 8$ and the sample sizes are $n = 100, 250, 500, 750, 1000, 1250, 1500$. Averages of 100 experiments with standard deviation error bars are shown for the training times (in minutes), the number of iterations, and the number of constraints used in the last iteration of the algorithm.

that the number of constraints necessary for solving (6.3) is far less than $O(n^2)$, in fact an $O(dn)$ bound looks reasonable for these examples. However, despite of this reduction of the computational effort, the training times presented by Figure 6.2 indicate that the CP methods scale poorly with the sample size n. The reason is that interior-point QCQP solvers form and factorize a sparse Hessian matrix, and this calculation eventually becomes too slow even with $O(dn)$ constraints. Hence, we consider a Hessian-free optimization

method next, which is capable to trade numerical accuracy for computational efficiency.

6.1.3 Alternating direction methods of multipliers

As an alternative to cutting plane (CP) methods, one can use alternating direction method of multipliers (ADMM) algorithms, which are easier to scale as the problem size grows by parallelization and trading numerical accuracy for computational efficiency. ADMM algorithms lie in the intersection of augmented Lagrangian methods (see for example Bertsekas, 1996), and decomposed optimization techniques (see for example Bertsekas and Tsitsiklis, 1997), and inherit their advantageous properties such as weak convergence requirements by making the dual problem differentiable, and efficient parallelization (Boyd et al., 2010). We are not the first to consider ADMM to solve (6.3): a decomposition method was previously proposed by Aybat and Wang (2014), and an ADMM algorithm was considered by Mazumder et al. (2015). Here we adapt the latter ADMM algorithm to solve the partitioned LSE problem (6.4).

Consider the augmented Lagrangian function of (6.4) defined as

$$\begin{aligned}\mathcal{L}_\rho(A, \boldsymbol{b}, \boldsymbol{v}, S, \eta) \doteq & \frac{1}{2}\sum_{i=1}^{n}(v_i - \mathcal{Y}_i)^2 \\ & + \sum_{i=1}^{n}\sum_{k=1}^{K}\eta_{ik}\left(s_{ik} + \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k - v_i\right) \\ & + \frac{\rho}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(s_{ik} + \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k - v_i\right)^2,\end{aligned} \tag{6.8}$$

where $A \doteq (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K)$, $\boldsymbol{b} \doteq (b_1, \ldots, b_K)$, $\boldsymbol{v} \doteq (v_1, \ldots, v_n)$, $\rho > 0$ is the penalty parameter, $S \doteq (s_{ik})_{i=1,\ldots,n,k=1,\ldots,K}$ are slack variables such that $s_{ik} \geq 0$ with $s_{ik} \doteq 0$ if $i \in \mathcal{C}_k$, and $\eta = (\eta_{ik})_{i=1,\ldots,n,k=1,\ldots,K}$ are the dual variables. The ADMM algorithm performs a primal-dual optimization of the augmented Lagrangian function just like augmented Lagrangian methods, further decomposing the primal step into blockwise operations in a way that the subproblems admit efficient (often closed form) solutions, also allowing easy parallelization. In order to solve problem (6.4), an ADMM algorithm can consider the

following subproblems in each iteration:

$$A \leftarrow \operatorname*{argmin}_{A} \mathcal{L}_\rho(A, \boldsymbol{b}, \boldsymbol{v}, S, \eta) \tag{6.9}$$
$$\text{such that } \|\boldsymbol{a}_k\| \le L \text{ for all } k = 1, \ldots, K\,,$$

$$(\boldsymbol{b}, \boldsymbol{v}) \leftarrow \operatorname*{argmin}_{\boldsymbol{b},\boldsymbol{v}} \mathcal{L}_\rho(A, \boldsymbol{b}, \boldsymbol{v}, S, \eta)\,, \tag{6.10}$$

$$S \leftarrow \operatorname*{argmin}_{S \ge 0} \mathcal{L}_\rho(A, \boldsymbol{b}, \boldsymbol{v}, S, \eta) \text{ such that } s_{ik} = 0 \text{ if } i \in \mathcal{C}_k \tag{6.11}$$
$$\text{for all } i = 1, \ldots, n,\ k = 1, \ldots, K\,,$$

$$\eta_{ik} \leftarrow \eta_{ik} + \rho\big(s_{ik} + \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k - v_i\big) \tag{6.12}$$
$$\text{for all } i = 1, \ldots, n,\ k = 1, \ldots, K\,.$$

This can be transformed into Algorithm 6.2 shown below by observing that problem (6.9) can be decomposed componentwise solving for each $\boldsymbol{a}_k$ separately, problems (6.10) and (6.11) admit closed-form solutions,[5] and by introducing the scaled dual-variables $\hat{\eta}_{ik} \doteq \eta_{ik}/\rho$.

Although the optimization problem (6.9) in step 4 of Algorithm 6.2 does not admit a closed-form solution in general when $L < \infty$, it has been well-studied for trust-region optimization algorithms and can be solved efficiently with high-accuracy by a variant of Newton's root-finding method such as Algorithm 4.3 of Nocedal and Wright (2006). Hence, the iteration cost of Algorithm 6.2 is dominated by $O(nK)$ in the usual case when $n \ge d^2$ (not counting the $O(d^3)$ factorization cost of matrix $\sum_{i=1}^n \boldsymbol{\mathcal{X}}_i \boldsymbol{\mathcal{X}}_i^\top$, which can be done offline).

Notice that we only used an iteration limit t_* as a termination condition for Algorithm 6.2. The reason is that although ADMM algorithms are guaranteed to converge to the global optimum, their convergence is quite slow and as such it becomes impractical to wait for convergence to "happen". Hence, instead of using optimality conditions to derive stopping rules (Boyd et al., 2010, Section 3.3), we terminate after a fixed number of iterations. Because of this, the empirical risk using ADMM is often larger than the excess risk of the

[5] To solve the quadratic optimization (6.10) for variable $[\boldsymbol{b}\ \boldsymbol{v}]^\top$, use the inverse Hessian matrix $\begin{bmatrix} n\rho\, I_K & -\rho\, \mathbf{1}_{K\times n} \\ -\rho\, \mathbf{1}_{n\times K} & (1+K\rho) I_n \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n\rho} I_K + \frac{1}{n}\mathbf{1}_{K\times K} & \frac{1}{n}\,\mathbf{1}_{K\times n} \\ \frac{1}{n}\,\mathbf{1}_{n\times K} & \frac{1}{1+K\rho}\big(I_n + \frac{K\rho}{n}\mathbf{1}_{n\times n}\big) \end{bmatrix}$. Then unvectorize the solution and use the z_i, w_k quantities for all $k = 1, \ldots, K$ and $i = 1, \ldots, n$ as defined by steps 5 and 6 in Algorithm 6.2 to form the solution (shown by steps 7 and 8).

1. **input:** training set $\mathcal{D}_n$, Lipschitz bound L, partition P,
   dual stepsize ρ, iteration number t_*
2. initialize $\boldsymbol{a}_k^0$, b_k^0, v_i^0, s_{ik}^0, $\hat{\eta}_{ik}^0$ appropriately (for example to zero)
3. **for** $t = 0, 1, \ldots, t_* - 1$ **do**
4. $\quad \boldsymbol{a}_k^{t+1} \leftarrow \underset{\boldsymbol{a}_k : \|\boldsymbol{a}_k\| \leq L}{\operatorname{argmin}} \frac{1}{2} \boldsymbol{a}_k^\top \Big(\sum_{i=1}^n \boldsymbol{\mathcal{X}}_i \boldsymbol{\mathcal{X}}_i^\top \Big) \boldsymbol{a}_k + \sum_{i=1}^n \big(\hat{\eta}_{ik}^t + s_{ik}^t + b_k^t - v_i^t \big) \boldsymbol{\mathcal{X}}_i^\top \boldsymbol{a}_k$
   for all $k = 1, \ldots, K$
5. $\quad z_i \leftarrow \frac{\mathcal{Y}_i}{\rho} + \sum_{k=1}^n \big(\hat{\eta}_{ik}^t + s_{ik}^t + \boldsymbol{\mathcal{X}}_i^\top \boldsymbol{a}_k^{t+1} \big)$ for all $i = 1, \ldots, n$
6. $\quad w_k \leftarrow - \sum_{i=1}^n \big(\hat{\eta}_{ik}^t + s_{ik}^t + \boldsymbol{\mathcal{X}}_i^\top \boldsymbol{a}_k^{t+1} \big)$ for all $k = 1, \ldots, K$
7. $\quad b_k^{t+1} \leftarrow \frac{w_k}{n} + \frac{\rho}{n} \Big(\sum_{i=1}^n z_i + \sum_{k=1}^K w_k \Big)$ for all $k = 1, \ldots, K$
8. $\quad v_i^{t+1} \leftarrow \frac{\rho}{1 + K\rho} \Big(z_i + \frac{K\rho}{n} \sum_{i=1}^n z_i \Big) + \frac{\rho}{n} \sum_{k=1}^K w_k$ for all $i = 1, \ldots, n$
9. $\quad s_{ik}^{t+1} \leftarrow \max \big\{ 0, -\big(\hat{\eta}_{ik}^t + \boldsymbol{\mathcal{X}}_i^\top \boldsymbol{a}_k^{t+1} + b_k^{t+1} - v_i^{t+1} \big) \big\} \mathbb{I}\{ i \notin \mathcal{C}_k \}$
   for all $i = 1, \ldots, n$ and $k = 1, \ldots, K$
10. $\quad \hat{\eta}_{ik}^{t+1} \leftarrow \hat{\eta}_{ik}^t + \big(s_{ik}^{t+1} + \boldsymbol{\mathcal{X}}_i^\top \boldsymbol{a}_k^{t+1} + b_k^{t+1} - v_i^{t+1} \big)$
    for all $i = 1, \ldots, n$ and $k = 1, \ldots, K$
11. **end for**
12. **output:** hyperplane slopes $\boldsymbol{a}_1^{t_*}, \ldots, \boldsymbol{a}_K^{t_*}$, and heights $b_1^{t_*}, \ldots, b_K^{t_*}$

Algorithm 6.2: Alternating direction methods of multipliers (ADMM) algorithm training a max-affine LSE with K hyperplanes over a fixed partition P.

estimator computed by the CP method (such as Algorithm 6.1 combined with an interior-point algorithm for step 7). However, besides the computational cost of ADMM scales better in terms of the dimension d and the sample size n, stopping ADMM early seems to significantly decrease the excess risk as well, an effect that was studied previously by a number of authors, including Zhang and Yu (2005) (in the context of boosting) and Yao et al. (2007) (in the context of gradient descent and linear prediction).

Parameter ρ has two roles in Algorithm 6.2 as scaling the penalty of the primal objective and being the step size of the dual update. Because we expect ADMM to slow down around constraint violation magnitude 10^{-4}, we use $\rho \doteq 0.01$, which seems to serve the two roles reasonably. We did not observe any improvement to this by using an adaptive technique for setting ρ as surveyed in Boyd et al. (2010, Section 3.4.1).

We compare ADMM and CP methods on the discussion problems (6.1)

and present the results by Figure 6.3. For this, we use ADMM with $P = P_{1:n}$,

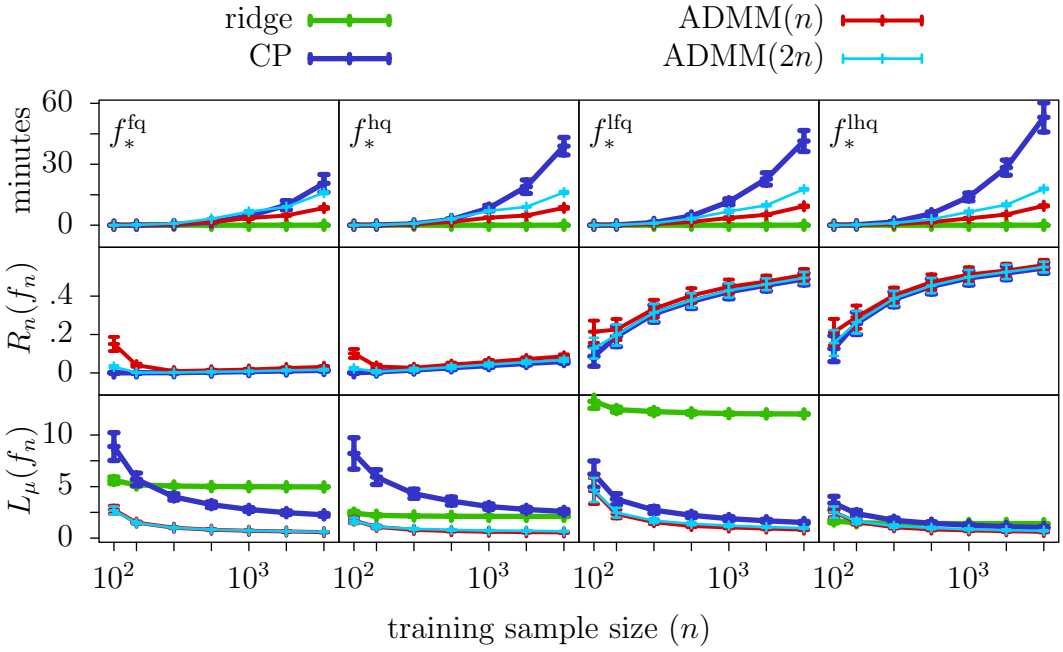

Figure 6.3: Comparison of linear and max-affine LSEs computed by cutting plane (CP) method and ADMM, running for n or $2n$ iterations on the discussion problems $f_*^{\mathrm{fq}}, f_*^{\mathrm{hq}}, f_*^{\mathrm{lfq}}, f_*^{\mathrm{lhq}}$. The dimension is $d = 8$ and the sample sizes are $n = 100, 250, 500, 750, 1000, 1250, 1500$. Averages of 100 experiments with standard deviation error bars are shown for the training times (in minutes), the empirical risk $R_n(f_n)$, and the excess risk $L_\mu(f_n)$ with $\ell = \ell_{\mathrm{sq}}$ measured on 10^6 new samples for each experiment.

a problem specific tight Lipschitz bound $L = L_*$, and terminate it after $t_* = n$ or $t_* = 2n$ iterations.[6] Furthermore, we use CP with $P = P_{1:n}$, $L = L_*$, and the constraint selection method (6.7) only, but the empirical and excess risks are very similar for (6.6) as well. Finally, we also include ridge regression (see Section 4.2.2) with a small regularization parameter $\beta = 10^{-6}$ for numerical stability to serve as a baseline. As Figure 6.3 shows, ADMM indeed scales better in terms of the sample size n by losing some on the training accuracy. More significantly, observe that the CP method provides a significantly worse excess risk than even ridge regression on problem f_*^{hq}, which indicates serious overfitting on the large flat plateau of the regression function. Thus, the less

[6] We also center the $\boldsymbol{\mathcal{X}}_i$, $\mathcal{Y}_i$ values and scale them by $\max_{i=1,\ldots,n} \|\boldsymbol{\mathcal{X}}_i\|$, and $\max_{i=1,\ldots,n} |\mathcal{Y}_i|$, respectively.

accurate solution of ADMM enjoys significantly smaller excess risk, and even doubling the number of iterations in the stopping conditions makes a little difference.

To conclude, it seems that early stopping of ADMM reduces the chance of overfitting, which is a great advantage compared to CP methods. To understand this effect better, in the next section we experiment with a problem where the upper bound on the Lipschitz factor is dropped (that is $L = \infty$), for which we expect that the chance of overfitting is much larger.

6.1.4 Training without the Lipschitz factor

So far we have used a tight Lipschitz upper bound L for each regression problem, but this value is often not available in practice. Dropping L, that is solving (6.3) with $L = \infty$ accurately, would usually lead to an estimator with a very large excess risk on many noisy problems such as (6.1). The cause of overfitting is that the slope of this estimator can grow unbounded and perfectly fit the training data at the boundary, as illustrated on Figure 5.1.

In fact, recent work by Han and Wellner (2016) based on Gao and Wellner (2015) shows that without a Lipschitz bound but assuming bounded functions, the complexity of the shape of the domain of the convex function influences the convergence rate of any method. In particular, for domains with smooth boundaries the minimax rate worsens by roughly a factor of two in the exponent to $\Theta(n^{-2/(d+1)})$, a strong indication that giving up on a known Lipschitz factor may present significant challenges.

An additional complication is that even if we could learn the Lipschitz bound L correctly (for example by cross-validation), the max-affine LSE (6.3) might still suffer from serious overfitting on problems with flat and steep regions at the boundary, similarly as shown by the results for f_*^{hq} on Figure 6.3. However, recall that ADMM solution of (6.3) provided significantly better excess risk by not optimizing the empirical risk too accurately. By the next experiment, we measure this tradeoff between empirical and excess risks and the overfitting robustness of (6.3) solved by ADMM for problems (6.1).

The experiment shown by Figure 6.4 measures the training accuracy (de-

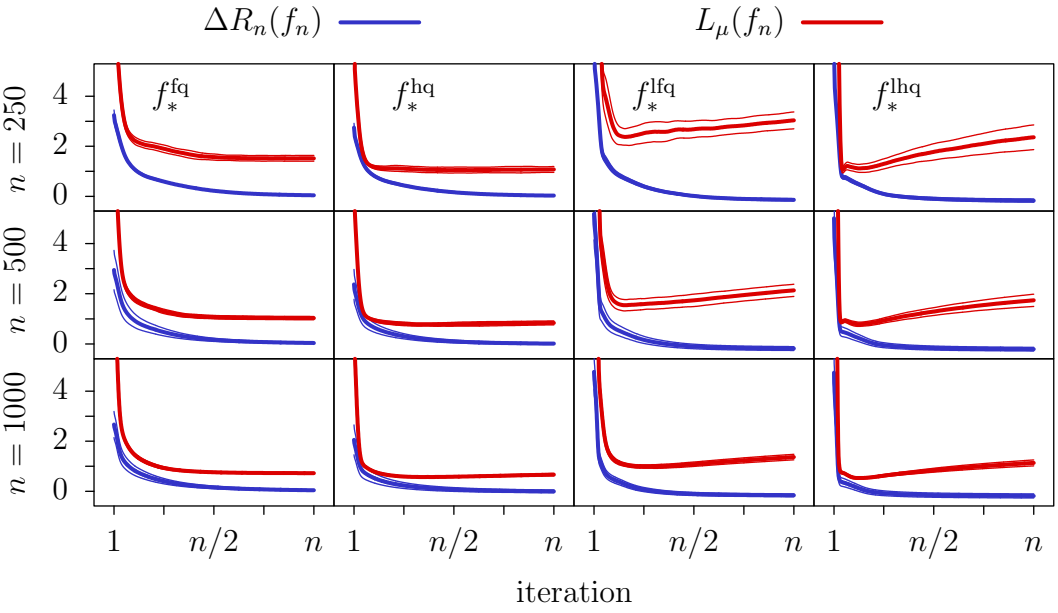

Figure 6.4: Measurements of training accuracy $\Delta R_n(f_n)$ and excess risk $L_\mu(f_n)$ as a function of the iteration number of ADMM. Averages of 25 experiments with standard deviation error ranges are shown on discussion problems $f_*^{\mathrm{fq}}, f_*^{\mathrm{hq}}, f_*^{\mathrm{lfq}}, f_*^{\mathrm{lhq}}$. The dimension is $d = 8$ and the sample sizes are $n = 250, 500, 1000$. The excess risk $L_\mu(f_n)$ is measured on 10^5 new samples for each experiment.

noted by $\Delta R_n(f_n)$) as the empirical risk difference of the ADMM and CP solutions, and the excess risk in each iteration of ADMM using $P = P_{1:n}$, $L = \infty$, and $t_* = n$. The baseline for the training accuracy is computed by the CP using $P = P_{1:n}$, (6.7), and $L = L_*$ set tightly for each problem. Observe that $\Delta R_n(f_n)$ is sometimes negative indicating that the ADMM empirical risk is smaller than the CP one due to ADMM's $L = \infty$ setting. We also note that ADMM's slope is bounded on these examples as $\max_{k=1,\dots,K} \|\boldsymbol{a}_k\| \leq 2.4\, L_*$ suggesting that a theoretical guarantee might be still valid.

Furthermore, Figure 6.4 also illustrates the overfitting behavior of ADMM. Notice that the saturation of the excess risk is "long" and overfitting is "slow", at least over the training horizon of n steps. The fact that overfitting is worse for the linearized problems f_*^{lfq} and f_*^{lhq} makes sense as the regression functions of these problems are max-affine functions with only $d + 3 = 11$ hyperplanes

implying faster convergence to the "saturation" of the excess risk and leaving more time for overfitting to accumulate. However, this also suggests that overfitting of ADMM can be controlled by the number of iterations which might be tuned by cross-validation.

6.1.5 Cross-validating ADMM algorithm

To tune the number of iterations for ADMM Algorithm 6.2, we consider cross-validation, where we train multiple ADMM instances simultaneously and measure their cross-validated error.

To describe the algorithm, we suppose that the data $\mathcal{D}_n$ is already randomly shuffled uniformly. First, the data is splitted into u_* equal subsets (the last one might be smaller) indexed with the nonempty, disjunct sets F_u, $u = 1, \ldots, u_*$ satisfying $\cup_{u=1}^{u_*} F_u = \{1, \ldots, n\}$. Then the algorithm runs u_* instances of ADMM producing estimates $f_n^1, \ldots, f_n^{u_*}$ such that the u-th instance uses all but the u-th subset for training and estimates the risk $R_\mu(f_n^u)$ on the "hold out data" indexed by F_u using $R_\mu(f_n^u) \approx \hat{R}_{F_u}(f_n^u) \doteq \frac{1}{|F_u|} \sum_{i \in F_u} (f_n^u(\boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i)^2$. In each iteration, the cross-validation error is computed by $\frac{1}{u_*} \sum_{u=1}^{u_*} \hat{R}_{F_u}(f_n^u)$ and the process is stopped if it has not been improved in the last t_{wait} steps. The final max-affine model is produced by running a single instance of ADMM on the full data set $\mathcal{D}_n$ for as many iterations as used for the lowest cross-validation error obtained previously.

The following experiment compares the CP method and ADMM, with and without using a tight problem-specific Lipschitz bound L_* and the above discussed cross-validation technique. The results are presented on Figure 6.5. Here the CP algorithm used $P = P_{1:n}$, $L = L_*$, and ADMM used $P = P_{1:n}$, $t_* = n$, and $L = L_*$ or $L = \infty$, respectively. The cross-validated ADMM (cvADMM) used $P = P_{1:n}$, $L = \infty$, and $t_{\text{wait}} = 100$ with $u_* = 5$ or $u_* = 10$, respectively (these values for u_* are chosen as a "standard choice", and t_{wait} is set to the reciprocal of the dual step size $1/\rho$). At the end of each training process, we dropped all hyperplanes that did not have any influence on the empirical risk, hence the max-affine models usually have slightly less than n hyperplanes.

