University of Alberta

Statistical analysis of L1-penalized linear estimation with applications

by

Bernardo Ávila Pires

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Fall 2011
Edmonton, Alberta

Abstract

We study linear estimation based on perturbed data when performance is measured by a matrix norm of the expected residual error, in particular, the case in which there are many unknowns, but the "best" estimator is sparse, or has small ℓ_1-norm. We propose a Lasso-like procedure that finds the minimizer of an ℓ_1-penalized squared norm of the residual. For linear regression we show $O\left(\sqrt{\frac{1}{n}}\right)$ uniform bounds for the difference between the residual error norm of our estimator and that of the "best" estimator. These also hold for on-policy value function approximation in reinforcement learning. In the off-policy case, we show $O\left(\sqrt{\frac{\ln n}{n}}\right)$ bounds for the expected difference. Our analysis has a unique feature: it is the same for both regression and reinforcement learning. We took care to separate the deterministic and probabilistic arguments, so as to analyze a range of seemingly different linear estimation problems in a unified way.

Contents

List of Tables

Chapter 1

Introduction

First of all, we state our problem in Section 1.1. After we do this, we will be able to properly present the goals and contributions of our work (Section 1.2), and a brief discussion about the choices we made to delimit the scope of our work, in Section 1.3. Section 1.4 briefly summarizes the content of the remaining chapters.

1.1 Problem statement

Suppose that there exist a matrix $A \in \mathbb{R}^{q \times d}$ and a vector $b \in \mathbb{R}^q$, and that we have a random matrix $\hat{A} \in \mathbb{R}^{q \times d}$ and a random vector $\hat{b} \in \mathbb{R}^q$ that are estimates of A and b, respectively. Consider the problem of minimizing the following *loss* [1]

$$L(\theta) = \|A\theta - b\|_M,$$

with respect to (w.r.t.) $\theta \in \mathbb{R}^d$, where $M \in \mathbb{R}^{q \times q}$ is a fixed (symmetric) positive-definite matrix. Of course, we do not have access to A or b, but only to $\hat{A}$ and $\hat{b}$. We will call this problem a linear estimation problem.

The study of certain types of linear estimation problems is not unprecedented. For example, much has been done in perturbation analysis (Bonnans and Shapiro (2000)), but there are certain intrinsic aspects of our problem, in particular the stochastic components of $\hat{A}$ and $\hat{b}$, that prevent us from easily applying results from perturbation analysis.

We are interested in bounding $L(\hat{\theta}_\lambda) - \inf_\theta L(\theta)$, where $\hat{\theta}_\lambda$ is a solution candidate defined as

$$\hat{\theta}_\lambda \doteq \arg\min_\theta \|\hat{A}\theta - \hat{b}\|_M^2 + \lambda\|\theta\|_1,$$

and is based on the Lasso estimator of Tibshirani (1996) (in Section 1.3 we will discuss why we chose to study a Lasso-like estimator). We call this gap $L(\hat{\theta}_\lambda) - \inf_\theta L(\theta)$ the *excess loss*.

1.2 Goals and contributions

This work is a theoretical study of a particular family of solution candidates $\hat{\theta}_\lambda$ (indexed by $\lambda \geq 0$) in three scenarios: under a generic linear estimation problem

[1] Recall that for $v \in \mathbb{R}^q$, $\|v\|_M^2 = v^\top M v$.

formulation as above, and in two specific applications, linear regression and value prediction in reinforcement learning (RL) Sutton and Barto (1998). The study of the (generic) linear estimation problem is further divided into three parts: a deterministic analysis, a stochastic analysis, and an investigation of the problem in a specific stochastic setting, which is meant to illustrate the applicability of our results.

In Section 1.2.1, we describe what the ideal form of our results would be, and summarize our theoretical findings. Of course, these findings will be properly detailed and explained in the coming chapters; in the following section we will only outline the results, so as to emphasize their important characteristics.

We believe that the analysis we have employed to obtain the results is also a contribution of this work. In Section 1.2.2 we discuss the novelty and the advantages of our "two-step" analysis.