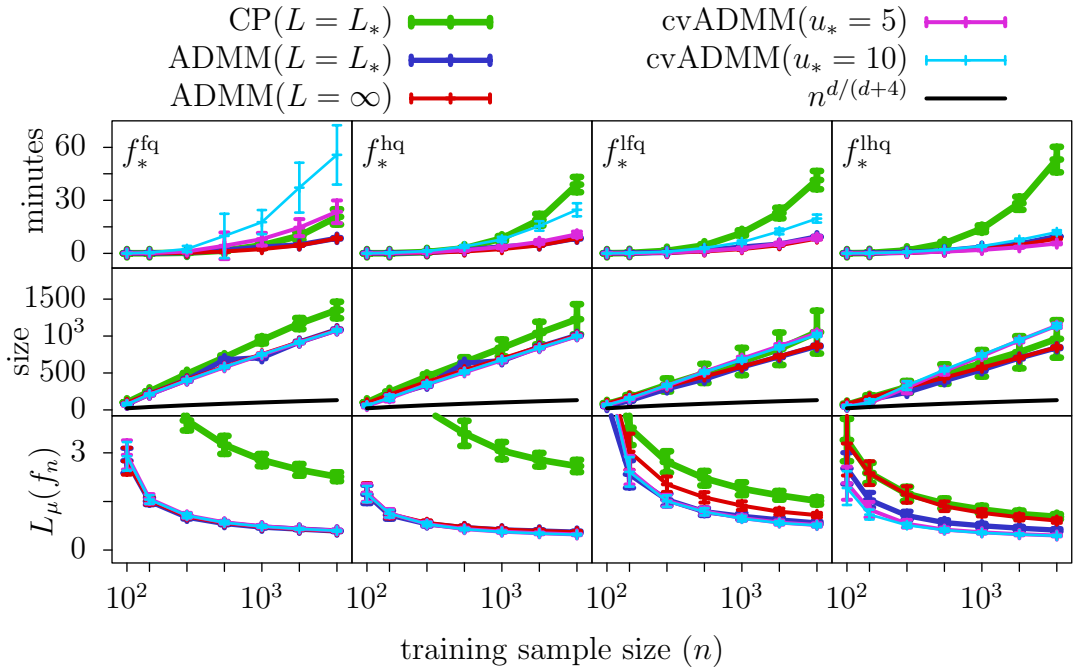

Figure 6.5: Performance of ADMM and cross-validated ADMM without using the Lipschitz bound on discussion problems f_*^{fq}, f_*^{hq}, f_*^{lfq}, f_*^{lhq}. The dimension is $d = 8$ and the sample sizes are $n = 100$, 250, 500, 750, 1000, 1250, 1500. Averages of 100 experiments with standard deviation error bars are shown for training times (in minutes), model size, and excess risk $L_\mu(f_n)$ measured on 10^6 new samples for each experiment.

Figure 6.5 shows that dropping the Lipschitz bound for ADMM makes it overfit more on the f_*^{lfq} and f_*^{lhq} problems, just as it was noted when discussing Figure 6.4 earlier. However, Figure 6.5 also confirms that the overfitting of ADMM can be "eliminated" by learning an appropriate iteration number using cross-validation.

We mention that our cross-validated ADMM implementation follows a "naive" approach, so its training time scales linearly in u_*. Although it seems unlikely that ADMM training could be turned entirely incremental in terms of reducing training time by combining models trained on separate subsets of the data, it might happen that a divide and conquer technique could significantly speed up this process, similar to the procedure of Joulani et al. (2015).

Finally, we point out that after dropping the "unused" hyperplanes (which do not influence the empirical risk), the size of the max-affine models on Fig-

ure 6.5 are still very close to n, far exceeding the optimal size $\lceil n^{d/(d+4)} \rceil$ suggested by Theorem 5.6. Hence, in the rest of the chapter, we discuss max-affine estimators that aim to improve the excess risk by reducing the number of hyperplanes.

6.1.6 Partitioning by the L_1–L_2 penalty term

One idea for reducing the number of hyperplanes in max-affine LSE models and keeping the theoretical guarantee is to intoduce a penalty term on the scaling of the error term $\alpha = O(n^{-4/(d+4)})$ of Theorem 5.6. The point is to choose the penalty term so to encourage sparsification of the hyperplanes. Here, we shortly discuss the L_1–L_2 penalty, which modifies the max-affine LSE training (6.2) as

$$\min_{\substack{\boldsymbol{a}_1,\dots,\boldsymbol{a}_n,\\ b_1,\dots,b_n}} \sum_{i=1}^{n} \Big(\max_{j=1,\dots,n} \boldsymbol{a}_j^\top \boldsymbol{\mathcal{X}}_i + b_j - \mathcal{Y}_i \Big)^2 + \beta \sum_{j=1}^{n} \|\boldsymbol{a}_j\| \ , \tag{6.13}$$

using some regularization parameter $\beta > 0$. This scheme applies an L_1-penalty over the hyperplanes and an L_2-penalty to their slopes individually, so it encourages sparsification and maintains rotational invariance.

Unfortunately, (6.13) cannot be rewritten to a convex form as (6.3) because the penalty term influences the locations where the hyperplanes are active (active means providing no lower value than any other hyperplane). To see why this is a problem, notice that the convex form (6.3) enforces the i-th hyperplane to be active at $\boldsymbol{\mathcal{X}}_i$, but usually this cannot be maintained when $\boldsymbol{a}_i = \boldsymbol{0}$ (only around the minimum). Hence, such a convex form cannot sparsify and so it is not equivalent to (6.13) which would surely start dropping hyperplanes as the regularization parameter β is set large enough.

One might still wonder whether (6.13) could be used for a postprocessing step. The idea is to train a max-affine LSE (6.3) in the first step, then filter the hyperplanes by locally solving (6.13), starting from the LSE solution. But as we observed, the local training in the second step does not maintain monotonicity (using a larger β should eliminate more hyperplanes), so it is unclear how one could select the appropriate level of regularization (learning β) in a reliable way to control the number of hyperplanes.

At this point, we reach the boundary of when a theoretical guarantee can be shown to hold on the risk of the studied estimators. From now, we turn to heuristic approaches that train max-affine models with significantly less than n hyperplanes. Although theory will be still used as a guide (for example to set some parameter values), the emphasis shifts to computational efficiency and adaptation for problems where max-affine estimators can exceed beyond worst-case excess risk convergence rates, at least empirically on problems needed for our applications.

6.2 Heuristic max-affine estimators

In this section we consider max-affine estimators using fewer hyperplanes than samples. By using smaller models, we hope for a computationally cheaper training process and for an improved excess risk on problems where "small" max-affine representations provide "sufficient" accuracy (for example f_*^{lfq} and f_*^{lhq}). Formally, we consider the following optimization problem,

$$\min_{M} \sum_{i=1}^{n} \Big(f_n(M; \boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i \Big)^2 \tag{6.14}$$

where $M \doteq \{(\boldsymbol{a}_k, b_k) : k = 1, \ldots, K\}$, $f_n(M, \boldsymbol{x}) \doteq \max_{k=1,\ldots,K} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k$, and the model size $K = |M|$ is less than the sample size n.

Let us first point out the difference between the partitioned LSE problem (6.4) and (6.14): the former searches among max-affine estimators that induce a fixed partition, while the latter considers any max-affine estimators inducing arbitrary partitions with size no larger than some fixed limit. Unfortunately, this generality does not come for free. In particular, we are not aware of any convex reformulation of (6.14). Hence, we only consider models that induce a partition in a finite partition space that is constructed in some heuristic way during the training process.

We also point out that the partition size K can also be learned from the data as well. Recall that max-affine LSEs having a near-optimal worst-case convergence rate exist with $K = \lceil n^{d/(d+4)} \rceil$ hyperplanes (Theorem 5.6). Furthermore, when the regression function admits a "good" max-affine approxi-

mation with "not too many" K_* hyperplanes, the estimators using about the same $O(K_*)$ number of hyperplanes can significantly improve the worst-case rate of Theorem 5.1 to $O(K_* \ln(n)/n)$. Hence, we limit the number of hyperplanes to $1 \leq K \leq \lceil n^{d/(d+4)} \rceil$.

Finally, we mention that one can solve the partitioned LSE problem (6.4) using the induced partition of a max-affine estimator to refine the estimate produced by a heuristic training method. We consider this approach for each training algorithm in the following sections.

6.2.1 Least squares partition algorithm

The *Least Squares Partition Algorithm* (LSPA, Magnani and Boyd, 2009) is a variation of the K-means clustering method that uses a greedy alternating optimization technique.

Initialized by some partition P_0 of $\{1, \ldots, n\}$, LSPA alternates between two steps. In the "update step" LSPA fits a max-affine model $M_t \doteq \{(\boldsymbol{a}_k^t, b_k^t) : k = 1, \ldots, K_t\}$ given a fixed partition $P_t = \{\mathcal{C}_1^t, \ldots, \mathcal{C}_{K_t}^t\}$ by fitting each cell individually using ridge regression (with some small regularizer β for stability):

$$\begin{aligned} \boldsymbol{a}_k^t &\doteq \Big(\sum_{i \in \mathcal{C}_k^t} \Delta_{ik} \Delta_{ik}^\top + \beta I_d \Big)^{-1} \sum_{i \in \mathcal{C}_k^t} \Delta_{ik} \mathcal{Y}_i \,, \quad \Delta_{ik} \doteq \boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mathcal{X}}}_k^t \,, \\ b_k^t &\doteq \frac{1}{|\mathcal{C}_k^t|} \sum_{i \in \mathcal{C}_k^t} \mathcal{Y}_i - (\bar{\boldsymbol{\mathcal{X}}}_k^t)^\top \boldsymbol{a}_k^t \,, \quad \bar{\boldsymbol{\mathcal{X}}}_k^t \doteq \frac{1}{|\mathcal{C}_k^t|} \sum_{i \in \mathcal{C}_k^t} \boldsymbol{\mathcal{X}}_i \,, \quad k = 1, \ldots, K_t \,. \end{aligned} \tag{6.15}$$

In the "assignment step" LSPA regroups the data according to the induced partition of the max-affine model M_t:

$$\begin{aligned} \hat{\mathcal{C}}_k^{t+1} &\doteq \big\{ i = 1, \ldots, n \,\big|\, \boldsymbol{\mathcal{X}}_i^\top \boldsymbol{a}_k^t + b_k = f_n(M_t; \boldsymbol{\mathcal{X}}_i) \big\} \,, \quad k = 1, \ldots, K_t \,, \\ P_{t+1} &= \big\{ \mathcal{C}_1^{t+1}, \ldots, \mathcal{C}_{K_{t+1}}^{t+1} \big\} \doteq \big\{ \hat{\mathcal{C}}_1^{t+1}, \ldots, \hat{\mathcal{C}}_{K_t}^{t+1} \big\} \setminus \{\emptyset\} \,, \end{aligned} \tag{6.16}$$

where ties are broken arbitrarily. The pseudocode of LSPA is given as Algorithm 6.3.

As the alternating optimization of LSPA is not guaranteed to converge, the algorithm is terminated after t_{wait} iterations counted from the last improvement of the emirical risk. An undesirable property of LSPA is that it is very sensitive to the initial partition P_0, because fitting a "bad cell" can produce

1. **input:** training set $\mathcal{D}_n$, partition P_0, regularizer β, iterations t_{wait}
2. $t \leftarrow 1$, $t_{\max} \leftarrow t_{\text{wait}}$, $M_* \leftarrow \emptyset$, $E_* \leftarrow \infty$
3. **while** $t \leq t_{\max}$ **do**
4. $\quad M_t \leftarrow$ fitting of partition P_t by (6.15) with β
5. $\quad E_t \leftarrow R_n\big(f_n(M_t; \cdot)\big)$ using $\ell = \ell_{\text{sq}}$
6. $\quad$ **if** $E_t < E_*$ **then**
7. $\quad\quad M_* \leftarrow M_t$, $E_* \leftarrow E_t$, $t_{\max} \leftarrow t + t_{\text{wait}}$
8. $\quad$ **end if**
9. $\quad P_{t+1} \leftarrow$ induced partition of M_t by (6.16)
10. $\quad t \leftarrow t + 1$
11. **end while**
12. **output:** model $M_* = \{(\boldsymbol{a}_k^*, b_k^*) : k = 1, \ldots, K_*\}$

Algorithm 6.3: Least Squares Partition Algorithm (LSPA) training a max-affine estimator by alternating optimization.

a hyperplane which becomes the only active one, thereby inducing a partition having only one cell and eliminating all other hyperplanes in a single step. For this reason, Magnani and Boyd (2009) propose restarting Algorithm 6.3 multiple times from random Voronoi partitions and tracking the best max-affine model having the smallest empirical risk over the whole training process. They did not provide any guideline for setting the size of the initial Voronoi partitions K_0.

Here, we propose initializing LSPA by random Voronoi partitions with $K_0 \doteq \lceil n^{d/(d+4)} \rceil$ cells which is the largest number needed for worst-case performance (Theorem 5.6). Furthermore, we draw the centers of the cells uniformly from $\{\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n\}$ without repetition as mentioned by Balázs et al. (2015, Section 5). Formally, we set

$$\begin{aligned} P_0 &\doteq \{\mathcal{C}_1^0, \ldots, \mathcal{C}_{K_0}^0\}\,, \text{ and for all } k = 1, \ldots, K_0, \\ \mathcal{C}_k^0 &\doteq \big\{ j = 1, \ldots, n \,\big|\, \|\boldsymbol{\mathcal{X}}_{i_k} - \boldsymbol{\mathcal{X}}_j\| = \min_{l=1,\ldots,K_0} \|\boldsymbol{\mathcal{X}}_l - \boldsymbol{\mathcal{X}}_j\| \big\}\,, \end{aligned} \tag{6.17}$$

where $i_1, \ldots, i_{K_0} \in \{1, \ldots, n\}$ are drawn uniformly without repetition (that is $i_k = i_l$ if and only if $k = l$, for all $k, l = 1, \ldots, K_0$).

The following experiment compares LSPA (with $t_{\text{wait}} = 10$, $\beta = 10^{-6}$, and various restart numbers $R = 10, 30, 50$) to cross-validated ADMM (cvADMM, Section 6.1.5) on problems (6.1). The results presented on Figure 6.6, show that LSPA improves the excess risk compared to cvADMM for problems f_*^{lfq}

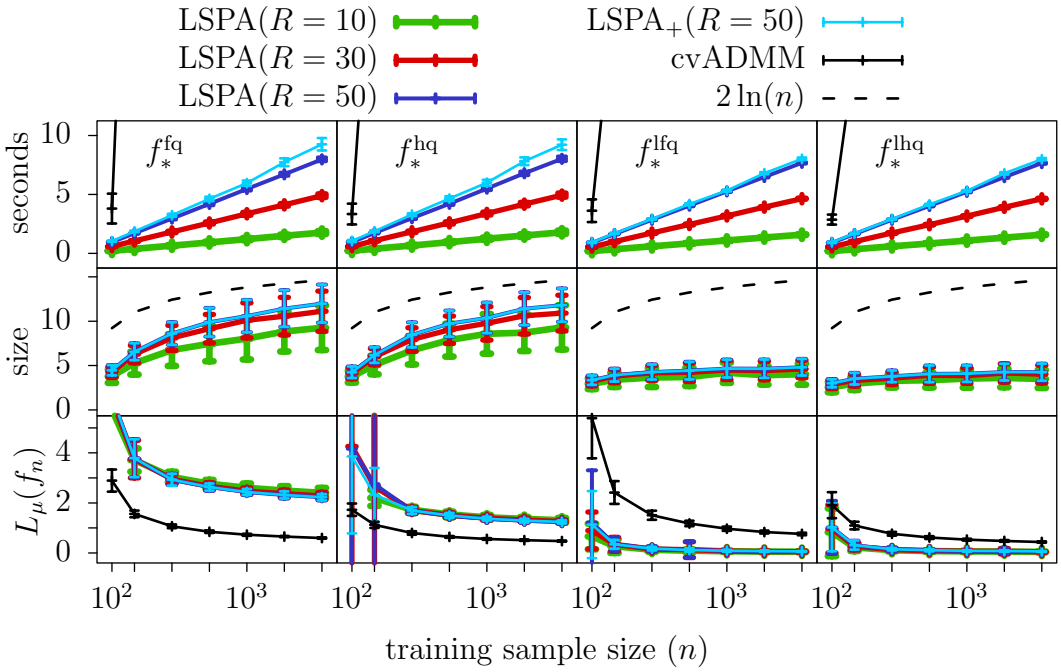

Figure 6.6: Comparison of LSPA and cross-validating ADMM (cvADMM) on discussion problems f_*^{fq}, f_*^{hq}, f_*^{lfq}, f_*^{lhq} for dimension $d = 8$ and sample sizes $n = 100, 250, 500, 750, 1000, 1250, 1500$. Averages of 100 experiments with standard deviation error bars are shown for training times (in seconds), model size, and excess risk $L_\mu(f_n)$ with $\ell = \ell_{\mathrm{sq}}$ measured on 10^6 new samples for each experiment.

and f_*^{lhq} where the max-affine representation is exact, otherwise it is worse for problems f_*^{fq} and f_*^{hq}. While the LSPA model size is significantly larger for f_*^{fq}, f_*^{hq} than for f_*^{lfq}, f_*^{lhq} as it is needed for a more accurate representation, it is still far below the bound $\lceil n^{d/(d+4)} \rceil$ suggested by Theorem 5.6 which explains the underfitting effect on these "difficult" problems. However, working with such reduced model sizes LSPA trains much faster than the algorithms we considered in Section 6.1, even for $R = 50$ repetitions which looked quite stable for these examples.

Figure 6.6 also presents an LSPA variant (LSPA$_+$) that performs a post-processing step by solving the partitioned LSE problem (Section 6.1.1) over the partition induced by the max-affine estimator obtained by LSPA. For this, we use ADMM with a slightly different stopping rule, that terminates the algorithm when the change in the empirical risk and the amount of constraint

violations drop under some threshold (we use 10^{-6}, dictated by that for 64 bit floats $10^{-8} \approx \sqrt{\epsilon}$ where ϵ is the machine precision). Surprisingly, LSPA$_+$ barely changes the performance of LSPA regarding empirical and excess risks, indicating that the least squares fit of the LSPA partition is "nearly" a local optimum.

To summarize, LSPA works well for the examples (6.1), but it is unclear how many restarts are necessary in general and the algorithm might overfit by producing cells with too few data points. The following methods try to fix these issues by constructing the partition "more carefully".

6.2.2 Convex adaptive partitioning algorithm

The *Convex Adaptive Partitioning* (CAP) algorithm (Hannah and Dunson, 2013) proposes an incremental cell splitting partitioning technique and combines it with an LSPA step. Over the iterations, CAP also maintains a minimum cell size which makes hyperplane fitting more "reliable" and improves robustness against overfitting.

CAP builds the partition incrementally in each iteration by splitting one cell using a linear cut. Because there are too many such linear cuts, CAP considers only a subset along specific directions and a few cut points. Formally, CAP splits the k-th cell along the j-th coordinate and v-th "knot" as

$$\mathcal{C}'_k \doteq \left\{i \in \mathcal{C}_k : \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k \geq c_{kjv}\right\}, \quad \mathcal{C}'_{K+1} \doteq \left\{i \in \mathcal{C}_k : \boldsymbol{a}_k^\top \boldsymbol{\mathcal{X}}_i + b_k > c_{kjv}\right\},$$
$$c_{kjv} \doteq \frac{v}{v_* + 1} \min\{\mathcal{X}_{ij} : i \in \mathcal{C}_k\} + \left(1 - \frac{v}{v_* + 1}\right) \max\{\mathcal{X}_{ij} : i \in \mathcal{C}_k\}, \tag{6.18}$$

where $\boldsymbol{\mathcal{X}}_i = [\mathcal{X}_{i1} \dots \mathcal{X}_{id}]^\top$, $j = 1, \dots, d$, and $v = 1, \dots, v_*$. Among the Kdv_* possible cuts, only those partitions $P_{K+1} \doteq (P_K \setminus \{\mathcal{C}_k\}) \cup \{\mathcal{C}'_k, \mathcal{C}'_{K+1}\}$ are considered for which the new cells $\mathcal{C}'_k$, $\mathcal{C}'_{K+1}$ also maintain the minimum size s_*. Then the new cells are fitted by ridge regression and the model with the lowest empirical risk is selected. Finally, at the end of the CAP iteration, an LSPA step is performed for the chosen model, when the induced partition of the hyperplanes maintains the minimum cell size requirement.

The CAP max-affine training method is shown as Algorithm 6.4. In each iteration, CAP adds one new cell to the current partition and also adds one

new hyperplane. The new partition is obtained by splitting a cell (step 5),

1. **input:** training set $\mathcal{D}_n$, minimum cell size s_*, knot number v_*
2. $P_1 \leftarrow$ single cell partition $\{\{1, \ldots, n\}\}$
3. $M_1 = \{(\boldsymbol{a}_1, b_1)\} \leftarrow$ hyperplane fitting P_1
4. **for** $K = 2, 3, \ldots$ **do**
5. $\quad \{\mathcal{P}_K\} \leftarrow$ propose new partitions by splitting the cells of P_{K-1} as described by (6.18) using v_* and s_*
6. $\quad$ stop if proposal set $\{\mathcal{P}_K\}$ is empty
7. $\quad \{M_K\} \leftarrow$ compute new models using M_{K-1} by fitting the new cells of the partitions in $\{\mathcal{P}_K\}$
8. $\quad (M_K, P_K) \leftarrow$ model and its partition in $\{M_K, \mathcal{P}_K\}$ with the smallest empirical risk
9. $\quad \hat{P}_K \leftarrow$ induced partition of M_K by (6.16)
10. $\quad$ **if** every cell of $\hat{P}_K$ has size at least s_* **then**
11. $\quad\quad P_K \leftarrow \hat{P}_K$, $M_K \leftarrow$ fitting of partition $\hat{P}_K$ by (6.15) with $\beta = 0$
12. $\quad$ **end if**
13. **end for**
14. $M_* \leftarrow$ select the best model from $M_1, \ldots, M_K$ which minimizes the approximate GCV error
15. **output:** model $M_* = \{(\boldsymbol{a}_k^*, b_k^*) : k = 1, \ldots, K_*\}$

Algorithm 6.4: Convex Adaptive Partitioning (CAP) algorithm training a max-affine estimator by incremental cell splitting.

and the new model is created by taking the old one and fitting the two cells of the split (step 7). Then the best model is selected among all proposed ones with the lowest empirical risk (step 8). At the end, an LSPA iteration is performed, but accepted only if the induced partition satisfies the minimum cell size requirements (steps 9 to 12). The algorithm terminates when there are no more possible splits due to the minimum cell size requirement (step 6) and finally returns the best model minimizing an approximate version of the Generalized Cross-Validation (GCV) error (step 14) as described by Hannah and Dunson (2013, Section 5).

We tested CAP on our discussion problems (6.1) and summarize the results on Figure 6.7. We ran CAP with knot number $v_* = 10$, and set the minimum cell size $s_* \doteq \max\{2(d+1), n/(D_* \ln(n))\}$ with $D_* = 3$ or $D_* = 10$, as suggested by Hannah and Dunson (2013, Section 3.1).[7] Here we point out that the choice of s_* is crucial for the behavior of CAP. First, the definition of s_* upper bounds

[7] Their paper has a typographical error having min instead of max in the definition of s_*.