1.2.1 Bounds

Recall the we wish to bound the excess loss $L(\hat{\theta}_\lambda) - \inf_\theta L(\theta)$. The quality of our results will depend on how we choose λ, so we will try to pick $\lambda = \hat{\lambda}$ to obtain a bound that holds with probability at least $1 - \delta$ for $0 < \delta < 1$ and has the form

$$L(\hat{\theta}_{\hat{\lambda}}) \leq c_\delta(\theta^*) + L(\theta^*), \tag{1.1}$$

where θ^* is a vector with minimum loss [2]:

$$\theta^* \doteq \arg\min_\theta L(\theta),$$

and $c_\delta(\theta^*)$ is an excess loss term that goes to zero as $\hat{A} - A$ and $\hat{b} - b$ concentrate around zero. We expect the "complexity" of θ^*, *i.e.*, its size in some norm, to affect $c_\delta(\theta^*)$, because we believe a θ^* with large norm can "amplify" the the error $\hat{A} - A$. Note that in (1.1) the multiplier of $L(\theta^*)$ is exactly one.

We would like c_δ to have a logarithmic relationship with δ, so that the bounds are high-probability bounds, and we want $\hat{\lambda}$ *not to be* a function of δ. We will refer to a bound in which $\hat{\lambda}$ does not depend on δ as a uniform bound (in δ). In other words, a bound for $L(\hat{\theta}_{\hat{\lambda}})$ will be uniform if, given $\hat{\theta}_{\hat{\lambda}}$, there will be a bound of the form of (1.1) that will hold with probability at least $1 - \delta$ for any $0 < \delta < 1$.

When $A\theta^* = b$, which holds iff $L(\theta^*) = 0$, we will say that *the system is consistent*, otherwise, we will say the system is *inconsistent*. When the system is consistent, the requirement of a $1 \cdot L(\theta^*)$ term in the bound is vacuous, so, for this case, proving the results we are after is easier, in the sense that there are less requirements to meet. In fact, we managed to prove a uniform bound with the desired aspects for the consistent case. In the more general case, *i.e.*, when we do not know whether or not $A\theta^* = b$, we showed a high-probability bound with all the desired requirements, except for uniformity, because $\hat{\lambda}$ is selected based on a choice of δ.

As an illustration of the form our results take under different stochastic assumptions, we present results for the case in which $\hat{A}$ and $\hat{b}$ are averages of n i.i.d. bounded random variables. For this particular scenario, it is well known that the excess loss is $\Omega\left(\sqrt{\frac{1}{n}}\right)$, so we aimed for tight bounds (w.r.t. to n), *i.e.*, for an excess

[2] If the minimizer is not unique, we will pick one with minimum ℓ_1-norm, for convenience.

loss that is also $O\left(\sqrt{\frac{1}{n}}\right)$. So, for this specific case, we also established that a fast rate is required, *i.e.*, we wish to have an $O\left(\sqrt{\frac{1}{n}}\right)$ rate.

We show that

$$L(\hat{\theta}_{\hat{\lambda}}) = O\left((\|\theta^*\|_1 + 1)^4 \sqrt{\frac{1}{n}}\right) \tag{1.2}$$

holds with high-probability (and uniformly in δ) when $A\theta^* = b$. In this case we propose a choice of $\hat{\theta}_{\hat{\lambda}}$ based on universal constants, but the results for the case in which we do not make assumptions about the consistency of the system require $\hat{\lambda}$ to be chosen in a peculiar way. In general, what one does is to obtain observations $\hat{A}'$ and $\hat{b}'$ of A and b that are independent of $\hat{A}$ and $\hat{b}$, then choose

$$\hat{\lambda} = \arg\min_{\lambda} \|\hat{A}'\hat{\theta}_{\lambda} - \hat{b}'\|_M.$$

We make our choice of $\hat{\lambda}$ based on $\hat{A}$, $\hat{b}$ and δ, in a way that will be explained in Chapter 2. What is important to say is that this manner of choosing $\hat{\lambda}$ is new, and that we also show how to, in computational terms, perform it efficiently, by modifying the procedure that computes $\hat{\theta}_{\lambda}$.