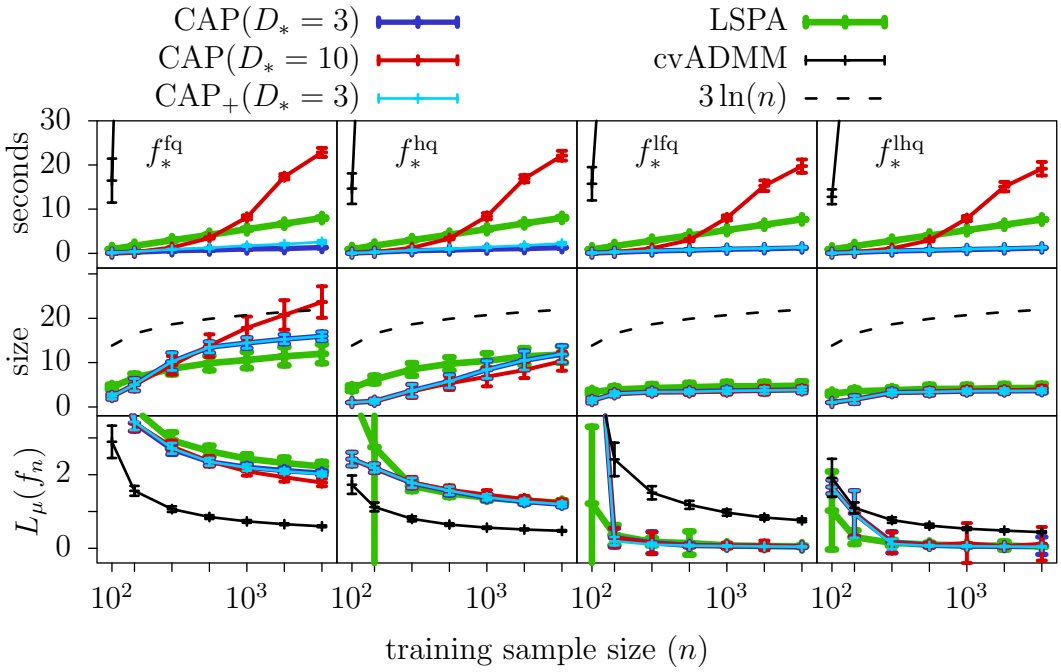

Figure 6.7: Comparison of CAP, LSPA ($R = 50$), and cvADMM ($u_* = 10$) on discussion problems f_*^{fq}, f_*^{hq}, f_*^{lfq}, f_*^{lhq}. The dimension is $d = 8$ and the sample sizes are $n = 100, 250, 500, 750, 1000, 1250, 1500$. Averages of 100 experiments with standard deviation error bars are shown for training times (in seconds), model size, and excess risk $L_\mu(f_n)$ measured on 10^6 new samples for each experiment.

the number of hyperplanes by $D_* \ln(n)$. Then, because CAP terminates only when there are no more splits available, this logartihmic bound on the model size is also the cause of its attractive speed. This can be observed on Figure 6.7 showing that on all problems regardless of the size of the final max-affine model using $D_* = 10$ is significantly slower than $D_* = 3$. Furthermore, the results also show that using larger models increase the variance of CAP for problems with small max-affine regression functions.

Figure 6.7 also shows a CAP variant (CAP$_+$) which performs a post–processing step by ADMM Algorithm 6.2 in the same way as it has been performed for LSPA in Section 6.2.1. Similar to LSPA$_+$, CAP$_+$ provides only a small variance improvement for the excess risk.

To summarize, CAP can provide a training speedup compared to LSPA by restricting its search process for models with smaller sizes. However, even

a slight change of the model size bound can increase the training time by much, even if the proposed bound is logarithmic in the sample size n, which is significantly less than $\lceil n^{d/(d+4)} \rceil$, the optimal value suggested by Theorem 5.6. To address this issue, we describe another algorithm in the next section which drops the logarithmic model size restriction.

6.2.3 Adaptive max-affine partitioning algorithm

In this section we present a new algorithm, called *Adaptive Max-Affine Partitioning* (AMAP), which combines the partitioning technique of CAP, LSPA, and the cross-validation scheme of cvADMM. Our goal is to adapt the model size and the training speed to the regression problem.

Similar to CAP, AMAP builds the model incrementally by splitting a cell and improving the partition using LSPA. The AMAP model improvement step is given by Algorithm 6.5. AMAP performs coordinatewise cell splitting (steps 5 to 15), just as CAP, however, AMAP makes the split always at the median (steps 6 and 7) instead of checking multiple cut points. This saves computation time, but can also create worse splits. To compensate for this loss in quality, AMAP runs a restricted version of LSPA (steps 18 to 23) not just for a single step as CAP, but until the candidate model improves the empirical risk and its induced partition satisfies the minimum cell requirement (step 23). We also mention that indices $\{i \in \mathcal{C}_k : \mathcal{X}_{ij} = m_j\}$ are assigned to $\mathcal{C}_{\text{le}}$ and $\mathcal{C}_{\text{gt}}$ (step 7) in order to preserve the minimum cell requirement.

Notice that the difference between M' and M is only two hyperplanes (step 10), so the number of arithmetic operations for computing E' (step 11) can be improved from $O(nKd)$ to $O(nd)$. Further, the cost of ridge regressions (steps 8 and 9) is $O(|\mathcal{C}_k|d^2)$. Hence, the computational cost of the entire cell splitting process (steps 2 to 17) is bounded by $O(\max\{K, d\}d^2n)$. For the LSPA part, the partition fitting (step 21) is $O(nd^2)$ and the error calculation (step 22) is $O(nKd)$. So, the cost of a single LSPA iteration (steps 19 to 22) is bounded by $O(\max\{K, d\}dn)$, implying that the cost of Algorithm 6.5 is bounded by $O\big(\max\{t_{\text{LSPA}}, d\}\max\{K, d\}dn\big)$, where t_{LSPA} denotes the number of LSPA iterations.

input: training set $\mathcal{D}_n$, model $M = \{(\boldsymbol{a}_k, b_k) : k = 1, \dots, K\}$,
partition $P = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$, empirical risk E,
minimum cell size s_*, regularizer β
{cell splitting}
$M'_* \leftarrow M, E'_* \leftarrow E, P'_* \leftarrow P$
for all $k = 1, \dots, K$ **do**
  if $|\mathcal{C}_k| \geq 2s_*$ **then**
    for all $j = 1, \dots, d$ **do**
      $m_j \leftarrow \text{median}(\{\mathcal{X}_{ij} : i \in \mathcal{C}_k\})$
      $\mathcal{C}_{\text{le}} \leftarrow \{i \in \mathcal{C}_k : \mathcal{X}_{ij} \leq m_j\}, \mathcal{C}_{\text{gt}} \leftarrow \{i \in \mathcal{C}_k : \mathcal{X}_{ij} \geq m_j\}$
      $(\boldsymbol{a}_{\text{le}}, b_{\text{le}}) \leftarrow$ ridge regression on $\{(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) : i \in \mathcal{C}_{\text{le}}\}$ with β
      $(\boldsymbol{a}_{\text{gt}}, b_{\text{gt}}) \leftarrow$ ridge regression on $\{(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) : i \in \mathcal{C}_{\text{gt}}\}$ with β
      $M' \leftarrow \big(M \setminus \{(\boldsymbol{a}_k, b_k)\}\big) \cup \{(\boldsymbol{a}_{\text{le}}, b_{\text{le}}), (\boldsymbol{a}_{\text{gt}}, b_{\text{gt}})\}$
      $E' \leftarrow R_n\big(f_n(M'; \cdot)\big)$
      if $E' < E'_*$ **then**
        $M'_* \leftarrow M', E'_* \leftarrow E', P'_* \leftarrow (P \setminus \{\mathcal{C}_k\}) \cup \{\mathcal{C}_{\text{le}}, \mathcal{C}_{\text{gt}}\}$
      end if
    end for
  end if
end for
{running LSPA}
repeat
  $M_* \leftarrow M'_*, E_* \leftarrow E'_*, P_* \leftarrow P'_*$
  $P'_* \leftarrow$ induced partition of M_* by (6.16)
  $M'_* \leftarrow$ fitting of partition P'_* by (6.15) with β
  $E'_* \leftarrow R_n\big(f_n(M'_*; \cdot)\big)$
until $\min_{\mathcal{C} \in P'_*} |\mathcal{C}| \geq s_*$ and $E'_* < E_*$
output: model M_*, partition P_*, empirical risk E_*

Algorithm 6.5: Adaptive max-affine partitioning (AMAP) model improvement step using incremental cell splitting and LSPA.

One problem with this algorithm is that coordinatewise cell splitting is not rotation invariant. To fix this, we run AMAP after a pre-processing step, which uses thin singular value decomposition (thin-SVD) to drop redundant coordinates and align the data along a rotation invariant basis. Formally, let the raw (but already centered) data be organized into $X \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$. Then, we scale the values $[\mathcal{Y}_1 \dots \mathcal{Y}_n]^\top \doteq \boldsymbol{y} / \max\{1, \|\boldsymbol{y}\|_\infty\}$, and decompose X by thin-SVD as $X = USV^\top$, where $U \in \mathbb{R}^{n \times d}$ is semi-orthogonal, $S \in \mathbb{R}^{d \times d}$ is diagonal with singular values in decreasing order, and $V \in \mathbb{R}^{d \times d}$

is orthogonal. Coordinates related to zero singular values are dropped[8] and the points are scaled by S as $[\boldsymbol{\mathcal{X}}_1 \dots \boldsymbol{\mathcal{X}}_n]^\top \doteq US/\max\{1, S_{11}\}$, where S_{11} is the largest singular value. Now observe that rotating the raw points X as XQ with some orthogonal $Q \in \mathbb{R}^{d\times d}$ only transforms V to $Q^\top V$ and does not affect the pre-processed points $\boldsymbol{\mathcal{X}}_1, \dots, \boldsymbol{\mathcal{X}}_n$. Finally, we note that thin-SVD can be computed using $O(nd^2)$ arithmetic operations (with $n \geq d$),[9] which is even less than the asymptotic cost of Algorithm 6.5.

AMAP is presented as Algorithm 6.6, and run using uniformly shuffled (and pre-processed) data $\mathcal{D}_n$ and a partition $\{F_1, \dots, F_{u_*}\}$ of $\{1, \dots, n\}$ having about equally sized cells as described in Section 6.1.5. As before, the regularizer β (we use 10^{-6}) is set to a small value only to ensure numerical stability.

For model selection, AMAP replaces the approximate GCV scheme of CAP by u_*-fold cross-validation (steps 9 to 20) and terminates when the cross-validation error (step 15) of the best model set $\mathcal{M}_*$ (steps 8 and 17) cannot be further improved for t_{wait} iterations, similarly as done by cvADMM. At the end, the final model is chosen from the model set $\mathcal{M}_*$ with the best cross-validation error, and minimizes the empirical risk on the entire data (step 21).

AMAP starts with models having a single hyperplane (steps 2 to 7) and increments each model by at most one hyperplane in every iteration (step 12). Notice that if AMAP cannot find a split for a model M_u to improve the empirical risk $\overline{E}_u$, the update for model M_u (steps 12 and 13) can be skipped in the subsequent iterations as Algorithm 6.5 is deterministic. We also mention that for the minimum cell size, we use $s_* \doteq \max\{2(d+1), \lceil \log_2(n) \rceil\}$ allowing model sizes up to $O\big(n^{d/(d+4)}/\ln(n)\big)$, which is enough for near-minimax performance (Theorem 5.6).

An empirical comparison of AMAP, CAP, LSPA, and cvADMM is provided by Figure 6.8 on the four discussion problems f_*^{fq}, f_*^{hq}, f_*^{lfq}, and f_*^{lhq}. Observe that the computational cost of AMAP grows with the model size, preserving

[8] By removing columns of U and V, and columns and rows of S.

[9] First decompose X by a thin-QR algorithm in $O(nd^2)$ time (Golub and Loan, 1996, Section 5.2.8) as $X = QR$, where $Q \in \mathbb{R}^{n\times d}$ has orthogonal columns and $R \in \mathbb{R}^{d\times d}$ is upper triangular. Then apply SVD for R in $O(d^3)$ time (Golub and Loan, 1996, Section 5.4.5).

input: training set $\mathcal{D}_n$, minimum cell size s_*,
folds $F_1, \dots, F_{u_*}$, iterations t_{wait}, regularizer β
{initialization}
for $u = 1, \dots, u_*$ **do**
  $P_u \leftarrow \{\bar{F}_u\}$ with $\bar{F}_u \doteq \{1, \dots, n\} \setminus F_u$
  $M_u = \{(\boldsymbol{a}_1^u, b_1^u)\} \leftarrow$ ridge regression on $\{(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) : i \in \bar{F}_u\}$ with β
  $\bar{E}_u \leftarrow |\bar{F}_u|^{-1} \sum_{i \in \bar{F}_u} |f_n(M_u; \boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i|^2$
  $E_u \leftarrow |F_u|^{-1} \sum_{i \in F_u} |f_n(M_u; \boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i|^2$
end for
$\mathcal{M}_* \leftarrow \{M_1, \dots, M_{u_*}\}$, $E_* \leftarrow \frac{1}{u_*} \sum_{u=1}^{u_*} E_u$
{cross-validation training}
$t \leftarrow 1$, $t_{\max} \leftarrow t_{\text{wait}}$
while $t \leq \min\{t_{\max}, \lceil n^{d/(d+4)} \rceil\}$ **do**
  for $u = 1, \dots, u_*$ **do**
    $(M_u, P_u, \bar{E}_u) \leftarrow$ update by Algorithm 6.5 using
    $\{(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) : i \in \bar{F}_u\}$, M_u, P_u, $\bar{E}_u$, s_*, β
    $E_u \leftarrow |F_u|^{-1} \sum_{i \in F_u} |f_n(M_u; \boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i|^2$
  end for
  $E \leftarrow \frac{1}{u_*} \sum_{u=1}^{u_*} E_u$
  if $E < E_*$ **then**
    $\mathcal{M}_* \leftarrow \{M_1, \dots, M_{u_*}\}$, $E_* \leftarrow E$, $t_{\max} \leftarrow t + t_{\text{wait}}$
  end if
  $t \leftarrow t + 1$
end while
{choosing the final model}
$M_* \leftarrow \operatorname{argmin}_{M \in \mathcal{M}_*} R_n\big(f_n(M; \cdot)\big)$
output: model M_*

Algorithm 6.6: Cross-validated adaptive max-affine partitioning (AMAP).

a fast training for the linearized problems f_*^{lfq} and f_*^{lhq}. Furthermore, the freedom of choosing larger models did not increase the variance of AMAP as it did for CAP on the linearized problems, while the cross-validation scheme significantly improved the excess risk for small sample sizes. On the "difficult" problems, AMAP also provides the best excess risk, although both CAP and LSPA are close.

Finally, notice that the post-trained version AMAP_+ provides almost identical results to AMAP, only a slight improvement can be observed on Figure 6.8 for problems f_*^{fq} and f_*^{hq}. Based on the empirical results of LSPA_+, CAP_+, and AMAP_+, it seems that there are many local optima of (6.14) which are close to an "LSPA equilibrium". Clearly, this raises the question whether "good

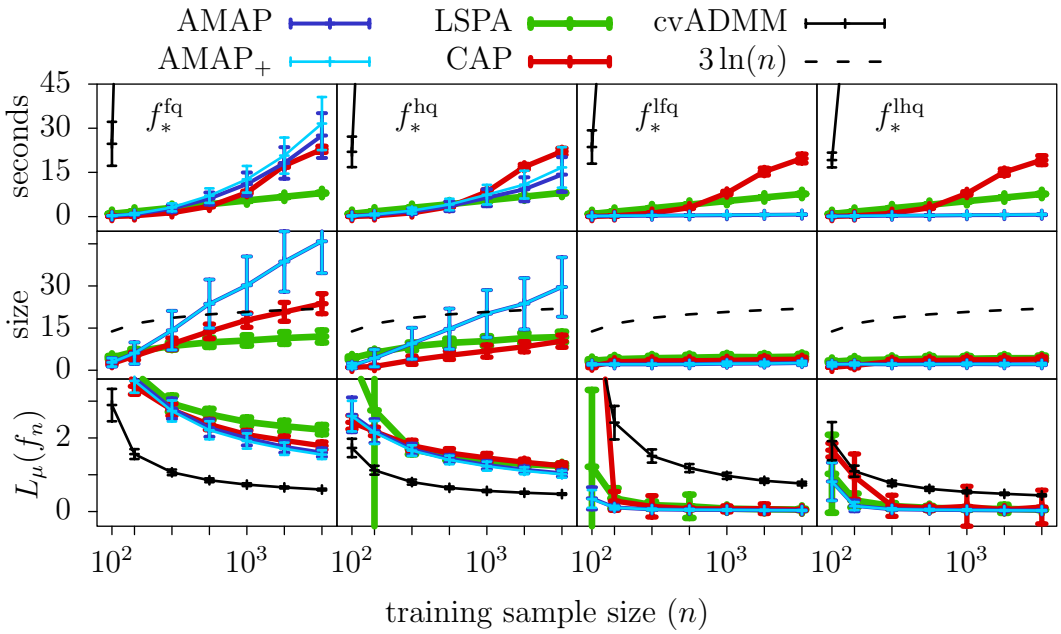

Figure 6.8: Comparison of AMAP, CAP ($D_* = 10$), LSPA ($R = 50$), and cvADMM ($u_* = 10$) on discussion problems f_*^{fq}, f_*^{hq}, f_*^{lfq}, f_*^{lhq}. The dimension is $d = 8$ and the sample sizes are $n = 100, 250, 500, 750, 1000, 1250, 1500$. Averages of 100 experiments with standard deviation error bars are shown for training times (in seconds), model size, and excess risk $L_\mu(f_n)$ measured on 10^6 new samples for each experiment.

quality" solutions are of this type or using LSPA restricts the partition search too much. We do not know the answer for sure, but some insight suggesting the former is provided later in Section 8.3.

Chapter 7

Evaluation of max-affine estimators

So far we studied max-affine training algorithms on the discussion problems f_*^{fq}, f_*^{hq}, f_*^{lfq}, f_*^{lhq} (6.1) for which we kept the dimension ($d = 8$) and the sample sizes ($100 \leq n \leq 1500$) small enough allowing us to compute max-affine LSEs (Section 6.1). In this chapter, we consider larger convex regression problems to evaluate heuristic max-affine estimators (Section 6.2) on randomized synthetic, real data sets, and stochastic programming problems.[1]

While the randomized synthetic problems (Section 7.1) aim to reveal the strengths and weaknesses of max-affine estimators, the real world data (Section 7.2) and the stochastic programming problems (Section 7.3) focus on the applications presented in Chapter 2.

To relate max-affine estimators to other regression techniques, we provide results for Multivariate Adaptive Regression Splines (MARS, Friedman, 1991) with a piecewise cubic model as implemented by Jekabsons (2016),[2] and Support Vector Regression (SVR, Vapnik, 1998, Chapter 11) with a radial basis function (RBF) kernel as implemented by Chang and Lin (2011).[3] For the training of the max-affine estimators, we consider AMAP, CAP ($D_* = 3$ or 5), and LSPA ($R = 50$).

[1] To run the experiments, we used the same hardware and software tools as mentioned at the beginning of Chapter 6.

[2] Default parameter values were used (ARESLab ver. 1.10.3).

[3] ϵ-SVR using RBF kernel were trained by 5-fold cross-validation using $C \in \{1, 5, 10\}$ and $\gamma \in \{1/(4d), 1/d, 4/d\}$. Defaults were used for the other parameters (LIBSVM ver. 3.21), and the data was centered and scaled (using $\boldsymbol{\mathcal{X}}_i / \max_{i,j} |\mathcal{X}_{ij}|$ and $\mathcal{Y}_i / \max_i |\mathcal{Y}_i|$).

7.1 Randomized synthetic problems

In this section we consider synthetic convex regression problems with randomly generated regression functions. The goal of these tests is to reveal the strengths and weaknesses of max-affine estimators providing a general convex regression benchmark which is (likely) harder to overfit.

7.1.1 Quadratic and max-affine targets

First, we consider randomized versions of the discussion problems (6.1) on a larger scale than in the previous sections by increasing the dimension d to 10 and 20, while increasing the sample size range to $10^3 \leq n \leq 10^4$. The parameters $H_*, \boldsymbol{f}_*, c_*$ of f_*^{fq} and f_*^{hq} and the discretization points $\{\boldsymbol{x}_k^* : k = 1, \ldots, L_*\}$ of f_*^{lfq} and f_*^{lhq} are generated randomly[4] for each problem instance (while they were kept fixed in the above sections). The number of discretization points is always set as $K_* \doteq 2d$.

The results for $d = 10$ are presented on Figure 7.1, showing that max-affine estimators perform worse than SVR on the quadratic problems f_*^{fq} and f_*^{hq}, and better on the max-affine ones f_*^{lfq} and f_*^{lhq}, as expected. Comparing Figure 7.1 to Figure 6.8 notice that AMAP performs much better than CAP on problems f_*^{fq} and f_*^{lfq} for the randomized setting than for the fixed configuration of (6.1) used in Chapter 6.

Surprisingly, MARS worked so poorly[5] that it appears only on the plot of f_*^{hq} with by far the worst result, which suggests that its default parameter setting is not universal and in particular does not fit to these problems well. As MARS was the slowest algorithm, we could not afford to use cross-validation to tune even one of its many parameters.

The differences among max-affine estimators (AMAP, LSPA, CAP) grow even further when the dimension is increased from $d = 10$ to $d = 20$ as presented on Figure 7.2, but the overall picture remains similar.

[4] The value of c_* is uniformly drawn from $[-1, 1]$, $H_* \doteq (1/d)\hat{H}_*^\top \hat{H}_*$, and the coordinates of $\hat{H}_*$ and $\boldsymbol{f}_*$ are drawn independently from the standard normal distribution.

[5] Although MARS was rarely comparable to the max-affine estimators and SVR, it was always significantly better than ridge regression as expected.

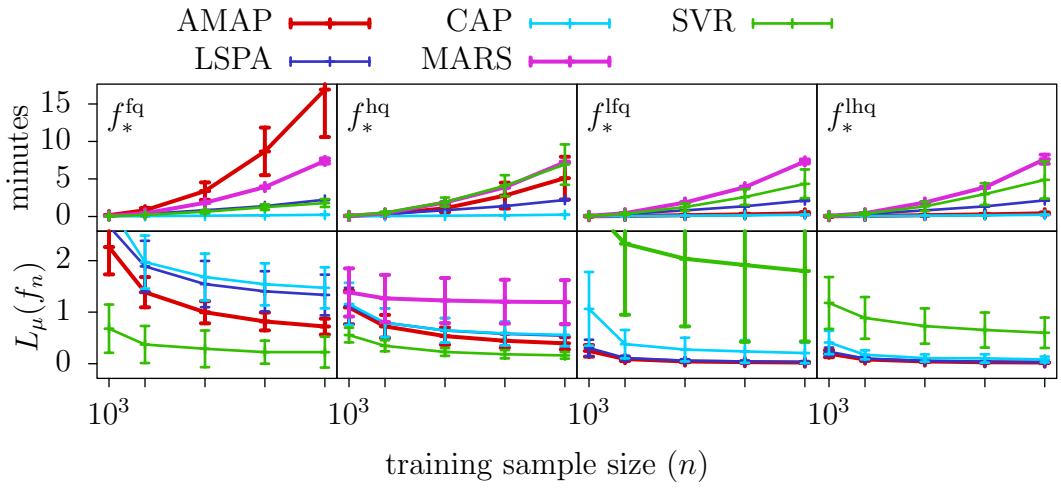

Figure 7.1: Performance of max-affine estimators (AMAP, LSPA, CAP), SVR and MARS on randomized quadratic (f_*^{fq}, f_*^{hq}) and max-affine (f_*^{lfq}, f_*^{lhq}) problems. The dimension is $d = 10$ and the sample sizes are $n = 10^3$, 2500, 5000, 7500, 10^4. Averages of 100 experiments with standard deviation error bars are shown for the training time (in minutes), and the excess risk $L_\mu(f_n)$ measured on 10^6 new samples for each experiment.

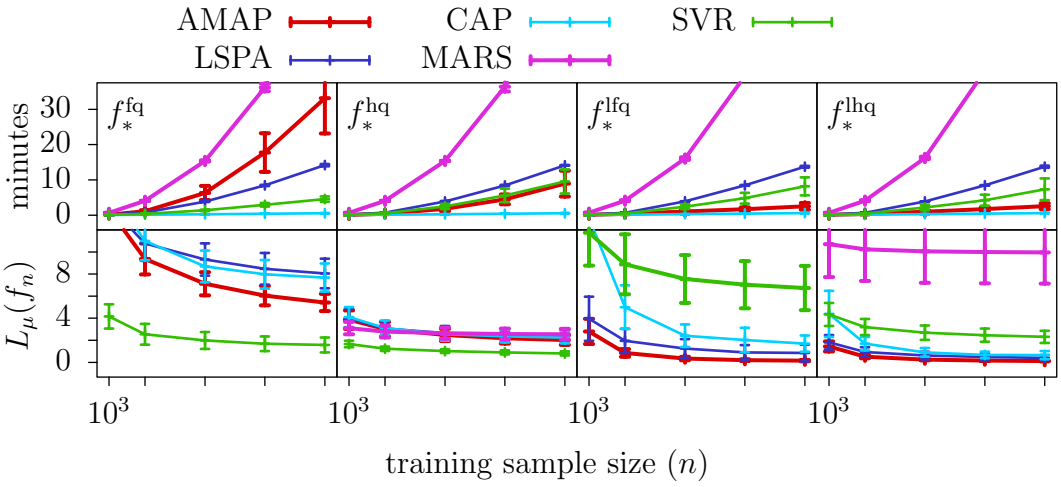

Figure 7.2: Performance of max-affine estimators (AMAP, LSPA, CAP), SVR and MARS on randomized quadratic (f_*^{fq}, f_*^{hq}) and max-affine (f_*^{lfq}, f_*^{lhq}) problems. The dimension is $d = 20$ and sample sizes are $n = 10^3$, 2500, 5000, 7500, 10^4. Averages of 100 experiments with standard deviation error bars are shown for the training time (in minutes), and the excess risk $L_\mu(f_n)$ measured on 10^6 new samples for each experiment.

7.1.2 Sum-max-affine and log-sum-exp targets

Based on the results of Section 7.1.1, one might think that max-affine estimators work better for piecewise linear targets than for smooth ones. In this section we show that it is not the case by testing the algorithms on sum-max-affine (f_*^{sma}, piecewise linear) and log-sum-exp (f_*^{lse}, smooth) convex functions, defined as

$$f_*^{\mathrm{sma}}(\boldsymbol{x}) \doteq \sum_{s=1}^{S_*} \max_{k=1,\dots,K_*} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k \,, \qquad f_*^{\mathrm{lse}}(\boldsymbol{x}) \doteq \ln \sum_{k=1}^{K_*} \exp(\boldsymbol{a}_k^\top \boldsymbol{x} + b_k) \,,$$

where the parameters $\boldsymbol{a}_k, b_k$ are generated randomly the same way as for the linearized quadratic function f_*^{lfq} in Section 7.1.1. Further, we set $S_* \doteq 2d$, $K_* \doteq d$, sample covariates uniformly $\boldsymbol{\mathcal{X}} \sim \mathcal{U}(\mathbb{X})$ over $\mathbb{X} \doteq [-2, 2]^d$, and use the same standard Gaussian noise model as for (6.1).

The results are presented by Figure 7.3 showing that max-affine estimators

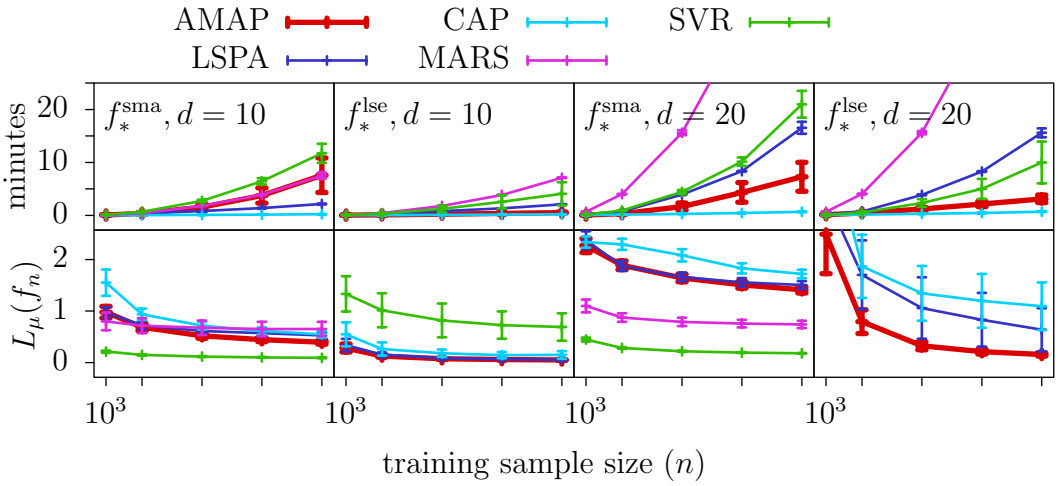

Figure 7.3: Performance of max-affine estimators (AMAP, LSPA, CAP), SVR and MARS on randomized sum-max-affine (f_*^{sma}) and log-sum-exp (f_*^{lse}) problems. The dimensions are $d = 10, 20$ and sample sizes are $n = 10^3$, 2500, 5000, 7500, 10^4. Averages of 100 experiments with standard deviation error bars are shown for the training time (in minutes), and the excess risk $L_\mu(f_n)$ measured on 10^6 new samples for each experiment.

are weaker for sum-max-affine functions and better for log-sum-exp targets compared to SVR and MARS, especially as the complexity of the regression

functions grow. In fact, noting that sum-max-affine representations can approximate quadratic functions well (see Section 8.2) and log-sum-exp functions are just the smoothened versions of max-affine ones, these results are not so surprising (except for MARS which works much better for f_*^{sma} than for f_*^{fq}), but rather match the conclusions of Section 7.1.1.

7.2 Real problems

Here we evaluate the algorithms using 100-fold cross-validation on a few real data sets coming from convex regression settings.

7.2.1 Convex estimation of average wages

We consider the estimation of wages ($\mathcal{Y}_i$) based on education and experience ($\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^2$) using two data sets. The first data set (BW, Verbeek, 2004, Section 3.5) contains 1471 entries (after removing one outlier) of Belgian hourly wages with education level and years of experience. The second data set (SL, Ramsey and Schafer, 2002, Chapter 10, Exercise 29) contains 25601 entries (after removing 31 outliers) of US weekly wages ($\mathcal{Y}_i$) with years of education and experience ($\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^2$). This was proposed as a convex regression benchmark by Hannah and Dunson (2013). The average output of this data is also shown by Figure 2.1 and explained in Section 2.1.

The results for the BW and SL data sets are presented by Figure 7.4, where the split plot on the left shows mean and standard deviation estimates at the bottom (SL values were scaled by 0.01) measured by 100-fold cross-validation, while the top part zooms ($\times 15$) the tip of the $R_\mu(f_n)$ mean estimates and provides them with standard errors (standard deviation divided by the root number of experiments, which is 10). The bottom part shows that both problems BW and SL have a significant amount of measurement noise ($\mathcal{Y}_i - f_*(\boldsymbol{\mathcal{X}}_i)$), and all algorithms are as good as linear LSE (ridge) for prediction tasks using the risk $R_\mu(f_n)$.

However, comparing the algorithms for estimation tasks requires the excess risk $L_\mu(f_n, f_*)$, which cannot be measured because f_* is unknown. Using the

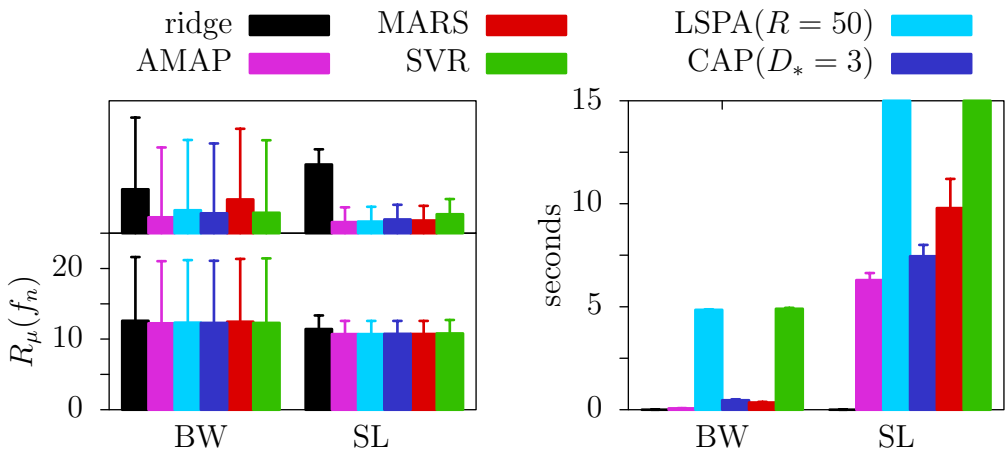

Figure 7.4: Comparison of max-affine estimators (AMAP, LSPA, and CAP), MARS, and SVR on the average wage estimation problems (BW and SL) using 100-fold cross-validation.

constant shift $L_\mu(f_n, f_*) = R_\mu(f_n) - R_\mu(f_*)$ for the squared loss $\ell = \ell_{\text{sq}}$, we can at least compare the algorithms by the mean estimates of $R_\mu(f_n)$, which is equivalent to ranking the algorithms by the (non-measurable) mean estimates of $L_\mu(f_n, f_*)$. Such ranking is provided by the left-top part of Figure 7.4 using standard error "confidence" bars, showing that linear estimates are indeed the best on the BW problem, but not on the SL one, where convex estimators compete with state-of-the-art non-convex regression techniques such as MARS and SVR. Of course, this picture is not quite complete without knowing the standard deviation of these algorithms regarding the excess risk $L_\mu(f_n, f_*)$, which unfortunately cannot be measured without evaluating the regression function f_* at the data points $\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n$.

Finally, notice on Figure 7.4 that AMAP is faster than CAP and much faster than LSPA (run for about 85 seconds) while its risk performance is about the same, which is due to its improved efficiency for problems fitted with small max-affine models (on the SL problem, the number of used hyperplanes is about 7 to 9 for AMAP, LSPA and CAP as well, while it is even less for the BW case).

7.2.2 Convex fitting of aircraft profile drag data

Now consider the XFOIL aircraft design problem which requires the max-affine approximation of a non-convex regression function (see Section 2.2 for more details). The regression function can be measured without noise by the XFOIL simulator and its shape is not far from being convex, hence using max-affine estimators is reasonable. We also point out that the absence of measurement noise ($\mathcal{Y}_i = f_*(\boldsymbol{\mathcal{X}}_i)$ a.s.) implies $R_\mu(f_n) = L_\mu(f_n, f_*)$, so we can observe the standard deviation of the estimates, not just their means as for the wage estimation problem above. The XFOIL data set (Hoburg and Abbeel, 2014) contains 3073 entries of profile drag coefficient data ($\mathcal{Y}_i$) with lift coefficient, Reynolds number and wing thickness ratio ($\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^3$).

The result is presented by Figure 7.5 showing that the problem is highly nonlinear ($L_\mu(f_n, f_*) \approx 0.06 \pm 0.015$ for ridge regression) and using max-affine

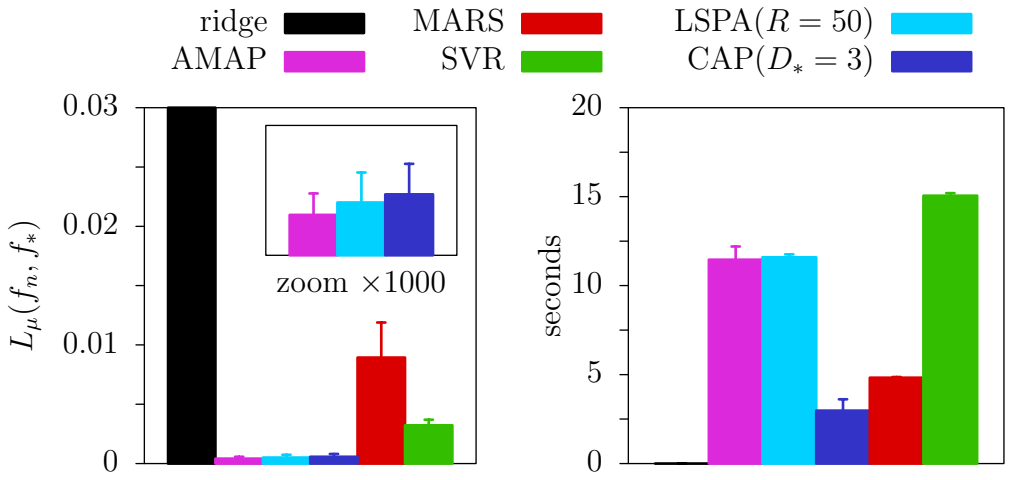

Figure 7.5: Comparison of max-affine estimators (AMAP, LSPA, CAP), MARS, and SVR on the XFOIL aircraft profile drag approximation problem using 100-fold cross-validation.

estimators is not just a necessary compromise for the application, but also very attractive as they significantly outperform the non-convex approaches (there is almost a factor of 8 between the excess risk of AMAP and SVR).

Finally, the zoomed picture inside the left part of Figure 7.5 compares the max-affine estimators by showing the tip of the excess risks for AMAP, LSPA,

and CAP with their standard deviations. To improve performance, AMAP uses about twice as many hyperplanes (28 ± 3) than LSPA (15 ± 3) and CAP (16 ± 1), which explains its longer training time compared to CAP.

7.3 Stochastic programming problems

In this section we use max-affine estimators to approximate the cost-to-go functions of convex stochastic programming (SP) problems. These SP problems were defined in Section 2.3 by (2.3), and they require the computation of $\pi_1(\boldsymbol{x}_0, \boldsymbol{z}_0)$, where

$$\begin{aligned} \pi_t(\boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}) &\doteq \operatorname*{argmin}_{\boldsymbol{x}_t \in \mathbb{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})} J_t(\boldsymbol{x}_t)\,, \\ J_t(\boldsymbol{x}_t) &= \mathbb{E}\big[c_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) + J_{t+1}(\pi_{t+1}(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t))\big]\,, \end{aligned} \tag{7.1}$$

for all $t = 1, \ldots, T$, and $\pi_{T+1}(\cdot, \cdot) \doteq \perp$ with $J_{T+1}(\perp) = 0$. The sequence $\pi \doteq (\pi_1, \ldots, \pi_T)$ represents an optimal policy.

We only consider SP problems with convex polyhedral decision constraints written as $\mathbb{X}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\mathcal{Z}}_t) = \{\boldsymbol{x}_{t+1} : Q_{t+1}\boldsymbol{x}_{t+1} + W_{t+1}(\boldsymbol{\mathcal{Z}}_t)\boldsymbol{x}_t \leq \boldsymbol{c}_{t+1}(\boldsymbol{\mathcal{Z}}_t)\}$ which are non-empty for all possible realizations of $\boldsymbol{x}_t$ and $\boldsymbol{\mathcal{Z}}_t$. As the coefficient Q_{t+1} of the decision variable $\boldsymbol{x}_{t+1}$ is independent of random disturbances $\boldsymbol{\mathcal{Z}}_t$ and the constraint $\boldsymbol{x}_t \in \mathbb{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})$ for policy π_t is feasible for any $\boldsymbol{x}_{t-1}$ and $\boldsymbol{z}_{t-1}$, these SP problems are said to have a fixed, relatively complete recourse (Shapiro et al., 2009, Section 2.1.3). We will exploit the fixed recourse property for sampling (7.2), while relatively complete recourse allows us not to deal with infeasibility issues which could make these problems very difficult.[6]

In order to construct approximations $\hat{J}_t$ to the cost-to-go function J_t, we need "realizable" decision samples $\boldsymbol{x}_{t,i}$ at stage t. We generate these incrementally during a forward pass for $t = 1, 2, \ldots, T$, where given m disturbances $\boldsymbol{z}_{t,1:m} \doteq \{\boldsymbol{z}_{t,j} : j = 1, \ldots, m\}$ and n decisions $\boldsymbol{x}_{t,1:n} \doteq \{\boldsymbol{x}_{t,i} : i = 1, \ldots, n\}$ at stage t, we uniformly sample new decisions for stage $t+1$ from the set

$$\hat{\mathbb{X}}_{t+1} \doteq \Big\{\boldsymbol{x}_{t+1} : Q_{t+1}\boldsymbol{x}_{t+1} \leq \max_{\substack{i=1,\ldots,n,\\ j=1,\ldots,m}} \big\{\boldsymbol{c}_{t+1}(\boldsymbol{z}_{t,j}) - W_{t+1}(\boldsymbol{z}_{t,j})\boldsymbol{x}_{t,i}\big\}\Big\}\,, \tag{7.2}$$

[6] Infeasible constraints can be equivalently modeled by cost-to-go functions assigning infinite value for points outside of the feasible region. Then notice that the estimation of functions with infinite magnitude and slope can be arbitrarily hard.

where the maximum is taken component-wise. To uniformly sample the convex polytope $\hat{\mathbb{X}}_{t+1}$, we use the Hit-and-run Markov-Chain Monte-Carlo algorithm (Smith, 1984, or see Vempala, 2005) by generating 100 chains (to reduce sample correlation) each started from the average of 10 randomly generated border points, and discarding d_{t+1}^2 samples on each chain during the burn-in phase,[7] where d_{t+1} is the dimension of $\boldsymbol{x}_{t+1}$.

Then, during a backward pass for $t = T, T-1, \ldots, 1$, we recursively use the cost-to-go estimate of the previous stage $\hat{J}_{t+1}(\cdot)$ to approximate the values of the cost-to-go function J_t at the decision samples $\boldsymbol{x}_{t,i}$ generated during the forward pass, that is

$$J_t(\boldsymbol{x}_{t,i}) \approx y_{t,i} \doteq \frac{1}{m}\sum_{j=1}^{m} c_t(\boldsymbol{x}_{t,i}, \boldsymbol{z}_{t,j}) + \min_{\boldsymbol{x}_{t+1} \in \mathbb{X}_{t+1}(\boldsymbol{x}_{t,i}, \boldsymbol{z}_{t,j})} \hat{J}_{t+1}(\boldsymbol{x}_{t+1}) \tag{7.3}$$

for all $t = 1, \ldots, T$, and $\hat{J}_{T+1}(\cdot) \equiv 0$. This allows us to set up a regression problem with data $\{(\boldsymbol{x}_{t,i}, y_{t,i}) : i = 1, \ldots, n\}$ which is used to construct an estimate $\hat{J}_t(\cdot)$ of the cost-to-go function $J_t(\cdot)$.

We call the resulting method, shown as Algorithm 7.1, full approximate dynamic programming (fADP) because global approximations to the cost-to-go functions are constructed.

When the cost-to-go functions J_t are approximated by "reasonably sized" convex piecewise linear representations $\hat{J}_t$, the minimization problem in (7.3) can be solved efficiently by linear programming (LP). In the following sections, we exploit the speed of LP solvers for fADP using either AMAP or CAP as the regression procedure. Then, the computational cost to run Algorithm 7.1 is mostly realized by solving Tnm LP tasks for (7.3) and training T estimators using the regression algorithm REG. Although using max-affine LSEs (Section 6.1) for REG is fast enough for sample sizes up to $n \leq 1500$, the provided max-affine representations using $\Theta(n)$ number of hyperplanes turn the LP tasks too costly to solve and preventing Algorithm 7.1 from terminating within a reasonable time (at least using our hardware and implementation).

[7] The choice was inspired by the $O(d_{t+1}^2)$ mixing result of the Hit-and-run algorithm (Lovász, 1999).

1. **input:** SP problem, number of trajectories n,
   number of evaluations m, regression algorithm REG
2. $\boldsymbol{x}_{0,i} \leftarrow \boldsymbol{x}_0$ for all $i = 1, \ldots, n$
3. $\boldsymbol{z}_{0,j} \leftarrow \boldsymbol{z}_0$ for all $j = 1, \ldots, m$
   {forward pass}
4. **for all** $t = 0, 1, \ldots, T-1$ **do**
5. sample $\boldsymbol{x}_{t+1,1:n}$ from $\hat{\mathbb{X}}_{t+1}$ by (7.2) using $\boldsymbol{x}_{t,1:n}$, and $\boldsymbol{z}_{t,1:m}$
6. sample $\boldsymbol{z}_{t+1,1:m}$ from the distribution of $\boldsymbol{\mathcal{Z}}_{t+1}$
7. **end for**
   {backward pass}
8. $\hat{J}_{T+1}(\cdot) \leftarrow 0$
9. **for all** $t = T, T-1, \ldots, 1$ **do**
10. compute $y_{t,1}, \ldots, y_{t,n}$ by (7.3) using $\hat{J}_{t+1}(\cdot)$, $\boldsymbol{x}_{t,1:n}$, and $\boldsymbol{z}_{t,1:m}$
11. $\hat{J}_t(\cdot) \leftarrow \mathrm{REG}(\{(\boldsymbol{x}_{t,i}, y_{t,i}) : i = 1, \ldots, n\})$
12. **end for**
13. **output:** cost-to-go functions $\hat{J}_t(\cdot)$, $t = 1, \ldots, T$

Algorithm 7.1: Full approximate dynamic programming (fADP) for constructing global approximations to the cost-to-go functions of a stochastic programming problem.

The situation is even worse for nonconvex estimators REG for which LP has to be replaced for (7.3) by a much slower nonlinear constrained optimization method using perhaps randomized restarts to minimize the chance of being trapped in a local minima. Furthermore, the MARS and SVR implementations we have access to, do not provide gradient information, so minimization over these representations require an even slower gradient-free nonlinear optimization technique. With this, both MARS and SVR are impractical to use in our test problems.

To evaluate the fADP algorithm on a SP problem for some regression algorithm REG, we evaluate the greedy policy with respect to the learned cost-to-go functions $\{\hat{J}_t : t = 1, \ldots, T\}$. More precisely, we run $\hat{\pi} \doteq (\hat{\pi}_1, \ldots, \hat{\pi}_T)$ with $\hat{\pi}_t(\boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}) \in \operatorname{argmin}_{\boldsymbol{x}_t \in \mathbb{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})} \hat{J}_t(\boldsymbol{x}_t)$ on 1000 episodes, and record the average revenue (negative cost) as $\mathrm{REV} \doteq -\frac{1}{1000} \sum_{e=1}^{1000} \sum_{t=1}^{T} c_t\big(\boldsymbol{x}_t^{(e)}, \boldsymbol{z}_t^{(e)}\big)$ over the episodes' trajectories $\big\{\big(\boldsymbol{x}_t^{(e)}, \boldsymbol{z}_t^{(e)}\big) : t = 1, \ldots, T\big\}$, $e = 1, \ldots, 1000$. We repeat this experiment 100 times for each regression algorithm REG,[8] and

[8]The random seeds are kept synchronized, so every algorithm is evaluated on the same set of trajectories. Furthermore, fADP algorithms with the same n and m parameters use the same training data $\boldsymbol{x}_{t,1:n}$ and $\boldsymbol{z}_{t,1:m}$ for all $t = 0, \ldots, T$.

show the mean and standard deviation of the resulting sample.