Under no assumption about the consistency of the system, we can show that

$$L(\hat{\theta}_{\hat{\lambda}}) = O\left((\|\theta^*\|_1 + 1)^2 \sqrt{\frac{1}{n}}\right) + L(\theta^*) \tag{1.3}$$

holds with high-probability, but non-uniformly in δ (because of the way $\hat{\lambda}$ is chosen).

These results have an interesting impact in the two applications we study in our work, linear regression and value prediction in reinforcement learning (RL). We study these applications because we wish to illustrate the simplicity of applying our general linear estimation analysis, and because we are interested in seeing the nature of the results we can prove with our analysis.

Our main contribution to linear regression comes from using (1.2) to show a fast, uniform rate for the excess risk Bartlett et al. (2009) of $\hat{\theta}_{\hat{\lambda}}$:

$$O\left((\|\theta^*\|_1 + 1)^2 \frac{1}{n\lambda_{\min}(C)}\right), \tag{1.4}$$

where $\lambda_{\min}(C)$ is the minimum eigenvalue of the covariance matrix of the input r.v..

There are three main contributions for value prediction RL. The first two are rates for the (unsquared) projected Bellman error (Antos et al. (2008)). In the on-policy case (Sutton and Barto (1998)), this rate is given by

$$O\left((\|\theta^*\|_1 + 1)\sqrt{\frac{1}{n\lambda_{\min}(C)}}\right), \tag{1.5}$$

and the bound for off-policy value prediction is has the form of

$$O\left((\|\theta^*\|_1 + 1)\sqrt{\frac{1}{n}}\right) + \sqrt{\frac{\lambda_{\max}(C)}{\lambda_{\min}(C)}}\,\|A\theta^* - b\|_{C^{-1}},$$

and is based on (1.3). The third contribution for value prediction in RL is that we can easily show that the least-squares temporal difference learning (LSTD, Bradtke and Barto (1996)) is solving a linear estimation problem, and therefore we can derive performance results for estimators computed by LSTD-like methods.

We would like to eliminate the above dependence on $\lambda_{\min}(C)$. Possible ways of doing so are discussed in 3.2, Remark 3.2.3.

1.2.2 Analysis

In order to show bounds for $L(\hat{\theta}_{\hat{\lambda}})$, we used a two-step analysis: first, we obtained bounds for the excess loss as a function of the errors $\hat{A} - A$ and $\hat{b} - b$, which we call a deterministic analysis, and which has similarities to perturbation analysis (Bonnans and Shapiro (2000)). Second, we applied stochastic assumptions (in the form of concentration inequalities) to obtain stochastic results. The main goal of doing the analysis this way is to decouple the stochastic arguments and properly delimit their role in the bounds. This way, the effect of the stochastic assumptions on the bound can be better understood, as well as the effect of the proof techniques employed in the deterministic step.

Moreover, the stochastic assumptions are used in a generic form that allows one to quickly derive results using different concentration inequalities. For example, we show results for when $\hat{A}$ and $\hat{b}$ are averages of i.i.d. bounded r.v.'s, but, *e.g.*, if one wants to obtain corresponding results for when these r.v.'s constitute a mixing process Doukhan (1994), it suffices to define a basic quantities according to the appropriate concentration inequalities to obtain the corresponding result.

We are also interested in the two-step analysis because it helps us choose $\hat{\lambda}$ to get uniform bounds. Normally, one first derives high-probability bounds in terms of $\hat{\lambda}$ and after applying the stochastic assumptions, chooses $\hat{\lambda}$ so as to minimize the right-hand side of the inequality. This will often lead to non-uniform bounds, because the "bound-optimal" choice of $\hat{\lambda}$ will be a function of δ.

The fact that we can easily modify our results by changing some of the underlying stochastic assumptions is a significant simplification of the commonplace method for deriving similar bounds, which is to perform the whole analysis from scratch for each different scenario, even if the only difference between them lies in the concentration inequalities used. This is why we consider our analysis to be another contribution of our work [3].