7.3.1 Energy storage optimization

In this section we consider the energy storage optimization problem of Section 2.3.1 using a solar energy source, a discounted nightly electricity pricing model (Economy 7 tariff), and planning for a two days horizon on hourly basis ($T \doteq 48$). Retail and wholesale price curves, along with the electricity demand and energy production distributions of this model are shown on Figure 7.6 for the two consecutive sunny days whose data is used in the experiments. The

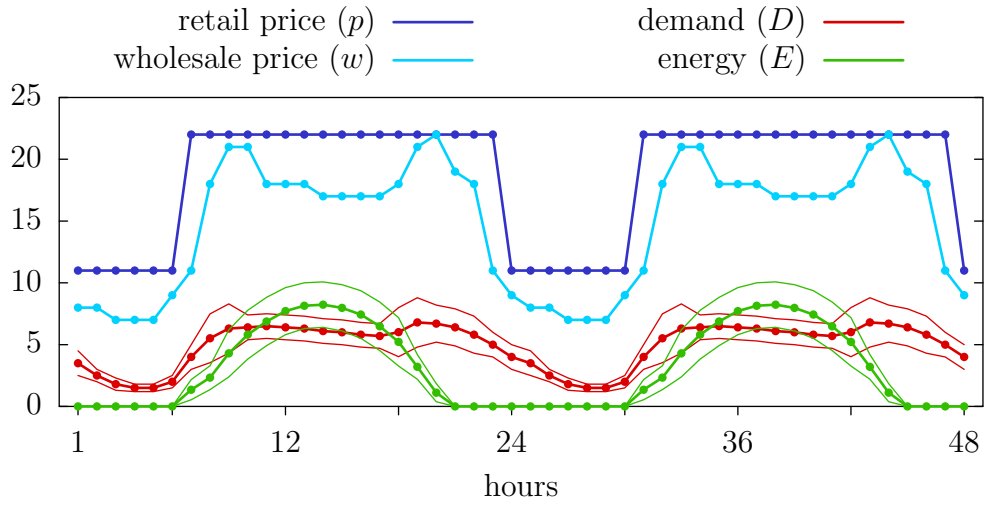

Figure 7.6: Parameters of the energy storage optimization problem. Retail (p) and wholesale (w) price curves, energy demand (D) and production (E) distributions with mean and standard deviation are shown for two-day long period.

distributions are truncated normal with support $D \in [0, 15]$ and $E \in [0, 12]$ for demand and energy production, respectively and the storage has capacity $s_{\max} \doteq 20$ with charge and discharge rates $r_c \doteq 4$ and $r_d \doteq 10$, respectively. The model is initialized by $s_0 \doteq 0$, $d_1 \doteq \mathbb{E}[D_1]$, and $e_1 \doteq \mathbb{E}[E_1]$.

To evaluate fADP on this problem, we use the CAP and AMAP convex regression techniques with multiple configurations determined by the number of trajectories n generated for training (which is the sample size for the regression tasks as well), and the number of evaluations m used to approximate the cost-to-go functions J_t at a single point (7.3). The result is presented on Figure 7.7, which also includes a "heuristic" algorithm to provide a baseline. The

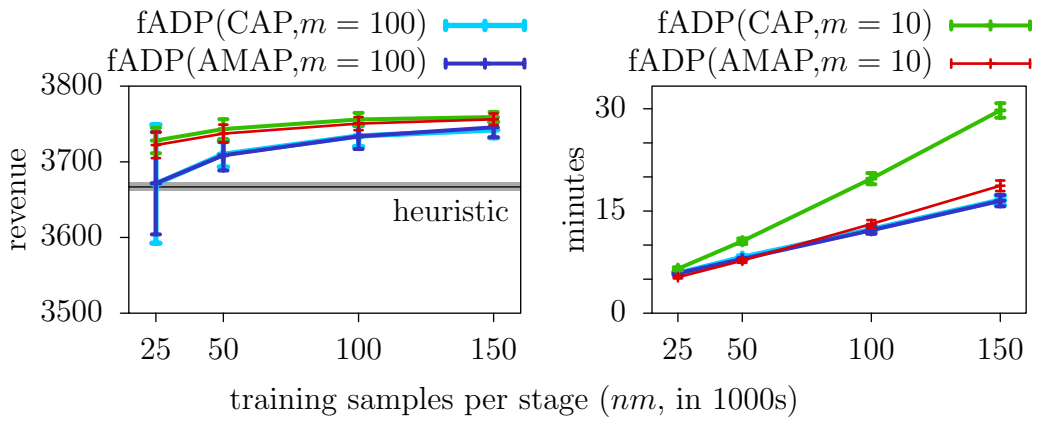

Figure 7.7: Energy storage optimization results for the fADP algorithm using AMAP or CAP as the inner convex regression procedure. Results show the total revenue (negative cost), and the training time in minutes for trajectories n and cost-to-go evaluations m.

heuristic uses a fixed policy of immediately selling the solar energy preferrably for demand ($f_{\text{ed}} \to \max$, $f_{\text{eg}} \geq 0$, $f_{\text{es}} = 0$), selling from the battery during the day when demand still allows ($f_{\text{gs}} = 0$, $f_{\text{sd}} \geq 0$), charging the battery overnight ($f_{\text{gs}} \to \max$, $f_{\text{sd}} = 0$), and selling everything close to the last stage ($f_{\text{gs}} = 0$, $f_{\text{sd}} \to \max$, $f_{\text{sg}} \to \max$). This policy is much better than the *optimal policy without storage*[9] which scores 3227 ± 6.

The results of Figure 7.7 show that fADP using convex regression significantly outperforms the heuristic baseline algorithm when the sample size is large enough. The regression algorithms prefer larger sample sizes n to better sample quality m, although this significantly increases the computation time for CAP to provide a small revenue increase comapred to AMAP.

7.3.2 Beer brewery optimization

Now consider operating a beer brewery as described in Section 2.3.2 by planning for a 48 weeks horizon on a fortnight basis ($T \doteq 24$). The demand distributions for lager and ale beers are shown on Figure 7.8 with mean and standard deviation. Both distributions are truncated normal with support $[0.1, 12]$. The cost vectors are set to fixed values for all $t = 1, \ldots, T$ as

[9]Because $p \geq w$, the optimal policy for $s_{\max} = 0$ minimizes the immediate cost by $f_{\text{ed}} \doteq \min\{E, D\}$, $f_{\text{eg}} \doteq \max\{0, E - f_{\text{ed}}\}$, and $f_{\text{es}} \doteq f_{\text{sd}} \doteq f_{\text{gs}} \doteq f_{\text{sg}} \doteq 0$.

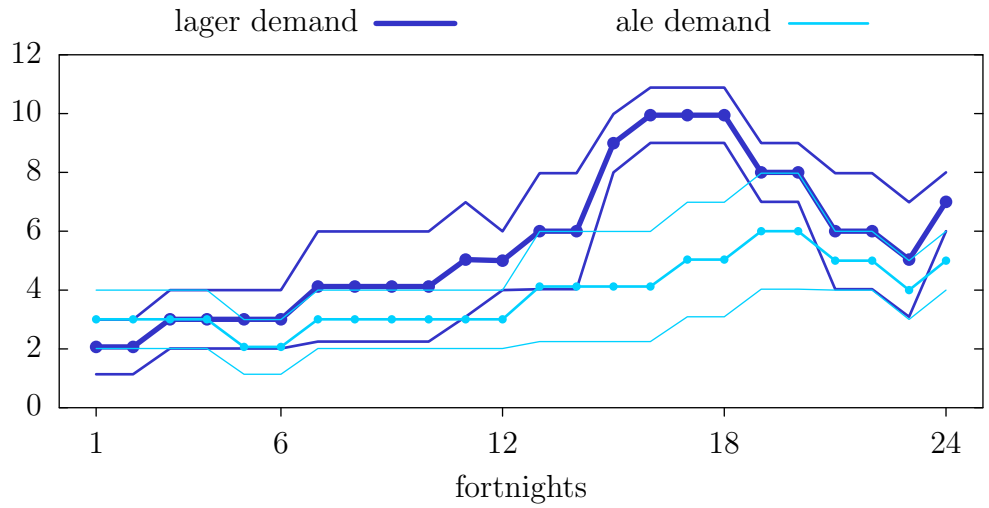

Figure 7.8: Lager and ale beer demand distributions for the beer brewery optimization problem with mean and standard deviation are shown for a 48 weeks horizon.

$\boldsymbol{h}_t \doteq [1\ 0.2\ 0.2\ 1\ 2\ 1\ 1\ 1\ 2]^\top$, $\boldsymbol{c}_t \doteq [20\ 10\ 5\ 1\ 1]^\top$, and $\boldsymbol{r}_t \doteq [90\ 50]^\top$. Furthermore, the ingredient requirement vectors for brewing are $\boldsymbol{b}_a \doteq [1\ 1\ 1]^\top$ and $\boldsymbol{b}_l \doteq [0.5\ 0.9\ 0.8]^\top$ for ale and lager, respectively, and the capacity vector is $\boldsymbol{k} \doteq [10\ 10\ 10\ 10\ \infty\ 10\ \infty\ \infty\ \infty]^\top$ to ensure the feasibility (relatively complete recourse) requirements, as discussed at the beginning of Section 7.3.

Similar to the energy optimization case, we use the CAP and AMAP estimators for fADP with various trajectory set sizes n and cost-to-go evaluation numbers m. The results are presented on Figure 7.9. In this case, AMAP improves the performance significantly by collecting revenue over 4100 compared to CAP which stays around 3600.

However, the result also shows that the running time of AMAP also become significantly larger than CAP. Based on the experience of Section 7.1, the increased running time of AMAP indicates that large max-affine models are needed to improve the accuracy of the cost-to-go approximations, which increases the computational cost of the LP tasks of (7.3), and eventually slows down the fADP algorithm significantly. Notice that using larger trajectory sets n for AMAP provide better quality at the beginning, but the improved sample quality with $m = 100$ eventually achieves the same using significantly less computational resources.

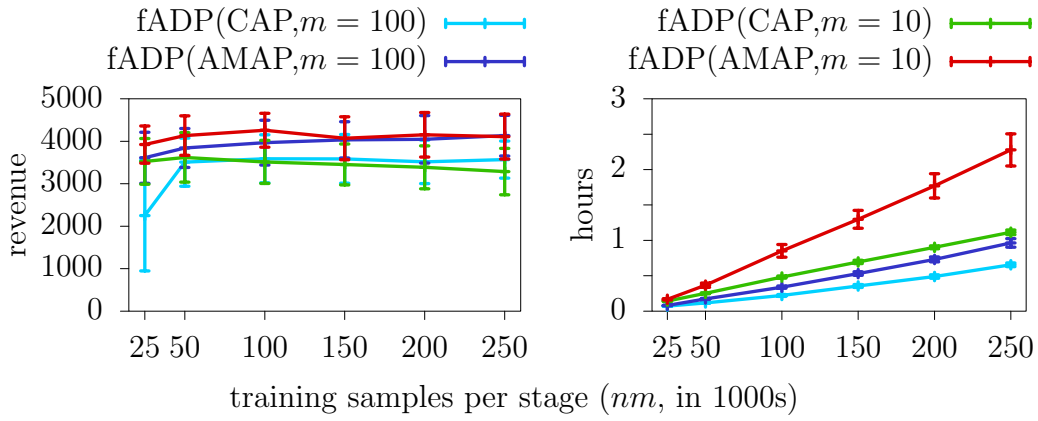

Figure 7.9: Beer brewery optimization results for fADP algorithm using AMAP and CAP convex regression to approximate the convex cost-to-go functions. Results show the revenue (negative cost), and the training time in minutes for trajectories n and cost-to-go evaluations m.

Finally, we point out that scaling up the fADP algorithm for larger problems would require many improvements. One of these could be using more expressive convex piecewise-linear representations (see Section 8.2), which might compress the LP tasks enough for the current solvers. Another important step could be to use a LP algorithm which can more efficiently solve large number of similar LP problems with different right hand sides (see for example Gassmann and Wallace, 1996). And eventually, it would become inevitable to localize cost-to-go approximations to a fraction of the decision space, perhaps by running fADP iteratively alternating between sampling and estimation phases, and exploring at the boundary of the accurately approximated region to find and avoid delayed rewards and costs. However, this goes far beyond the scope of this thesis, and is thus left as future work.

Chapter 8

Conclusions and future work

In this chapter we review the content of the previous chapters in the context of a few selected future research directions in convex regression.

8.1 Beyond convexity

As our main regression result (Theorem 3.2) is general and valid beyond the scope of convex regression, it is likely that the theoretical results of Chapter 5 can be generalized to the estimation of Lipschitz continuous functions. To prove such result, one could consider using maxima of minima of affine (max-min-affine) estimates (Bagirov et al., 2010, Section 2.2) which can represent any continuous piecewise linear function (Gorokhovik et al., 1994).

As max-min-affine functions can be also rewritten as the difference of two max-affine functions, these estimates are in the class of delta-convex mappings (difference of two convex maps), which can uniformly approximate any continuous function (Bačák and Borwein, 2011, Proposition 2.2).

Then, the "only" work left is to prove an approximation rate for the class of max-min-affine functions (or equivalently for the difference of two max-affine maps) to Lipschitz continuous targets and replace Lemma 5.2 in the analysis of Section 5.4. We believe this method is capable to deliver the minimax convergence rate $n^{-2/(2+d)}$ (Györfi et al., 2002, Theorem 3.2 with $p = 1$) up to logarithmic factors for max-min-affine estimators if their size parameter (number of hyperplanes) are chosen appropriately, similarly as for the convex case (Theorem 5.6).

8.2 Sum-max-affine representations

For the estimation of convex functions, we focused on max-affine representations for which we have a theoretical analysis (Chapter 5), convex training algorithms with a theoretical guarantee on their sample complexity (Section 6.1), and scalable heuristic training methods (Section 6.2).

However, max-affine representations can be very inefficient for approximating "important" convex functions. For example, the Manhattan norm $\boldsymbol{x} \mapsto \|W\boldsymbol{x}\|_1$ as a target (with some scaling matrix $W \in \mathbb{R}^{d\times d}$) can be represented exactly only by the maximum of 2^d affine maps. Perhaps for a similar reason, the empirical results of Section 7.1 showed that max-affine representations are not efficient for quadratic and sum-max-affine targets which are important for many applications (Sections 7.2 and 7.3).

Now consider the sum of S max-affine functions with $K_1, \dots, K_S$ hyperplanes given as $h(\boldsymbol{x}) \doteq \sum_{s=1}^{S} \max_{k=1,\dots,K_s} \boldsymbol{a}_{sk}^\top \boldsymbol{x} + b_{sk}$. Clearly, the class of sum-max-affine functions includes the set of max-affine functions as a special case by $S = 1$, hence ERM estimators (3.1) over sum-max-affine functions also enjoy near-minimax rates (Theorem 5.6) as long as $\sum_{s=1}^{S} K_s = \Theta(n^{d/(d+4)})$ and $S = 1$ is allowed.

Further, observe that the Manhattan norm is a sum-max-affine function using only $2d$ planes as $\|W\boldsymbol{x}\|_1 \doteq \sum_{s=1}^{d} |\boldsymbol{w}_s^\top \boldsymbol{x}| = \sum_{s=1}^{d} \max\{-\boldsymbol{w}_s^\top \boldsymbol{x}, \boldsymbol{w}_s^\top \boldsymbol{x}\}$, where $\boldsymbol{w}_s$ is the s-th row of W, that is $W^\top \doteq [\boldsymbol{w}_1 \dots \boldsymbol{w}_d]$. This is a huge improvement compared to the 2^d hyperplanes used by max-affine maps.

More generally, the approximation result of max-affine maps (Lemma 5.2) could be extended to convex targets having an additive structure written as $f(\boldsymbol{x}) \doteq \sum_{s=1}^{S} f_s(W_s\, \boldsymbol{x})$ with each $f_s : \mathbb{R}^{d_s} \to \mathbb{R}$ being convex and $W_s \in \mathbb{R}^{d_s \times d}$, providing an approximation bound $O\big(\sum_{s=1}^{S} K_s^{-2/d_s}\big)$. Then notice that by rewriting a convex quadratic norm as $\|\boldsymbol{x}\|_Q^2 \doteq \boldsymbol{x}^\top Q \boldsymbol{x} = \sum_{s=1}^{d} (\boldsymbol{q}_s^\top \boldsymbol{x})^2$ for some positive semi-definite matrix $0 \preceq Q \doteq \sum_{s=1}^{d} \boldsymbol{q}_s \boldsymbol{q}_s^\top \in \mathbb{R}^{d\times d}$, such approximation result provides an $O(dK^{-2})$ bound, which is again a huge improvement compared to the $O(K^{-2/d})$ rate of max-affine maps.

However, we are not aware of any training algorithm for sum-max-affine estimators which could adaptively set the S and the $K_1, \ldots, K_S$ parameters based on the training data $\mathcal{D}_n$. The only work we know in this direction is due to Hannah and Dunson (2012) who studied various ensemble methods (bagging, smearing, and random forests) to build sum-max-affine estimators combining a fixed S number of max-affine maps trained by CAP (Section 6.2.2). As these ensemble techniques require a relatively large S, they do not build a compact representation, so they can be sensitive to overfitting and provide computationally too expensive models for many applications including convex stochastic programming (Section 7.3). Hence, we believe that adaptive training of compact sum-max-affine estimators are likely to significantly improve the effectiveness and applicability of conex regression algorithms in the future.

8.3 Searching convex partitions

During the discussion of heuristic max-affine estimators (Section 6.2), the estimates of the considered training algorihms (LSPA, CAP and AMAP) could not be substiantially improved by the partitioned LSE convex reformulation (Section 6.1.1) over the induced partition. Hence, it seems that there exist partitions $P = \{\mathcal{C}_k : k = 1, \ldots, K\}$ for which the cellwise least squares fit (that is ridge regression on $\{(\boldsymbol{\mathcal{X}}_i, \mathcal{Y}_i) : i \in \mathcal{C}_k\}$ with a small β for each $k = 1, \ldots, K$) is "nearly" a local optimum.

To investigate this issue, consider the following smooth approximation of max-affine functions:

Lemma 8.1. *Let* $h(\boldsymbol{z}) \doteq \max_{k=1,\ldots,K} \boldsymbol{w}_k^\top \boldsymbol{z}$ *and* $\hat{h}_t(\boldsymbol{z}) \doteq t \ln \sum_{k=1}^K \exp(\boldsymbol{w}_k^\top \boldsymbol{z}/t)$ *for some* $t > 0$. *Then* $0 \leq \hat{h}_t(\boldsymbol{z}) - h(\boldsymbol{z}) \leq t \ln(K)$ *holds for all* $\boldsymbol{z}$.

Proof. Fix $\boldsymbol{z}$ arbitrarily and notice that $h(\boldsymbol{z}) \leq \hat{h}_t(\boldsymbol{z})$ simply follows from Jensen's inequality. For the other side, let $\boldsymbol{w}_*$ be such that $\boldsymbol{w}_*^\top \boldsymbol{z} = h(\boldsymbol{z})$ and observe that $\hat{h}_t(\boldsymbol{z}) = t \ln \left(e^{\boldsymbol{w}_*^\top \boldsymbol{z}/t} \sum_{k=1}^K e^{(\boldsymbol{w}_k - \boldsymbol{w}_*)^\top \boldsymbol{z}/t} \right) \leq h(\boldsymbol{z}) + t \ln(K)$. ■

Then, by replacing the max-affine functions $f(\boldsymbol{x}) \doteq \max_{k=1,\ldots,K} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k$ of (6.14) with $\hat{f}_t(\boldsymbol{x}) \doteq t \ln \sum_{k=1}^K e^{\boldsymbol{w}_k^\top \boldsymbol{z}/t}$ using $\boldsymbol{z}^\top \doteq [1 \ \boldsymbol{x}^\top]$ and $\boldsymbol{w}_k^\top \doteq [b_k \ \boldsymbol{a}_k^\top]$, we

can compute the gradient of the modified objective $\hat{L}_t \doteq \sum_{i=1}^{n}(\hat{f}_t(\boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i)^2$ with respect to the estimate parameters $\{\boldsymbol{w}_k : k = 1, \ldots, K\}$ as

$$\nabla_{\boldsymbol{w}_k} \hat{L}_t = 2 \sum_{i=1}^{n} \big(\hat{f}_t(\boldsymbol{\mathcal{X}}_i) - \mathcal{Y}_i\big) \frac{\boldsymbol{\mathcal{Z}}_i \, e^{\boldsymbol{w}_k^\top \boldsymbol{\mathcal{Z}}_i / t}}{\sum_{l=1}^{K} e^{\boldsymbol{w}_l^\top \boldsymbol{\mathcal{Z}}_i / t}} \to 2 \sum_{i \in \mathcal{C}_k} \big(\boldsymbol{w}_k^\top \boldsymbol{\mathcal{Z}}_i - \mathcal{Y}_i\big) \mathcal{W}_{ik}(f) \boldsymbol{\mathcal{Z}}_i \quad \text{as } t \to 0 \,,$$

where $\boldsymbol{\mathcal{Z}}_i^\top \doteq [1 \ \boldsymbol{\mathcal{X}}_i^\top]$, $\mathcal{W}_{ik}(f) = |\{l = 1, \ldots, K : \boldsymbol{w}_l^\top \boldsymbol{\mathcal{Z}}_i = f(\boldsymbol{\mathcal{X}}_i)\}|^{-1}$, and $P = \{\mathcal{C}_k : k = 1, \ldots, K\}$ is the induced partition of f. As breaking the ties can be done by losing an arbitrarily small accuracy, we can assume without loss of generality that $\mathcal{W}_{ik}(f) = 1$ for all i and k. Then setting the right side to zero and solving it for $\boldsymbol{w}_k$ (by adding a small regularizer $\beta \boldsymbol{w}_k$), we recover a ridge regression estimate (Section 4.2.2) over the cell $\mathcal{C}_k$, which is the cellwise least squares fit over the induced partition, and also the "preferred" estimate type of the training mehods in Section 6.2.

Although this limiting analysis is not precise (it does not imply that such optimum exist), it still suggests that our empirical observations about "nearly locally optimal" cellwise fitted max-affine estimators might admit a general rule. Research in this direction could perhaps answer how much we might lose by limiting the search of (6.14) for these cellwise fitted estimators and provide further information on the "quality" of max-affine training algorithms in Section 6.2.

In summary, while the thesis advanced our knowledge of nonparametric shape constrained regression and sharpened the tools available for studying these and related problems, much work remains to be done, especially in connection to designing estimators which are efficient *both* in terms of their computational cost and they way they use the samples. Such better techniques, when used in multistage stochastic programming as outlined in this thesis perhaps with clever sampling techniques, have the potential to give rise to exciting novel, practical applications that may have major impact in several industries. It shall be interesting to see whether these will indeed happen.

Bibliography

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77.

Aybat, N. S. and Wang, Z. (2014). A parallel method for large scale convex regression problems. *IEEE Conference on Decision and Control (CDC)*.

Bagirov, A., Clausen, C., and Kohler, M. (2010). An algorithm for the estimation of a regression function by continuous piecewise linear functions. *Computational Optimization and Applications*, 45(1):159–179.

Balázs, G., György, A., and Szepesvári, C. (2015). Near-optimal max-affine estimators for convex regression. In Lebanon, G. and Vishwanathan, S., editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *JMLR W&CP*.

Balázs, G., György, A., and Szepesvári, C. (2016a). Chaining bounds for empirical risk minimization. http://arxiv.org/abs/1609.01872v1.

Balázs, G., György, A., and Szepesvári, C. (2016b). Max-affine estimators for convex stochastic programming. http://arxiv.org/abs/1609.06331v1.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334.

Bačák, M. and Borwein, J. M. (2011). On difference convexity of locally lipschitz functions. *Optimization: A Journal of Mathematical Programming and Operations Research*, 60(8–9):961–978.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition.

Bertsekas, D. P. (1996). *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific.

Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control, Volume I*. Athena Scientific, 3rd edition.

Bertsekas, D. P. and Tsitsiklis, J. N. (1997). *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific.

Birge, J. R. and Louveaux, F. (2011). *Introduction to Stochastic Programming*. Springer.

Bisschop, J. (2016). *AIMMS Optimization Modeling*. http://www.aimms.com.

Boucheron, S., Lugosi, G., and Massart, P. (2012). *Concentration Inequalities: A nonasymptotic theory of independence*. Clarendon Press.

Boyd, S., Kim, S.-J., Vandenberghe, L., and Hassibi, A. (2007). A tutorial on geometric programming. *Optimization and Engineering*, 8(1):67–127.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bronshteyn, E. M. and Ivanov, L. D. (1975). The approximation of convex sets by polyhedra. *Siberian Mathematical Journal*, 16(5):852–853.

Buldygin, V. V. and Kozachenko, Y. V. (2000). *Metric characterization of random variables and random processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI.