1.3 Motivation

In this section we briefly state some of the reasons for the important choices made to delimit the scope of our work. Though interesting, a thorough discussion of pros and cons of these decisions is orthogonal to our work, so we limit ourselves to elaborate on why we chose to address ℓ_1-regularized linear estimation the way we did.

[3] A similar two-step analysis has been studied by Rosasco (2006), but our analysis was developed independently.

1.3.1 Loss

Using $\|\cdot\|_M$, with $M \succ 0$, as our loss is convenient for developing results: in our case, for example, we have heavily relied upon triangle inequalities for this norm. The fact that we are using $\|\cdot\|_M^2$ in the definition of $\hat{\theta}_\lambda$ has computational advantages (we will discuss advantages of using other losses in Section 2.4).

Furthermore, this loss also works well with concentration inequalities and assumptions available in many scenarios [4].

All in all, although by no means the only one or the best one, the least squares loss is a *convenient* choice.

1.3.2 ℓ_1-norm penalization

Regularization is often a principled way of biasing estimators[5], and it is generally regarded as a useful way of mitigating overfitting.

Moreover, minimizing sample error with an ℓ_1-norm penalty $\lambda\|\theta\|_1$ is known to yield sparse estimators in certain cases (Tibshirani (1996); Hastie et al. (2009)). Sparsity may reduce the computational cost of estimation, and ℓ_1-norm regularization is suitable when a small number of covariates can be used to attain small excess loss.

Appropriate ℓ_1-norm regularization can often increase performance even in overcomplete spaces, *i.e.*, even when d is sensibly larger than q, but "appropriate regularization", in our case, choosing $\hat{\lambda}$ well, may not be simple at a first glance. In Section 2.4 we discuss the use of penalty functions different from $\lambda\|\theta\|_1$

Still, it is possible to make it so that $\hat{\lambda}$ is chosen automatically, and that is done using model selection.

1.4 Organization

This dissertation is meant to do more than just present our theoretical findings and some of their implications. We are also interested in guiding the reader through the process of building these results, a process which is valuable on its own. That is to say that not only do the main lemmas and corollaries give us insight about certain problems, but also does the analysis itself teach us how to develop theory with a good number of extensions to different scenarios.

This dissertation is divided into this introductory chapter, three main chapters, a conclusion, and two appendix chapters.

In Chapter 2 we state our core results, which are of two types, deterministic and stochastic. We also present extensions for the particular case in which $\hat{A}$ and $\hat{b}$ are averages of bounded i.i.d. r.v.'s . At the end of the chapter we summarize how to extend the stochastic results to other sampling models and other i.i.d. scenarios, and discuss a few generalizations.

We have chosen to present only the lemmas and corollaries in Chapter 2, and to append the proofs at the end of this dissertation, in appendix B.

[4] Its often criticized sensitivity to outliers is not a problem for when the quantities are bounded with probability one, but dealing with outliers is out of the scope of our work.

[5] Recall the James-Stein estimator James and Stein (1961) as an example of the advantages of using biased estimators to reduce the mean squared estimation error.

The two following chapters, 3 and 4, cover consequences of the i.i.d. results to regression and value prediction in RL. Each chapter is introduced by a formulation of the respective problem as a linear estimation problem. Then, the consequences of our results and relevant related works are discussed.

In the conclusion chapter we summarize our findings, and pose the questions that could be studied as a continuation of this work. We also include a few less related, but interesting, questions that arose during our work.

The first appendix chapter, A, discusses computing $\hat{\theta}_\lambda$ and $\hat{\theta}_{\hat{\lambda}}$ for our specific model selection scheme. There are some simple observations as well as previous works in the literature that considerably improve performance over naive implementations.

Appendix B, as said, contains proofs for the results in Chapter 2.

Chapter 2

Core Results

This chapter has the most essential contributions of our work. The results in chapters 3 and 4 are built upon those results of this chapter, especially those of Section 2.3, which are the corollaries for scenarios with i.i.d. sampling and sub-gaussian noise.

Section 2.1 encompasses the basic lemmas, which are stated in terms of key deterministic quantities. In Section 2.2 we adapt these results under generic concentration assumptions about the key quantities.