Cai, Y. (2009). *Dynamic programming and its application in economics and finance*. PhD thesis, Stanford University.

Cai, Y. and Judd, K. L. (2013). Shape-preserving dynamic programming. *Mathematical Methods of Operations Research*, 77(3):407–421.

Cesa-Bianchi, N. and Lugosi, G. (1999). Minimax regret under log loss for general classes of experts. In Ben-David, S. and Long, P. M., editors, *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, pages 12–18.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cooper, D. A. (1995). Learning lipschitz functions. *International Journal of Computer Mathematics*, 59(1–2):15–26.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, 2nd edition.

Drela, M. (1989). XFOIL: An analysis and design system for low reynolds number airfoils. In *Lecture Notes in Engineering*, volume 54 of *Low Reynolds Number Aerodynamics*, pages 1–12. Springer-Verlag. Software: http://web.mit.edu/drela/Public/web/xfoil/.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141.

Gao, F. and Wellner, J. A. (2015). Entropy of convex functions on $\mathbb{R}^d$. http://arxiv.org/abs/1502.01752v2.

Gassmann, H. I. and Wallace, S. W. (1996). Solving linear programs with multiple right-hand sides: Pricing and ordering schemes. *Annals of Operation Research*, 64:237–259.

Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations*. The Johns Hopkins University Press, 3rd edition.

Gorokhovik, V. V., Zorko, O. I., and Birkhoff, G. (1994). Piecewise affine functions and polyhedral sets. *Optimization: A Journal of Mathematical Programming and Operations Research*, 31(3):209–221.

Guntuboyina, A. (2011). *Minimax Lower Bounds*. PhD thesis, Yale University.

Guntuboyina, A. and Sen, B. (2013). Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag.

Györfi, L. and Wegkamp, M. (2008). Quantization for nonparametric regression. *IEEE Transactions on Information Theory*, 54(2):867–874.

Han, Q. and Wellner, J. A. (2016). Multivariate convex regression: global risk bounds and adaptation. https://arxiv.org/abs/1601.06844v1.

Hannah, L. A. and Dunson, D. B. (2011). Approximate dynamic programming for storage problems. In Getoor, L. and Scheffer, T., editors, *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pages 337–344.

Hannah, L. A. and Dunson, D. B. (2012). Ensemble methods for convex regression with applications to geometric programming based circuit design. In Langford, J. and Pineau, J., editors, *Proceedings of The 29th International Conference on Machine Learning (ICML)*, pages 369–376.

Hannah, L. A. and Dunson, D. B. (2013). Multivariate convex regression with adaptive partitioning. *Journal of Machine Learning Research*, 14:3261–3294.

Hannah, L. A., Powell, W. B., and Dunson, D. B. (2014). Semiconvex regression for metamodeling-based optimization. *SIAM Journal on Optimization*, 24(2):573–597.

Hoburg, W. and Abbeel, P. (2014). Geometric programming for aircraft design optimization. *American Institute of Aeronautics and Astronautics*, 52(11):2414–2426.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holloway, C. A. (1979). On the estimation of convex functions. *Operations Research*, 27(2):401–407.

Hsu, D., Kakade, S. M., and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14:569–600.

Huang, R. and Szepesvári, C. (2014). A finite-sample generalization bound for semiparametric regression: Partially linear models. In Kaski, S. and Corander, J., editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33 of *JMLR W&CP*.

Jekabsons, G. (2016). ARESLab: Adaptive regression splines toolbox for Matlab/Octave (ver. 1.10.3). http://www.cs.rtu.lv/jekabsons/.

Jiang, D. R. and Powell, W. B. (2015). An approximate dynamic programming algorithm for monotone value functions. *CoRR*. http://arxiv.org/abs/1401.1590v6.

Joulani, P., György, A., and Szepesvári, C. (2015). Fast cross-validation for incremental learning. *International Joint Conference on Artificial Intelligence*, pages 3597–3604.

Keshavarz, A. (2012). *Convex Methods for Approximate Dynamic Programming*. PhD thesis, Stanford University.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *The Econometrics Journal*, 11(2):308–325.

Lecué, G. and Mendelson, S. (2013). Learning subgaussian classes: Upper and minimax bounds. *CoRR*. http://arxiv.org/abs/1305.4825v1.

Lee, C.-Y., Johnson, A. L., Moreno-Centeno, E., and Kuosmanen, T. (2013). A more efficient algorithm for convex nonparametric least squares. *European Journal of Operational Research*, 227(2):391–400.

Liang, T., Rakhlin, A., and Sridharan, K. (2015). Learning with squared loss: Localization through offset rademacher complexity. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of The 28th Conference on Learning Theory (COLT)*, volume 40 of *JMLR W&CP*.

Lim, E. (2014). On convergence rates of convex regression in multiple dimensions. *INFORMS Journal of Computing*, 26(3):616–628.

Lim, E. and Glynn, P. W. (2012). Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208.

Lovász, L. (1999). Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461.

Magnani, A. and Boyd, S. P. (2009). Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1):1–17.

Mazumder, R., Choudhury, A., Iyengar, G., and Sen, B. (2015). A computational framework for multivariate convex regression and its variants. http://arxiv.org/abs/1509.08165.

Mendelson, S. (2014). Learning without concentration. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory (COLT)*, volume 35 of *JMLR W&CP*, pages 25–39.

Merton, R. C. (1992). *Continuous-Time Finance*. Wiley-Blackwell.

Moler, C. (1994). Benchmarks – LINPACK and MATLAB. *MATLAB Notes and News*.

MOSEK ApS (2015). *The MOSEK optimization toolbox for MATLAB manual. Version 7.0 (Revision 139)*. http://docs.mosek.com/7.0/toolbox/index.html.

Nascimento, J. and Powell, W. B. (2013). An optimal approximate dynamic programming algorithm for concave, scalar storage problems with vector-valued controls. *IEEE Transactions on Automatic Control*, 58(12):2995–3010.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2nd edition.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, 2nd edition.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Ramsey, F. L. and Schafer, D. W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury, 2nd edition. Data sets: https://cran.r-project.org/web/packages/Sleuth2/.

Rockafellar, R. T. (1972). *Convex Analysis*. Princeton University Press.

Ruszczyński, A. P. and Shapiro, A. (2003). *Stochastic Programming*. Elsevier.

Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657.

Shamir, O. (2015). The sample complexity of learning predictors with the squared loss. *Journal of Machine Learning Research*, 16:3475–3486.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on Stochastic Programming, Modeling and Theory*. Society for Industrial and Applied Mathematics and the Mathematical Programming Society.

Smith, R. L. (1984). Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308.

Stallman, R. M. et al. (2007). *GCC, the GNU Compiler Collection 4.1.2*. https://gcc.gnu.org/gcc-4.1/.

Stewart, G. W. (1980). The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409.

Sutton, R. S. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning: Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers.

The MathWorks, Inc. (2010). *MATLAB and Statistics Toolbox Release 2010b*. Natick, Massachusetts, United States. http://www.mathworks.com/products/new_products/release2010b.html.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society, Series B*, 73(3):273–282.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica*, 50(4):945–973.

Varian, H. R. (1984). The nonparametric approach to production analysis. *Econometrica*, 52(3):579–598.

Vempala, S. (2005). Geometric random walk: A survey. *Combinatorial and Computational Geometry*, 52:573–612.

Verbeek, M. (2004). *A Guide to Modern Econometrics*. John Wiley & Sons, 2nd edition. Data sets: http://feb.kuleuven.be/GME/.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579.

Appendix A

Sub-Gaussian random vectors and their Orlicz space

In this appendix we review a few important properties of sub-Gaussian random vectors and their Orlicz space equipped with the norm $\|\cdot\|_{\Psi_2}$.

Recall that a random variable $\mathcal{W}$ is called B-sub-Gaussian when it satisfies

$$\sup_{s\in\mathbb{R}} \mathbb{E}\big[e^{s(\mathcal{W}-\mathbb{E}\mathcal{W})-s^2B^2/2}\big] \leq 1\,.$$

Examples of sub-Gaussian random variables include every Gaussian random variable (see Buldygin and Kozachenko, 2000, Remark 1.3), and all bounded random variables due to Hoeffding's lemma (see Boucheron et al., 2012, Section 2.6). The basic properties of sub-Gaussian random variables are summarized by Lemma A.1.

Lemma A.1. *Let the random variables $\mathcal{W}, \mathcal{W}_1, \ldots, \mathcal{W}_n$ be centered and sub-Gaussian with $B, B_1, \ldots, B_n$, respectively. Then the following statements hold:*

(a) $\max\big\{\mathbb{P}\{\mathcal{W}\geq\gamma\}, \mathbb{P}\{\mathcal{W}\leq-\gamma\}\big\} \leq e^{-\gamma^2/(2B^2)}$, $\gamma\geq 0$,

(b) $\mathbb{P}\{|\mathcal{W}|\geq\gamma\} \leq 2e^{-\gamma^2/(2B^2)}$, $\gamma\geq 0$,

(c) $\mathbb{E}\big[|\mathcal{W}|^s\big] \leq 2(s/e)^{s/2}\,B^s$, $s>0$, *and* $\mathbb{E}\big[\mathcal{W}^{2k}\big] \leq (2^{k+1}/e)k!\,B^{2k}$, $k\in\mathbb{N}$,

(d) $\mathbb{E}\big[e^{s\mathcal{W}^2/(2B^2)}\big] \leq 1/\sqrt{1-s}$, $s\in[0,1)$,

(e) $c\mathcal{W}$ *is* $(|c|B)$*-sub-Gaussian,* $c\in\mathbb{R}$, *and* $\sum_{i=1}^n \mathcal{W}_i$ *is* $\big(\sum_{i=1}^n B_i\big)$*-sub-Gaussian,*

(f) if $\mathcal{W}_1, \ldots, \mathcal{W}_n$ *are independent,* $\sum_{i=1}^n \mathcal{W}_i$ *is sub-Gaussian with* $\sqrt{\sum_{i=1}^n B_i^2}$.

Proof. See Section 1.1 of Buldygin and Kozachenko (2000): Lemma 1.3 for (a) and (b), Lemma 1.4 for (c) with $s = 2k$ and $e(k/e)^k \le k!$, Lemma 1.6 for (d), Theorem 1.2 for (e), and Lemma 1.7 for (f). ■

The sub-Gaussian property can be also characterized by the Ψ_2 Orlicz norm, which we extend to random vectors $\boldsymbol{\mathcal{W}} \in \mathbb{R}^d$ as

$$\|\mathcal{W}\|_{\Psi_2} \doteq \inf\left\{B > 0 : \mathbb{E}[\Psi_2(\mathcal{W}/B)] \le 1\right\}$$

with $\Psi_2(\boldsymbol{x}) \doteq e^{\|\boldsymbol{x}\|^2} - 1$ and $\inf \emptyset \doteq \infty$. Notice that Lemma A.1d with $s = 1/2$ implies that every B-sub-Gaussian random variable $\mathcal{W}$ satisfies $\|\mathcal{W}\|_{\Psi_2} \le 2B$. Then, Lemma A.2 provides the opposite direction and the basic properties of $\|\cdot\|_{\Psi_2}$ extended to random vectors.

Lemma A.2. *Let $\boldsymbol{\mathcal{W}} \in \mathbb{R}^d$ be a random vector with $\|\boldsymbol{\mathcal{W}}\|_{\Psi_2} \le B$. Then the following statements hold:*

(a) $\mathbb{P}\{\|\boldsymbol{\mathcal{W}}\| \ge \gamma\} \le 2e^{-\gamma^2/B^2}$, $\gamma \ge 0$,

(b) $\mathbb{E}\big[\,\|\boldsymbol{\mathcal{W}}\|^{2s}\,\big] \le 2(s/e)^s\, B^{2s}$, $s > 0$, *and* $\mathbb{E}\big[\,\|\boldsymbol{\mathcal{W}}\|^{2k}\,\big] \le (2/e)k!\, B^{2k}$, $k \in \mathbb{N}$,

(c) $\sup_{\boldsymbol{s}\in\mathbb{R}^d} \mathbb{E}\big[e^{\boldsymbol{s}^\top(\boldsymbol{\mathcal{W}}-\mathbb{E}\boldsymbol{\mathcal{W}})-\|\boldsymbol{s}\|^2(2B^2)/2}\big] \le 1$,

(d) $\|c\boldsymbol{\mathcal{W}}\|_{\Psi_2} \le |c|B$, $c \in \mathbb{R}$, *and* $\|\sum_{i=1}^n \boldsymbol{\mathcal{W}}_i\|_{\Psi_2} \le \sum_{i=1}^n B_i$.

Proof. For (a), simply use the Chernoff bound as

$$\mathbb{P}\{\|\boldsymbol{\mathcal{W}}\| \ge \gamma\} = \mathbb{P}\big\{e^{\|\boldsymbol{\mathcal{W}}\|^2/B^2} \ge e^{\gamma^2/B^2}\big\} \le \mathbb{E}\big[e^{\|\boldsymbol{\mathcal{W}}\|^2/B^2}\big]e^{-\gamma^2/B^2} \le 2e^{-\gamma^2/B^2}\,.$$

For (b), use $x^s \le (s/e)^s\, e^x$ for $x \ge 0$, $s > 0$ with $x = \|\boldsymbol{\mathcal{W}}\|^2/B^2$, and take the expectation of both sides to get

$$\mathbb{E}\big[\,\|\boldsymbol{\mathcal{W}}\|^{2s}\,\big] \le (s/e)^s\, \mathbb{E}\big[e^{\|\boldsymbol{\mathcal{W}}\|^2/B^2}\big]B^{2s} \le 2(s/e)^s B^{2s}\,.$$

For the second part, simply use $s = k$ and $e(k/e)^k \le k!$.

For (c), fix $\boldsymbol{s} \in \mathbb{R}^d$ arbitrarily, and take an independent copy $\boldsymbol{\mathcal{W}}$, denoted by $\widehat{\boldsymbol{\mathcal{W}}}$ (so $\boldsymbol{\mathcal{W}}, \hat{\boldsymbol{\mathcal{W}}}$ are i.i.d.). Notice that $\boldsymbol{s}^\top(\boldsymbol{\mathcal{W}} - \widehat{\boldsymbol{\mathcal{W}}})$ is a symmetric random variable, so its odd moments are zero, that is $\mathbb{E}[\boldsymbol{s}^\top(\boldsymbol{\mathcal{W}} - \widehat{\boldsymbol{\mathcal{W}}})^{2k-1}] = 0$

for all $k \in \mathbb{N}$. Then, using Jensen's inequality, the exponential series expansion, the monotone convergence theorem, the Cauchy-Schwartz inequality, $(a+b)^k \leq 2^{k-1}(a^k + b^k)$, $(2k)! \geq 2^k k!^2$, and (b), we get

$$\begin{aligned}\mathbb{E}\big[e^{\boldsymbol{s}^\top(\boldsymbol{\mathcal{W}}-\mathbb{E}\boldsymbol{\mathcal{W}})}\big] \leq \mathbb{E}\big[e^{\boldsymbol{s}^\top(\boldsymbol{\mathcal{W}}-\widehat{\boldsymbol{\mathcal{W}}})}\big] &= \sum_{k=0}^{\infty} \frac{\|\boldsymbol{s}\|^{2k}\ \mathbb{E}\big[\|\boldsymbol{\mathcal{W}}-\widehat{\boldsymbol{\mathcal{W}}}\|^{2k}\big]}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} \frac{\|\boldsymbol{s}\|^{2k}\ 2^k\,\mathbb{E}[\|\boldsymbol{\mathcal{W}}\|^{2k}]}{k!^2} \leq \sum_{k=0}^{\infty} \frac{(2\,\|\boldsymbol{s}\|^2\,B^2)^k}{k!} = e^{\|\boldsymbol{s}\|^2(2B)^2/2}\,.\end{aligned}$$

For (d), simply observe the first claim by the definition of $\|\cdot\|_{\Psi_2}$. For the second part, notice that $\boldsymbol{x} \mapsto e^{\|\boldsymbol{x}\|^2}$ is a convex function, so using weights $\lambda_i = B_i^2 / \sum_{j=1}^n B_j^2$ for Jensen's inequality, we get

$$\mathbb{E}\Big[e^{(\sum_{i=1}^n \boldsymbol{\mathcal{W}}_i)^2/(\sum_{j=1}^n B_j^2)}\Big] \leq \sum_{i=1}^n \lambda_i\,\mathbb{E}\Big[e^{\boldsymbol{\mathcal{W}}_i^2/(\lambda_i \sum_{j=1}^n B_j^2)}\Big] = \sum_{i=1}^n \lambda_i\,\mathbb{E}\big[e^{\boldsymbol{\mathcal{W}}_i^2/B_i^2}\big] \leq 2\,.$$

■

Finally, Lemma A.3 proves a large deviation bound for the average of independent, centered sub-Gaussian random vectors.

Lemma A.3. *Let $\boldsymbol{\mathcal{W}}_1, \dots, \boldsymbol{\mathcal{W}}_n \in \mathbb{R}^d$ be random vectors with $\mathbb{E}[\boldsymbol{\mathcal{W}}_i] = \mathbf{0}$ and $\|\boldsymbol{\mathcal{W}}_i\|_{\Psi_2} \leq B_i$ for all $i = 1, \dots, n$. Then $\big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\mathcal{W}}_i\big\|_{\Psi_2} \leq \sqrt{\frac{8d}{n}\big(\frac{1}{n}\sum_{i=1}^n B_i^2\big)}$.*

Proof. First, write $\boldsymbol{\mathcal{W}}_i = [\mathcal{W}_{i1} \dots \mathcal{W}_{id}]^\top \in \mathbb{R}^d$ for all $i = 1, \dots, n$, and set $C \doteq 8\big(\frac{1}{n}\sum_{i=1}^n B_i^2\big)/n$. Then (c) implies that each random variable $\mathcal{W}_{ij}$ is $(\sqrt{2}\,B_i)$-sub-Gaussian. Hence, for every $j = 1, \dots, d$, Lemma A.1f used for the centered and independent random variables $\mathcal{W}_{1j}, \dots, \mathcal{W}_{nj}$ provides that $\frac{1}{n}\sum_{i=1}^n \mathcal{W}_{ij}$ is sub-Gaussian with C. Finally using Hölder's inequality, we get the claim by

$$\begin{aligned}\mathbb{E}\Big[e^{\big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\mathcal{W}}_i\big\|^2/(dC)}\Big] &= \mathbb{E}\Big[e^{\sum_{j=1}^d |\frac{1}{n}\sum_{i=1}^n \mathcal{W}_{ij}|^2/(dC)}\Big] \\ &\leq \prod_{j=1}^d \mathbb{E}\Big[e^{|\frac{1}{n}\sum_{i=1}^n \mathcal{W}_{ij}|^2/C}\Big]^{\frac{1}{d}} \leq 2\,.\end{aligned}$$

■

When $\mathcal{W}, \mathcal{Z} \in \mathbb{R}$ are sub-Gaussian, their product $\mathcal{W}\mathcal{Z}$ is a so-called "subexponential" random variable. To derive upper bounds for ERM estimators in Chapter 3, we apply Bernstein's inequality (Lemma A.4) for such random variables $\mathcal{W}\mathcal{Z}$, which requires an "appropriate" bound on the higher moments as provided by Lemma A.5. Here, "appropriate" means that the bound has to scale with the second moment of one multiplier, say $\mathcal{W}$, replacing $\|\mathcal{W}\|_{\Psi_2}^2$ by $\mathbb{E}[\mathcal{W}^2]$. The price we pay for this is only logarithmic in the kurtosis $\mathbb{K}_0[\mathcal{W}] \doteq \mathbb{E}[\mathcal{W}^4]/\mathbb{E}[\mathcal{W}^2]^2$, which is crucial to our analysis deriving near-minimax upper bounds for arbitrary sub-Gaussian regression problems.

Lemma A.4 (Bernstein's lemma)**.** *Let $\mathcal{W}$ be a real valued random variable such that $\mathbb{E}\big[|\mathcal{W}|^k\big] \le (k!/2)v^2c^{k-2}$ for all $2 \le k \in \mathbb{N}$. Then, for all $|s| < 1/c$,*

$$\ln \mathbb{E}\Big[e^{s(\mathbb{E}[\mathcal{W}]-\mathcal{W})}\Big] \le \frac{s^2\, v^2}{2(1-|s|c)}\,.$$

Proof. See, for example Boucheron et al. (2012, Theorem 2.10) with $n = 1$, and use $X_1 = -\mathcal{W}$, $\lambda = -s$ when $s < 0$. ∎

Lemma A.5. *Let $\mathcal{W}$, $\mathcal{Z}$ be two random variables such that $\mathbb{E}[\mathcal{W}^2] > 0$, and $\|\mathcal{W}\|_{\Psi_2} \le B$, $\|\mathcal{Z}\|_{\Psi_2} \le R$ with some $B, R > 0$. Then for all $2 \le k \in \mathbb{N}$,*

$$\mathbb{E}\big[|\mathcal{W}\mathcal{Z}|^k\big] \le 3\ln\big(4\sqrt{\mathbb{K}_0[\mathcal{W}]}\big)\,\mathbb{E}[\mathcal{W}^2]R^2\,k!\,\Big(4\ln\big(4\sqrt{\mathbb{K}_0[\mathcal{W}]}\big)BR\Big)^{k-2}.$$

Proof. Let $c > 0$ to be chosen later. Then by the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\mathbb{E}\big[|\mathcal{W}\mathcal{Z}|^k\big] &= \mathbb{E}\big[|\mathcal{W}\mathcal{Z}|^k\,\mathbb{I}\{|\mathcal{W}| \le cB\}\,\mathbb{I}\{|\mathcal{Z}| \le cR\}\big] \\
&\qquad + \mathbb{E}\big[|\mathcal{W}\mathcal{Z}|^k\,\mathbb{I}\{|\mathcal{W}| \le cB\}\,\mathbb{I}\{|\mathcal{Z}| > cR\}\big] \\
&\qquad + \mathbb{E}\big[|\mathcal{W}\mathcal{Z}|^k\,\mathbb{I}\{|\mathcal{W}| > cB\}\big] \\
&\le \mathbb{E}[\mathcal{W}^2](cR)^2(c^2BR)^{k-2} \\
&\qquad + \mathbb{E}[\mathcal{W}^4]^{\frac{1}{2}}\,(cB)^{k-2}\,\mathbb{E}\big[\mathcal{Z}^{2k}\,\mathbb{I}\{|\mathcal{Z}| > cR\}\big]^{\frac{1}{2}} \\
&\qquad + \mathbb{E}[\mathcal{W}^4]^{\frac{1}{2}}\,\mathbb{E}\big[\mathcal{W}^{4(k-2)}\mathcal{Z}^{4k}\big]^{\frac{1}{4}}\,\mathbb{P}\{|\mathcal{W}| > cB\}^{\frac{1}{4}} \\
&\le \mathbb{E}[\mathcal{W}^2]\Big((cR)^2(c^2BR)^{k-2} \\
&\qquad\qquad + \mathbb{K}_0[\mathcal{W}]^{\frac{1}{2}}(cB)^{k-2}\,\mathbb{E}\big[\mathcal{Z}^{4k}\big]^{\frac{1}{4}}\,\mathbb{P}\{|\mathcal{Z}| > cR\}^{\frac{1}{4}} \\
&\qquad\qquad + \mathbb{K}_0[\mathcal{W}]^{\frac{1}{2}}\,\mathbb{E}\big[\mathcal{W}^{8(k-2)}\big]^{\frac{1}{8}}\,\mathbb{E}\big[\mathcal{Z}^{8k}\big]^{\frac{1}{8}}\,\mathbb{P}\{|\mathcal{W}| > cB\}^{\frac{1}{4}}\Big)\,.
\end{aligned}$$

Further, if $c \geq 2\sqrt{\ln(4)}$, using Lemma A.2(b) gives us

$$\begin{aligned}\mathbb{E}\big[|\mathcal{W}\mathcal{Z}|^2\big] &\leq \mathbb{E}[\mathcal{W}^2]R^2\Big(c^2 + \mathbb{K}_0[\mathcal{W}]^{\frac{1}{2}}\, 2e^{-c^2/4}\big((2(\tfrac{4}{e})^4)^{\frac{1}{4}} + (2(\tfrac{8}{e})^8)^{\frac{1}{8}}\big)\Big) \\ &\leq \mathbb{E}[\mathcal{W}^2](cR)^2\Big(1 + 2!\,\mathbb{K}_0[\mathcal{W}]^{\frac{1}{2}}\, e^{-c^2/4}\Big)\,,\end{aligned}$$

and for $k \geq 3$, with also using $e(k/e)^k \leq k!$, $\sqrt{k!} \leq k!/2$, we get

$$\begin{aligned}\mathbb{E}\big[|\mathcal{W}\mathcal{Z}|^k\big] &\leq \mathbb{E}[\mathcal{W}^2](cR)^2(c^2BR)^{k-2}\cdot \\ &\qquad \Big(1 + \mathbb{K}_0[\mathcal{W}]^{\frac{1}{2}} e^{-c^2/4}\cdot 2^{5/4}\Big(\big(\tfrac{2k}{ec^2}\big)^{\frac{k}{2}} + \big(\tfrac{4(k-2)}{ec^2}\big)^{\frac{k-2}{2}}\big(\tfrac{4k}{ec^2}\big)^{\frac{k}{2}}\Big)\Big) \\ &\leq \mathbb{E}[\mathcal{W}^2](cR)^2(c^2BR)^{k-2}\Big(1 + k!\,\mathbb{K}_0[\mathcal{W}]^{\frac{1}{2}}\, e^{-c^2/4}\Big)\,.\end{aligned}$$

Finally, set $c \doteq 2\sqrt{\ln(4\mathbb{K}_0[\mathcal{W}]^{1/2})} \geq 2\sqrt{\ln(4)}$ by $\mathbb{K}_0[\mathcal{W}] \geq 1$ (due to Jensen's inequality). Then we get the claim for all $k \geq 2$ by using $\mathbb{K}_0[\mathcal{W}]^{1/2}\, e^{-c^2/4} = 1/4$ and $1 \leq k!/2$. ■

Appendix B

The scaled cumulant-generating function

The results of this appendix review a few important properties of the scaled cumulant-generating function.

Lemma B.1. *Let $\mathcal{W}, \mathcal{Z}$ be arbitrary random variables and $t > 0$. Then the following statements hold:*

(a) $\mathbb{E}[\mathcal{W}] \leq \mathbb{C}_{t'}[\mathcal{W}] \leq \mathbb{C}_t[\mathcal{W}]$ *for any* $t \geq t' > 0$,

(b) $\lim_{t \downarrow 0} \mathbb{C}_t[\mathcal{W}] = \mathbb{E}[\mathcal{W}]$,

(c) $\mathbb{C}_t[\lambda W + (1-\lambda)Z] \leq \lambda\, \mathbb{C}_t[W] + (1-\lambda)\, \mathbb{C}_t[Z]$ *for all* $\lambda \in (0,1)$,

(d) $\mathbb{C}_t[\mathcal{W} + \mathcal{Z}] \leq \mathbb{C}_{2t}[\mathcal{W}] + \mathbb{C}_{2t}[\mathcal{Z}]$.

Proof. For (a), let $s \doteq t/t' \geq 1$ so that $t's = t$. Then by Jensen's inequality,

$$\mathbb{E}[\mathcal{W}] = \frac{1}{t'}\mathbb{E}[t'\mathcal{W}] \leq \mathbb{C}_{t'}[\mathcal{W}] \leq \frac{1}{t'} \ln \mathbb{E}\big[\exp(t'\mathcal{W})^s\big]^{1/s} = \mathbb{C}_{t's}[\mathcal{W}] = \mathbb{C}_t[\mathcal{W}]\,.$$

For (b), use L'Hôpital's rule and the dominated convergence theorem to obtain

$$\lim_{t \downarrow 0} \mathbb{C}_t[\mathcal{W}] = \lim_{t \downarrow 0} (1/t) \ln \mathbb{E}\big[e^{t\mathcal{W}}\big] = \lim_{t \downarrow 0} \mathbb{E}\big[\big(e^{t\mathcal{W}} / \mathbb{E}[e^{t\mathcal{W}}]\big)\mathcal{W}\big] = \mathbb{E}[\mathcal{W}]\,.$$

For (c), pick any $\lambda \in (0,1)$, and use Hölder's inequality to get

$$\begin{aligned}\mathbb{C}_t\big[\lambda\mathcal{W} + (1-\lambda)\mathcal{Z}\big] &\leq \frac{1}{t} \ln\Big(\mathbb{E}\big[\big(e^{t\lambda\mathcal{W}}\big)^{1/\lambda}\big]^{\lambda}\, \mathbb{E}\big[\big(e^{t(1-\lambda)\mathcal{Z}}\big)^{1/(1-\lambda)}\big]^{1-\lambda}\Big) \\ &= \lambda\, \mathbb{C}_t[\mathcal{W}] + (1-\lambda)\mathbb{C}_t[\mathcal{Z}]\,.\end{aligned}$$

For (d), use the convexity property (c) with $\lambda = 1/2$ to get $\mathbb{C}_t[\mathcal{W} + \mathcal{Z}] = \mathbb{C}_t[(2\mathcal{W} + 2\mathcal{Z})/2] \leq (\mathbb{C}_t[2\mathcal{W}] + \mathbb{C}_t[2\mathcal{Z}])/2 = \mathbb{C}_{2t}[\mathcal{W}] + \mathbb{C}_{2t}[\mathcal{Z}]$. ■

Lemma B.2. *If $\mathcal{W}, \mathcal{W}_1, \ldots, \mathcal{W}_n$ are i.i.d. random variables and $t > 0$, then $\mathbb{C}_t\big[(1/n)\sum_{i=1}^n \mathcal{W}_i\big] = \mathbb{C}_{t/n}[\mathcal{W}]$.*

Proof. By independence, we have

$$\begin{aligned}\mathbb{C}_t\left[\frac{1}{n}\sum_{i=1}^n \mathcal{W}_i\right] &= \frac{1}{t}\ln\mathbb{E}\left[\exp\left(\frac{t}{n}\sum_{i=1}^n \mathcal{W}_i\right)\right] \\ &= \frac{1}{t}\ln\prod_{i=1}^n \mathbb{E}\big[\exp\big((t/n)\mathcal{W}_i\big)\big] = \mathbb{C}_{t/n}[\mathcal{W}]\,.\end{aligned}$$

■

Lemma B.3. *For any random variables $\mathcal{W}_1, \ldots, \mathcal{W}_n$ and $t > 0$,*

$$\mathbb{C}_t\big[\max_{i=1,\ldots,n} \mathcal{W}_i\big] \le \inf_{s\ge t}\left\{\frac{\ln(n)}{s} + \max_{i=1,\ldots,n} \mathbb{C}_s[\mathcal{W}_i]\right\}.$$

Proof. Using Lemma B.1a for an arbitrary $s \ge t$, we get

$$\begin{aligned}\mathbb{C}_t\Big[\max_{i=1,\ldots,n} \mathcal{W}_i\Big] &\le \mathbb{C}_s\Big[\max_{i=1,\ldots,n} \mathcal{W}_i\Big] = \frac{1}{s}\ln\mathbb{E}\Big[\max_{i=1,\ldots,n} e^{s\mathcal{W}_i}\Big] \\ &\le \frac{1}{s}\ln\sum_{i=1}^n \mathbb{E}\big[e^{s\mathcal{W}_i}\big] = \frac{\ln(n)}{s} + \max_{i=1,\ldots,n} \mathbb{C}_s[\mathcal{W}_i]\,.\end{aligned}$$

Taking infimum over $s \ge t$, we get the first claim. ■

Lemma B.4. *If $\mathcal{W}$ is a σ-sub-Gaussian random variable and $t > 0$, then $\mathbb{C}_t[\mathcal{W}] \le \mathbb{E}[\mathcal{W}] + t\sigma^2/2$ and $\mathbb{C}_t[|\mathcal{W}|] \le |\mathbb{E}\mathcal{W}| + \max\big\{\sigma\sqrt{2\ln 2}, t\sigma^2\big\}$. Furthermore, if $\mathcal{Z}$ is a random variable with $\mathbb{E}\big[e^{|\mathcal{Z}|/R}\big] \le 2$, then for all $t \le 1/R$, $\mathbb{C}_t[|\mathcal{Z}|] \le R\,\ln 2$.*

Proof. For the first claim, simply use the σ-sub-Gaussian property,

$$\mathbb{C}_t[\mathcal{W}] = (1/t)\ln\mathbb{E}\big[e^{t\mathcal{W}}\big] \le (1/t)\big(t\mathbb{E}[\mathcal{W}] + t^2\sigma^2/2\big) = \mathbb{E}[\mathcal{W}] + t\sigma^2/2\,.$$

For the second claim, use the monotonicity of $s \mapsto \mathbb{C}_s[\mathcal{W}]$ with any $s \ge t$, and $e^{|x|} \le e^x + e^{-x}$, to get

$$\begin{aligned}\mathbb{C}_t[|\mathcal{W}|] &\le (1/s)\ln\big(\mathbb{E}\big[e^{s\mathcal{W}}\big] + \mathbb{E}\big[e^{-s\mathcal{W}}\big]\big) \\ &\le (1/s)\ln\Big(2\,e^{s|\mathbb{E}\mathcal{W}| + s^2\sigma^2/2}\Big) = |\mathbb{E}\mathcal{W}| + s\sigma^2/2 + \ln(2)/s\,.\end{aligned}$$

If $t\sigma^2/2 \le \ln(2)/t$, then we can choose $t \le s = \sqrt{2\ln 2}/\sigma$ and get a $\sigma\sqrt{2\ln 2}$ term on the right hand side. Otherwise, use $s = t$ and $t\sigma^2/2 + \ln(2)/t \le t\sigma^2$.

For the last claim, use the monotonicity of $s \mapsto \mathbb{C}_s[\mathcal{Z}]$ with $t \le 1/R$ to get $\mathbb{C}_t[|\mathcal{Z}|] \le R\ln\mathbb{E}\big[e^{|\mathcal{Z}|/R}\big] \le R\ln 2$. ■

Appendix C

Auxiliary results on covering numbers

In this section we review a few useful results on the covering numbers of finite dimensional bounded spaces.

For the derivations here, we need packing numbers. Let $(\mathcal{P}, \psi)$ be a metric space and $\epsilon > 0$. Then the set $\{p_1, \ldots, p_k\} \in \mathcal{P}$ is an *ϵ-packing* of $\mathcal{P}$ under ψ if any two distinct elements in $\{p_1, \ldots, p_k\}$ are farther away from each other than ϵ: for any $i, j = 1, \ldots, k$, $i \neq j$, $\psi(p_i, p_j) > \epsilon$. The *$\epsilon$-packing number* of $\mathcal{P}$ under ψ, $\mathcal{M}_\psi(\epsilon, \mathcal{P})$, is the cardinality of the largest ϵ-packing:

$$\mathcal{M}_\psi(\epsilon, \mathcal{P}) \doteq \sup \left\{ k \in \mathbb{N} \,\middle|\, \exists p_1, \ldots, p_k \in \mathcal{P} : \min_{\substack{i,j=1,\ldots,k \\ i \neq j}} \psi(p_i, p_j) > \epsilon \right\}. \tag{C.1}$$

The relation between covering and packing numbers are given for all $\epsilon > 0$ by (for example, Dudley, 1999, Theorem 1.2.1)

$$\mathcal{N}_\psi(\epsilon, \mathcal{P}) \leq \mathcal{M}_\psi(\epsilon, \mathcal{P}) \leq \mathcal{N}_\psi(\epsilon/2, \mathcal{P}). \tag{C.2}$$

Next, we review a useful tool for bounding the entropy of finite dimensional vector spaces. Its proof is based on the volume argument (for example, Pollard, 1990, Lemma 4.1), which we provide here for completeness.

Lemma C.1. *Let $t \in \mathbb{N} \cup \{\infty\}$ and $\mathcal{P} \subset \mathbb{R}^v$ with a finite radius under $\|\cdot\|_t$, that is suppose there exists $\boldsymbol{q}_* \in \mathbb{R}^v$ such that $\mathcal{P} \subseteq \mathcal{B}_t(\boldsymbol{q}_*, R)$ for some $R > 0$. Then $\mathcal{H}_{\|\cdot\|_t}(\epsilon, \mathcal{P}) \leq v \ln(3R/\epsilon)$ for all $\epsilon \in (0, 3R]$. Furthermore, if $\mathcal{P}$ contains a ball with radius $r > 0$, that is $\mathcal{B}_t(\boldsymbol{p}_*, r) \subseteq \mathcal{P}$ holds for some $\boldsymbol{p}_* \in \mathcal{P}$, then $\mathcal{H}_{\|\cdot\|_t}(\epsilon, \mathcal{P}) \geq v \ln(r/\epsilon)$ for all $\epsilon > 0$.*

Proof. Let $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_M\}$ be an ϵ-packing of $\mathcal{P}$ under $\|\cdot\|_t$ with maximum cardinality. Consider the balls around the packing elements $\boldsymbol{q}_i$ of radius $\epsilon/2$, and notice that by the packing property ($\|\boldsymbol{q}_i - \boldsymbol{q}_j\|_t > \epsilon$ for all $i \neq j$), these balls are disjunct, that is $\mathcal{B}_t(\boldsymbol{q}_i, \epsilon/2) \cap \mathcal{B}_t(\boldsymbol{q}_j, \epsilon/2) = \emptyset$ if $i \neq j$. Also notice that each ball $\mathcal{B}_t(\boldsymbol{q}_i, \epsilon/2)$ lies within $\mathcal{B}_t(\boldsymbol{q}_*, R + \epsilon/2)$. Putting these together, we get

$$\begin{aligned} M(\epsilon/2)^v \operatorname{vol}(\mathcal{B}_0) &= \operatorname{vol}\left(\cup_{i=1}^M \mathcal{B}_t(\boldsymbol{q}_i, \epsilon/2)\right) \\ &\leq \operatorname{vol}\big(\mathcal{B}_t(\boldsymbol{q}_*, R + \epsilon/2)\big) = (R + \epsilon/2)^v \operatorname{vol}(\mathcal{B}_0)\,, \end{aligned}$$

where $\mathcal{B}_0$ is the unit ball in $\mathbb{R}^v$ under $\|\cdot\|_t$, $\operatorname{vol}(\mathcal{B})$ is the (v-dimensional) volume of set $\mathcal{B}$, and we used that $\operatorname{vol}(\mathcal{B}_t(\boldsymbol{q}, z)) = z^v \operatorname{vol}(\mathcal{B}_0)$ for all $\boldsymbol{q} \in \mathbb{R}^d$. Dividing both sides by $(\epsilon/2)^v$, we get $M \leq (1 + 2R/\epsilon)^v \leq (3R/\epsilon)^v$ for all $\epsilon \in (0, R]$. As $\mathcal{H}_{\|\cdot\|_t}(\epsilon, \mathcal{P}) = 0$ for $\epsilon > R$, the claimed upper bound follows from (C.2).

Now let $\{\boldsymbol{p}_1, \dots, \boldsymbol{p}_N\}$ be an ϵ-cover of $\mathcal{P}$ with minimum cardinality. Then by using $\mathcal{B}(\boldsymbol{p}_*, r) \subseteq \mathcal{P} \subseteq \cup_{i=1}^N \mathcal{B}(\boldsymbol{p}_i, \epsilon)$, we get

$$r^v \operatorname{vol}(\mathcal{B}_0) = \operatorname{vol}\big(\mathcal{B}(\boldsymbol{p}_*, r)\big) \leq \operatorname{vol}\left(\cup_{i=1}^N \mathcal{B}(\boldsymbol{p}_i, \epsilon)\right) \leq N\epsilon^v \operatorname{vol}(\mathcal{B}_0)\,.$$

Dividing both sides by ϵ^v, we get the claimed lower bound. ■

Appendix D

Density estimation and minimax lower bounds

In this section, we prove our lower bound on the minimax risk, Theorem 3.1, and provide the necessary background for this on information theory and density estimation.

For the derivations here, we need packing numbers (C.1) and a few information theoretic developments. We denote (σ-finite) reference measures for probability densities by ν, which can change its meaning based on the context. We use $P_{\mathcal{X}}$, $P_{\mathcal{Y}}$, $P_{\mathcal{X},\mathcal{Y}}$, $P_{\mathcal{X}|\mathcal{Y}}$ to denote the densities of the random variables $\mathcal{X}$, $\mathcal{Y}$, $(\mathcal{X},\mathcal{Y})$ and $\mathcal{X}|\mathcal{Y}$, appropriately.

The *entropy* of a random variable $\mathcal{X}$ is $H(\mathcal{X}) \doteq -\int \ln(P_{\mathcal{X}}(x))P_{\mathcal{X}}(x)\,\nu(dx)$. Similarly, let $H(\mathcal{X},\mathcal{Y}) \doteq -\int \ln(P_{\mathcal{X},\mathcal{Y}}(x,y)P_{\mathcal{X},\mathcal{Y}}(x,y)\,\nu(dx,dy)$ be the *joint entropy of* $\mathcal{X}$ *and* $\mathcal{Y}$, and denote the *conditional entropy of* $\mathcal{X}$ *given* $\mathcal{Y}$ by $H(\mathcal{X}|\mathcal{Y}) \doteq -\int \ln(P_{\mathcal{X}|\mathcal{Y}}(x,y)P_{\mathcal{X}|\mathcal{Y}}(x,y)\,\nu(dx,dy)$. Furthermore, the *mutual information of* $\mathcal{X}$ *and* $\mathcal{Y}$ is defined as $I(\mathcal{X};\mathcal{Y}) \doteq D_{\mathrm{KL}}(P_{\mathcal{X},\mathcal{Y}}\|P_{\mathcal{X}}\cdot P_{\mathcal{Y}})$, where $D_{\mathrm{KL}}(P\|Q) \doteq \int \ln(P(z)/Q(z))P(z)\,\nu(dz)$ is the *Kullback-Leibler (KL) divergence* between two densities P and Q. Then, consider the following result:

Lemma D.1 (Fano's inequality)**.** *Let $\mathcal{X}$ be a discrete random variable on the finite set $\mathbb{X}$ and $\mathcal{Y}$ be an arbitrary random variable. Furthermore, let $\hat{\mathcal{X}}$ be a discrete random variable such that $\hat{\mathcal{X}}|\mathcal{Y}$ is independent of $\mathcal{X}$. Then*

$$\mathbb{P}\{\mathcal{X} \neq \hat{\mathcal{X}}\} \geq \frac{H(\mathcal{X}|\mathcal{Y}) - \ln 2}{\ln|\mathbb{X}|}\,.$$

Proof. See Theorem 2.10.1 of Cover and Thomas (2006) for discrete $\mathcal{Y}$, and ex-

tend it to the continuous setting by noting that the data processing inequality (Cover and Thomas, 2006, Theorem 2.8.1) carries through to the continuous case along with the properties of the mutual information (Cover and Thomas, 2006, Section 8.5). ■

When $\mathcal{X}$ is uniformly distributed on $\mathcal{X}$, that is $P_{\mathcal{X}}(x) = 1/|\mathbb{X}|$ for all $x \in \mathbb{X}$, we have $H(\mathcal{X}) = \ln|\mathbb{X}|$, and so we can rewrite the result of Lemma D.1 as

$$\mathbb{P}\{\mathcal{X} \neq \hat{\mathcal{X}}\} \geq 1 - \frac{I(\mathcal{X};\mathcal{Y}) + \ln 2}{\ln|\mathbb{X}|}, \tag{D.1}$$

where we used that $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$ as explained by Cover and Thomas (2006, Section 2.4). Now let $\bar{\mathcal{X}}$ be a uniform random variable on an arbitrary finite set $\bar{\mathbb{X}}(x)$, which might depend on some $x \in \mathbb{X}$, and define $P_{\mathcal{Y}|\mathcal{X}=x}(y) \doteq P_{\mathcal{Y}|\mathcal{X}}(x,y)$. Then, we can upper bound $I(\mathcal{X};\mathcal{Y})$ as

$$\begin{aligned}
I(\mathcal{X};\mathcal{Y}) &= D_{\mathrm{KL}}(P_{\mathcal{X},\mathcal{Y}} \| P_{\mathcal{X}} \cdot P_{\mathcal{Y}}) \\
&= \sum_{x\in\mathbb{X}} P_{\mathcal{X}}(x) \int P_{\mathcal{Y}|\mathcal{X}}(x,y) \ln\left(\frac{P_{\mathcal{Y}|\mathcal{X}}(x,y)}{P_{\mathcal{Y}}(y)}\right) \nu(dy) \\
&\leq \max_{x\in\mathbb{X}} D_{\mathrm{KL}}(P_{\mathcal{Y}|\mathcal{X}=x} \| P_{\mathcal{Y}}) \\
&= \max_{x\in\mathbb{X}} \int P_{\mathcal{Y}|\mathcal{X}=x}(y) \ln\left(\frac{P_{\mathcal{Y}|\mathcal{X}=x}(y)}{(1/|\bar{\mathbb{X}}(x)|)\sum_{x'\in\bar{\mathbb{X}}(x)} P_{\mathcal{Y}|\bar{\mathcal{X}}}(x',y)}\right) \nu(dy) \\
&\leq \max_{x\in\mathbb{X}} \min_{\bar{x}\in\bar{\mathbb{X}}(x)} \int P_{\mathcal{Y}|\mathcal{X}=x}(y) \ln\left(\frac{P_{\mathcal{Y}|\mathcal{X}=x}(y)}{(1/|\bar{\mathbb{X}}(x)|)P_{\mathcal{Y}|\bar{\mathcal{X}}}(\bar{x},y)}\right) \nu(dy) \\
&= \max_{x\in\mathbb{X}} \left\{ \ln|\bar{\mathbb{X}}(x)| + \min_{\bar{x}\in\bar{\mathbb{X}}(x)} D_{\mathrm{KL}}(P_{\mathcal{Y}|\mathcal{X}=x} \| P_{\mathcal{Y}|\bar{\mathcal{X}}=\bar{x}}) \right\}.
\end{aligned} \tag{D.2}$$

Now let $d_{\mathrm{KL}} \doteq \sqrt{D_{\mathrm{KL}}}$ be the *square root KL divergence* and consider the next result (Lemma D.2), which is a modified version of Theorem 1 in Yang and Barron (1999), using local entropies in a slightly different way than presented in Yang and Barron (1999, Section 7). This result (Lemma D.2) extends the previous ideas to a probabilistic lower bound on the minimax rate of density estimators.

Lemma D.2. *Let $\mathbb{M}$ be a class of probability densities on some set $\mathbb{W}$ and $\mathcal{D}_n = (\mathcal{W}_1, \dots, \mathcal{W}_n) \sim P^n$ be an i.i.d. sample of size $n \in \mathbb{N}$ from $P \in \mathbb{M}$. Suppose there exist $\epsilon, \epsilon_* > 0$, and two functions $h, h_* : \mathbb{N} \to \mathbb{R}_{>0}$ such that $\mathcal{H}_{d_{KL}}(\epsilon_*, \mathbb{M}) \geq h_*(n)$ and $\mathcal{H}^*_{d_{KL}}(\epsilon, \epsilon_*, \mathbb{M}) \leq h(n)$. Furthermore, if we also have $4h(n) + \ln 4 \leq h_*(n)$ and $n\epsilon^2 \leq h(n)$, then*

$$\inf_{Q_n} \sup_{P \in \mathbb{M}} \mathbb{P}\big\{ D_{KL}\big(P \| Q_n(\mathcal{D}_n)\big) \geq \epsilon_*^2 \big\} \geq 1/2 \,,$$

where the probability is taken with respect to the random sample $\mathcal{D}_n \sim P^n$, and the infimum over Q_n scans through all estimators mapping to any probability density on $\mathbb{W}$.

Proof. Let $\mathbb{M}_{\epsilon_*}$ be an ϵ_*-packing of $\mathbb{M}$ under d_{KL} with maximum cardinality, and for any density Q on $\mathbb{W}$, define $Q^* \in \operatorname{argmin}_{P' \in \mathbb{M}_{\epsilon_*}} D_{\mathrm{KL}}(P' \| Q)$ with an arbitrary tie-breaking. Then notice that the definition of an ϵ_*-packing implies for any $Q, Q^* \in \mathbb{M}_{\epsilon_*}$ that $D_{\mathrm{KL}}(Q \| Q^*) \geq \epsilon_*^2$ if and only if $Q \neq Q^*$. Hence, by using the $P_n \doteq Q_n(\mathcal{D}_n)$ shorthand notation, we have

$$\begin{aligned}
\inf_{Q_n} \sup_{P \in \mathbb{M}} \mathbb{P}\big\{ D_{\mathrm{KL}}\big(P \| P_n\big) \geq \epsilon_*^2 \big\} &\geq \inf_{Q_n} \max_{P \in \mathbb{M}_{\epsilon_*}} \mathbb{P}\big\{ D_{\mathrm{KL}}\big(P \| P_n\big) \geq \epsilon_*^2 \big\} \\
&\geq \inf_{Q_n} \max_{P \in \mathbb{M}_{\epsilon_*}} \mathbb{P}\big\{ D_{\mathrm{KL}}\big(P \| P_n^*\big) \geq \epsilon_*^2 \big\} \\
&= \inf_{Q_n} \max_{P \in \mathbb{M}_{\epsilon_*}} \mathbb{P}\big\{ P \neq P_n^* \big\} \\
&\geq \inf_{Q_n} \frac{1}{|\mathbb{M}_{\epsilon_*}|} \sum_{P \in \mathbb{M}_{\epsilon_*}} \mathbb{P}\big\{ P \neq P_n^* \big\} \\
&= \inf_{Q_n} \mathbb{P}\big\{ P_* \neq P_n^* \big\} \,,
\end{aligned} \tag{D.3}$$

where in the last line P_* is a uniform random variable on $\mathbb{M}_{\epsilon_*}$ and the probability is taken with respect to $P \sim P_*$ and $\mathcal{D}_n \sim P^n$.

Then notice that P_* and $P_n^* | \mathcal{D}_n$ are independent, so the requirements of Fano's inequality (Lemma D.1) hold (with $\mathcal{X} \leftarrow P_*$, $\hat{\mathcal{X}} \leftarrow P_n^*$, $\mathcal{Y} \leftarrow \mathcal{D}_n$, $\mathbb{X} \leftarrow \mathbb{M}_{\epsilon_*}$). Hence, using that P_* is uniform, we get by (D.1) that

$$\mathbb{P}\big\{ P_* \neq P_n^* \big\} \geq 1 - \frac{I(P_*; \mathcal{D}_n) + \ln 2}{\ln |\mathbb{M}_{\epsilon_*}|} \geq 1 - \frac{I(P_*; \mathcal{D}_n) + \ln 2}{h_*(n)} \,, \tag{D.4}$$

where we used $\ln |\mathbb{M}_{\epsilon_*}| = \ln \mathcal{M}_{d_{\mathrm{KL}}}(\epsilon_*, \mathbb{M}) \geq \mathcal{H}_{d_{\mathrm{KL}}}(\epsilon_*, \mathbb{M}) \geq h_*(n)$ by (C.2).

Now, let $\mathbb{M}^*_{\epsilon,P}$ be an ϵ-cover of $\{Q \in \mathbb{M} : d_{\mathrm{KL}}(P\|Q) \le \epsilon_*\}$ under d_{KL} with minimum cardinality, so $\ln|\mathbb{M}^*_{\epsilon,P}| = \mathcal{H}^*_{d_{\mathrm{KL}}}(\epsilon, \epsilon_*, \mathbb{M})$. Then, applying (D.2) (with $\bar{\mathbb{X}} \leftarrow \mathbb{M}^*_{\epsilon,P}$) to (D.4), we obtain

$$\begin{aligned}\mathbb{P}\big\{P_* \neq P^*_n\big\} &\ge 1 - \frac{\max_{P\in\mathbb{M}_{\epsilon_*}} \ln|\mathbb{M}^*_{\epsilon,P}| + \min_{\hat{P}\in\mathbb{M}^*_{\epsilon,P}} D_{\mathrm{KL}}(P_{\mathcal{D}_n|P}\|P_{\mathcal{D}_n|\hat{P}}) + \ln 2}{h_*(n)} \\ &\ge 1 - \frac{h(n) + n\epsilon^2 + \ln 2}{h_*(n)} \ge 1/2\,, \end{aligned} \tag{D.5}$$

where $\min_{\hat{P}\in\mathbb{M}^*_{\epsilon,P}} D_{\mathrm{KL}}(P_{\mathcal{D}_n|P}\|P_{\mathcal{D}_n|\hat{P}}) = n \min_{\hat{P}\in\mathbb{M}^*_{\epsilon,P}} D_{\mathrm{KL}}(P_{\mathcal{W}_1|P}\|P_{\mathcal{W}_1|\hat{P}}) \le n\epsilon^2$ followed from the i.i.d. property of the sample $\mathcal{D}_n$ and the definition of an ϵ-cover with $P \in \mathbb{M}^*_{\epsilon,P}$ for all $P \in \mathbb{M}$.

Finally, combining (D.3) with (D.5), we get the claim. ■

The following result (Lemma D.3) relates the conditions of the previous density estimation lower bound (Lemma D.2) to the usual linear (a) and non-linear settings (b).

Lemma D.3. *Let $\mathbb{M}$ be a set of probability densities, $v > 0$ and $c_2 \ge c_1 > 0$. Then consider the following cases:*

(a) If for some $c_0 > 0$, $v\ln(c_1/z) \le \mathcal{H}_{d_{KL}}(z, \mathbb{M})$, $\mathcal{H}^*_{d_{KL}}(z, s, \mathbb{M}) \le v\ln(c_2\, s/z)$ for all $s \in (0, c_0]$ and $z \in (0, c_2\, s]$, then the conditions of Lemma D.2 are satisfied with $\epsilon = (13/20)\sqrt{v/n}$ and $\epsilon_* = (1/c_2)\sqrt{v/n}$ for every $n \ge (v/c_2^2)\max\{32\cdot 2^{4/v}/c_1^2, 1/c_0^2\}$.

(b) If $c_1 z^{-v} \le \mathcal{H}_{d_{KL}}(z, \mathbb{M}) \le c_2 z^{-v}$ is satisfied for all $z \in (0, \epsilon_0]$, then the conditions of Lemma D.2 hold for all $n \in \mathbb{N}$ with $\epsilon = \epsilon_0\, n^{-1/(v+2)}$ and $\epsilon_* = \big(6\max\{1, c_2\,\epsilon_0^{-v}, \epsilon_0^2\}/c_1\big)^{-1/v} n^{-1/(v+2)}$.

Proof. To prove (a), notice that $\epsilon = (13/20)\sqrt{v/n} < c_2\,\epsilon_*$ and $\epsilon_* \le c_0$ is satisfied if $n \ge v/(c_0\, c_2)^2$. Now choose

$$\begin{aligned} h_*(n) &\doteq v\ln\big(c_1 c_2\sqrt{n/v}\big) = v\ln\big(c_1/\epsilon_*\big) \le \mathcal{H}_{d_{\mathrm{KL}}}(\epsilon_*, \mathbb{M})\,, \\ h(n) &\doteq v\ln\big(20/13\big) = v\ln\big(c_2\,\epsilon_*/\epsilon\big) \ge \mathcal{H}^*_{d_{\mathrm{KL}}}(\epsilon, \epsilon_*, \mathbb{M})\,. \end{aligned}$$

Then, $n\epsilon^2 = v(13/20)^2 < h(n)$ and $4h(n) + \ln 4 = v\ln\big(4^{1/v}(20/13)^4\big) \le h_*(n)$ if $n \ge 4^{2/v}(20/13)^8\, v/(c_1 c_2)^2$.

To show (b), set $\epsilon \doteq \epsilon_0\, n^{-1/(v+2)} \le \epsilon_0$ The conditions for $h(n)$ hold if

$$h(n) \ge c_2\, \epsilon_0^{-v} n^{v/(v+2)} = c_2\, \epsilon^{-v} \ge \mathcal{H}_{d_{\mathrm{KL}}}(\epsilon, \mathbb{M})\,,$$
$$h(n) \ge \epsilon_0^2\, n^{v/(v+2)} = n\epsilon^2\,.$$

Hence, we can set $h(n) \doteq b\, n^{v/(v+2)}$ with $b \doteq \max\{1, c_2\, \epsilon_0^{-v}, \epsilon_0^2\}$ and also get $h(n) \ge 1$. Then, $4h(n) + \ln 4 < 6h(n) \doteq h_*(n) = c_1\, \epsilon_*^{-v} \le \mathcal{H}_{d_{\mathrm{KL}}}(\epsilon_*, \mathbb{M})$ holds too with $\epsilon_* = (c_1/(6b))^{1/v}\, n^{-1/(v+2)} \le \epsilon_0$ as $(c_1/(6b))^{1/v} \le \epsilon_0(c_1/(6c_2)) < \epsilon_0$. ■

Then, we simply reduce the derivation of Theorem 3.1 to the previously discussed results.

Proof of Theorem 3.1. First observe that the regression function is always in $\mathcal{F}_*$, hence $f_{\mu,\mathcal{F}} = f_{\mu,\mathcal{F}_*}$ implying $\mathcal{R}_n(\mathbb{M}^\sigma_{\mathrm{gs}}, \ell_{\mathrm{sq}}, \mathcal{F}) = \mathcal{R}_n(\mathbb{M}^\sigma_{\mathrm{gs}}, \ell_{\mathrm{sq}}, \mathcal{F}_*)$.

Now let $P_f, P_g \in \mathbb{M}^\sigma_{\mathrm{gs}}(\mathcal{F}_*, P_{\mathbb{X}})$ be two densities corresponding to $f, g \in \mathcal{F}_*$, respectively. Then notice that the KL divergence between P_f and P_g, due to the gaussian noise, satisfies

$$D_{\mathrm{KL}}(P_f \| P_g) = \mathbb{E}\big[D_{\mathrm{KL}}(P_{f|\boldsymbol{\mathcal{X}}} \| P_{g|\boldsymbol{\mathcal{X}}})\big] = \frac{1}{2\sigma^2}\mathbb{E}\big[|f(\boldsymbol{\mathcal{X}}) - g(\boldsymbol{\mathcal{X}})|^2\big] = \frac{\|f-g\|^2_{P_{\mathbb{X}}}}{2\sigma^2}\,,$$

where $P_{f|\boldsymbol{\mathcal{X}}}, P_{g|\boldsymbol{\mathcal{X}}}$ are the conditional distributions given $\boldsymbol{\mathcal{X}}$, respectively. Hence, we have $\mathcal{H}_{P_{\mathbb{X}}}(\sqrt{2}\sigma\epsilon, \mathcal{F}_*) = \mathcal{H}_{d_{\mathrm{KL}}}(\epsilon, \mathbb{M})$ for all $\epsilon > 0$.

Then we use Lemma D.2 with Lemma D.3, substituting $c_0 \leftarrow c_0/(\sqrt{2}\sigma)$, $c_1 \leftarrow c_1/(\sqrt{2}\sigma)$, $c_2 \leftarrow c_2/(\sqrt{2}\sigma)$ for Lemma D.3 (a) to obtain Theorem 3.1 (a), and use $\epsilon_0 \leftarrow \epsilon_0/(\sqrt{2}\sigma)$, $c_1 \leftarrow c_1(\sqrt{2}\sigma)^{-v}$, $c_2 \leftarrow c_2(\sqrt{2}\sigma)^{-v}$ for Lemma D.3 (b) to get Theorem 3.1 (b). ■

Appendix E

Optimization tools

In this appendix we shortly review a few optimization results.

Let $\mathbb{X}, \mathbb{Y}$ be two vector spaces over $\mathbb{R}$, and the graph of a set–valued function $C : \mathbb{X} \to 2^{\mathbb{Y}}$ be defined as $\text{graph}(C) \doteq \big\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X} \times \mathbb{Y} : \boldsymbol{y} \in C(\boldsymbol{x})\big\}$. The next lemma summarizes a few properties of $\text{graph}(C)$.

Lemma E.1. *Let $\mathbb{X}, \mathbb{Y}$ be two convex sets and $C : \mathbb{X} \to 2^{\mathbb{Y}}$ be some set–valued function. If graph(C) is convex, then $C(\boldsymbol{x})$ is convex for all $\boldsymbol{x} \in \mathbb{X}$, but the converse is not true in general.*

Furthermore, if $C(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{Y} : g_j(\boldsymbol{x}, \boldsymbol{y}) \leq 0,\ j = 1, \ldots, m\}$ for all $\boldsymbol{x} \in \mathbb{X}$ with some $g_j : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ jointly–convex functions in their arguments, then graph(C) is convex.

Proof. Let $\boldsymbol{x} \in \mathbb{X}$, $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{Y}$ and $\lambda \in (0, 1)$. Then by the convexity of $\text{graph}(C)$, we have $\lambda(\boldsymbol{x}, \boldsymbol{y}_1) + (1-\lambda)(\boldsymbol{x}, \boldsymbol{y}_2) = (\boldsymbol{x}, \lambda\boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2) \in \text{graph}(C)$, implying that $\lambda\boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2 \in C(\boldsymbol{x})$, so proving the first claim.

To show that the converse is not true, consider $\mathbb{X} = \mathbb{Y} = [0, 1]$ and $C(x) = \mathbb{I}\{x = 1\}$ with $(x_1, y_1) = (0, 0)$ and $(x_2, y_2) = (1, 1)$. Then $C(x)$ is convex for all $x \in [0, 1]$, but $\lambda(x_1, y_1) + (1-\lambda)(x_2, y_2) = (1-\lambda, 1-\lambda) \notin \text{graph}(C)$, because $1 - \lambda \notin C(1-\lambda) = \{0\}$.

To prove the last claim, let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}$, $\boldsymbol{y}_1 \in C(\boldsymbol{x}_1)$, $\boldsymbol{y}_2 \in C(\boldsymbol{x}_2)$. Then by the joint–convexity of the g_j functions, $g_j(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \lambda\boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2) \leq \lambda g_j(\boldsymbol{x}_1, \boldsymbol{y}_1) + (1-\lambda) g_j(\boldsymbol{x}_2, \boldsymbol{y}_2) \leq 0$, which implies that $\lambda\boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2 \in C(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)$, and so proves the claim. ■

Next, consider the following result, which is a slight generalization of Theorem 5.3 in Rockafellar (1972).

Lemma E.2. *Let $\mathbb{X}, \mathbb{Y}$ be two convex sets and $f : \mathbb{X} \times \mathbb{Y}$ be a jointly–convex function in its arguments. Additionally, let $C : \mathbb{X} \to 2^{\mathbb{Y}}$ be a set–valued function for which graph(C) is convex. Then $g(\boldsymbol{x}) \doteq \inf_{\boldsymbol{y} \in C(\boldsymbol{x})} f(\boldsymbol{x}, \boldsymbol{y})$ is a convex function.*

Proof. Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}$, $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{Y}$ and $\lambda \in (0, 1)$. As graph(C) is convex,

$$\boldsymbol{y}_1 \in C(\boldsymbol{x}_1), \boldsymbol{y}_2 \in C(\boldsymbol{x}_2) \quad \Rightarrow \quad \lambda \boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2 \in C\big(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2\big)\,.$$

Using this with the fact that the infimum on a subset becomes larger, and the joint–convexity of f, we get

$$\begin{aligned} g\big(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2\big) &= \inf_{\boldsymbol{z} \in C(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)} f\big(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \boldsymbol{z}\big) \\ &\leq \inf_{\boldsymbol{y}_1 \in C(\boldsymbol{x}_1)} \inf_{\boldsymbol{y}_2 \in C(\boldsymbol{x}_2)} f\big(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \lambda \boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2\big) \\ &\leq \inf_{\boldsymbol{y}_1 \in C(\boldsymbol{x}_1)} \inf_{\boldsymbol{y}_2 \in C(\boldsymbol{x}_2)} \lambda f(\boldsymbol{x}_1, \boldsymbol{y}_1) + (1-\lambda) f(\boldsymbol{x}_2, \boldsymbol{y}_2) \\ &= \lambda g(\boldsymbol{x}_1) + (1-\lambda) g(\boldsymbol{x}_2)\,, \end{aligned}$$

which proves the convexity of g. ■

Appendix F

Miscellaneous

This appendix is a collection of a few technical results which did not fit into any other context.

Lemma F.1. *For all $n \in \mathbb{N}$ and $c \geq 1$, $\frac{c-1}{cn} \leq 1 - c^{-1/n} \leq \frac{\ln(c)}{n}$.*

Proof. We prove the lower bound by induction on n. Notice that the claim holds for $n = 1$. Then let $n > 1$ and suppose that the claim holds for $n - 1$. Observe that $(x^q - qx)' = q(x^{q-1} - 1) \geq 0$ for $x \in (0, 1]$, $q \in [0, 1]$, hence by the monotonicity of $x \mapsto x^q - qx$, we have for any $0 < b \leq a \leq 1$ and $q \in [0, 1]$ that $a^q - b^q \geq q(a - b)$. Using this with $c \geq 1$ and the induction hypothesis, we obtain

$$1 - c^{-1/n} = 1 - \left(c^{-1/(n-1)}\right)^{(n-1)/n} \geq \left(1 - c^{-1/(n-1)}\right)\frac{n-1}{n} \geq \frac{c-1}{cn}\,,$$

which proves the induction step and so the lower bound.

For the upper bound, by $e^x \geq 1 + x$ for all $x \in \mathbb{R}$, we get

$$1 - c^{-1/n} = 1 - e^{-\ln(c)/n} \leq \frac{\ln(c)}{n}\,.$$

■

Lemma F.2. *For $A \in \mathbb{R}^{n\times d}$, $\boldsymbol{b} \in \mathbb{R}^n$ and $r > 0$, $\left\|(rI_d + A^\top A)^{-1}A^\top \boldsymbol{b}\right\| \leq \frac{\|\boldsymbol{b}\|}{2\sqrt{r}}$.*

Proof. Consider a thin singular value decomposition of A given as $A = USV^\top$, where $U \in \mathbb{R}^{n\times d}$ is semi-orthogonal ($U^\top U = I_d$), $S \in \mathbb{R}^{d\times d}$ is diagonal with nonegative elements, and $V \in \mathbb{R}^{d\times d}$ is orthogonal ($V^\top V = VV^\top = I_d$). Then

$$\left\|(rI_d + A^\top A)^{-1}A^\top \boldsymbol{b}\right\| = \left\|V(rI_d + S^2)^{-1}SU^\top \boldsymbol{b}\right\| \leq \left\|(rI_d + S^2)^{-1}S\right\| \|\boldsymbol{b}\|\,,$$

where we used $\|U\| = \|V\| = 1$ (as $U^\top U = I_d$ implies $\|U\boldsymbol{x}\|^2 = \|\boldsymbol{x}\|^2$). Finally notice that

$$\big\|(rI_d + S^2)^{-1}S\big\| \le \max_{s\ge 0} \frac{s}{r+s^2} = \frac{1}{2\sqrt{r}}\,,$$

which proves the claim. ■

Lemma F.3. *Consider a regression problem (as in Section 3.1) with some distribution μ and the squared loss ($\ell = \ell_{sq}$) such that f_* is a regression function and $\mathbb{E}\big[e^{|\mathcal{Y}-f_*(\boldsymbol{\mathcal{X}})|^2/\sigma^2}\big|\boldsymbol{\mathcal{X}}\big] \le 2$ holds for some $\sigma > 0$. Take a class $\hat{\mathcal{F}}_n \subseteq \{\mathbb{X} \to \mathbb{R}\}$ which is independent of $\mathcal{Y}_1, \dots, \mathcal{Y}_n$, and a random variable $\mathcal{Z}$ such that $\mathbb{E}[e^{\mathcal{Z}/R}] \le c$ for some $c > 0$, and $\inf_{f\in\hat{\mathcal{F}}_n} \frac{2}{n}\sum_{i=1}^n \mathcal{W}_{f,i}^2 + T \le \mathcal{Z}$ a.s. holds with $\mathcal{W}_{f,i} \doteq f(\boldsymbol{\mathcal{X}}_i) - f_*(\boldsymbol{\mathcal{X}}_i)$ and some random variable T. Then for $B_* \doteq \max\{R, 2\sigma^2/n\}\ln(4c/\gamma)$, we have*

$$\mathbb{P}\Big\{\inf_{f\in\hat{\mathcal{F}}_n} L_n(f, f_*) + T > B_*\Big\} \le \frac{\gamma}{4}\,.$$

Proof. Let $R_n \doteq \max\{R, 2\sigma^2/n\}$ and write $B_* = R_n \ln(4c/\gamma)$. Then, by using $\ell_{\mathrm{sq}}\big(\mathcal{Y}_i, f(\boldsymbol{\mathcal{X}}_i)\big) - \ell_{\mathrm{sq}}\big(\mathcal{Y}_i, f_*(\boldsymbol{\mathcal{X}}_i)\big) = \big(\mathcal{W}_{f,i} - 2(\mathcal{Y}_i - f_*(\boldsymbol{\mathcal{X}}_i))\mathcal{W}_{f,i}$ with Markov's inequality, the tower rule with the independece of $\mathcal{Y}_1, \dots, \mathcal{Y}_n$, and the bound $2\sigma^2/(nR_n) \le 1$, we obtain

$$\begin{aligned}
&\mathbb{P}\Big\{\inf_{f\in\hat{\mathcal{F}}_n} L_n(f, f_*) + T > B_*\Big\} \\
&\quad\le \frac{\gamma}{4c}\mathbb{E}\Big[\inf_{f\in\hat{\mathcal{F}}_n} e^{\left(\frac{1}{n}\sum_{i=1}^n \mathcal{W}_{f,i}^2 + 2\mathcal{W}_{f,i}(f_*(\boldsymbol{\mathcal{X}}_i)-\mathcal{Y}_i)+T\right)/R_n}\Big] \\
&\quad\le \frac{\gamma}{4c}\mathbb{E}\Big[\inf_{f\in\hat{\mathcal{F}}_n} e^{\left(\frac{1}{n}\sum_{i=1}^n \mathcal{W}_{f,i}^2 + T\right)/R_n} \prod_{i=1}^n \mathbb{E}\big[e^{\frac{2}{nR_n}\mathcal{W}_{f,i}(f_*(\boldsymbol{\mathcal{X}}_i)-\mathcal{Y}_i)}\big|\boldsymbol{\mathcal{X}}_1, \dots, \boldsymbol{\mathcal{X}}_n\big]\Big] \\
&\quad\le \frac{\gamma}{4c}\mathbb{E}\Big[\inf_{f\in\hat{\mathcal{F}}_n} e^{\left(T + (1+\frac{2\sigma^2}{nR_n})\frac{1}{n}\sum_{i=1}^n \mathcal{W}_{f,i}^2\right)/R_n}\Big] \le \frac{\gamma}{4c}\mathbb{E}\Big[e^{\mathcal{Z}/R}\Big] \le \frac{\gamma}{4}\,.
\end{aligned}$$

■