We have structured the analysis in such a way that the role of stochastic assumptions becomes clearer in our results. The evidence of where these assumptions affect our bounds allows us to bound estimation errors using different concentration inequalities and to choose λ uniformly w.r.t. to the confidence parameter of the bounds, δ.

In Section 2.3, we illustrate the application of our results to a scenario in which $\hat{A}$ and $\hat{b}$ are averages of n i.i.d. random variables. The study of this kind of process is quite common, and so the instantiation of the results from Section 2.2 will emphasize the simplicity of our results and help provide a better understanding of the meaning and the behavior of the stochastic results.

Let us state beforehand the definitions that will allow us to decouple the deterministic and the stochastic results (recall that $A, \hat{A} \in \mathbb{R}^{q\times d}$, $b, \hat{b} \in \mathbb{R}^q$ and $M \in \mathbb{R}^{q\times q}$).

Definition 2.0.1. *Given a matrix $M \succ 0$, let*

$$\Delta_A \doteq \|M^{\frac{1}{2}}(A - \hat{A})\|_F,$$
$$\Delta_b \doteq \|M^{\frac{1}{2}}(b - \hat{b})\|_2.$$

In Chapter 1, we used the notation $L(\theta)$ to denote $\|A\theta - b\|_M$. We will use these two terms interchangeably, because while in some places it is more convenient to use $L(\theta)$, some of the lemmas and corollaries are easier to interpret if we write them explicitly in terms of $\|A\theta - b\|_M$.

2.1 Deterministic analysis

In the first part of this section, Section 2.1.1, we derive results and properties of $\hat{\theta}_\lambda$ that hold for any fixed λ, and then in the second part (Section 2.1.2), we bound the behavior of $L(\hat{\theta}_{\hat{\lambda}})$, and explain how we choose $\hat{\lambda}$ from a set Λ.

2.1.1 Fixed choice of λ

The following lemma displays a very useful technique to relate different kinds of losses in a way that the quantities in Definition 2.0.1, as well as an ℓ_1-norm term, emerge. This association will be needed because we know how to control $\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M$, and we want to use this in order to bound $\|A\hat{\theta}_\lambda - b\|_M$.

Lemma 2.1.1. *For any $\theta \in \mathbb{R}^d$,*

$$\left| \|A\theta - b\|_M - \|\hat{A}\theta - \hat{b}\|_M \right| \leq \Delta_A \|\theta\|_1 + \Delta_b.$$

To see that this holds, observe that that $\|(A - \hat{A})\theta\|_M \leq \|M^{-\frac{1}{2}}(A - \hat{A})\|_F \|\theta\|_2$ and $\|\theta\|_2 \leq \|\theta\|_1$, then use triangle inequalities w.r.t. to $\|\cdot\|_M$.

The next lemma shows how we control $\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M$: we use the definition of $\hat{\theta}_\lambda$ to relate its empirical loss to that of θ^*. This "optimality" property will allow us upper-bound $\|\hat{\theta}_\lambda\|_1$ (provided that $\lambda > 0$) as well as $\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M^2$ by functions of θ^* (*e.g.*, we can apply Lemma 2.1.1 to θ^* and $\|A\theta^* - b\|_M$ will emerge).

Lemma 2.1.2. *We have*

$$\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M^2 \leq \|\hat{A}\theta^* - \hat{b}\|_M^2 + \lambda(\|\theta^*\|_1 - \|\hat{\theta}_\lambda\|_1).$$

Our first result is a bound on $L(\hat{\theta}_\lambda)$ with a specific λ chosen as a function of A, b, $\|\theta^*\|_1$ and $L(\theta^*)$, which are in practice unknown to us. We these choices that are functions of unknown quantities as *oracle* choices. This kind of result is important to us because it may provide insight about the hardness of the problem and the effectiveness of our proof techniques. Moreover, this oracle choice is useful for proving Lemma 2.1.6. The result for the oracle choice of λ is stated in Corollary 2.1.4, which follows from Lemma 2.1.3. This lemma itself is the common starting point for the proof of the remaining deterministic results, as well as of the stochastic ones.

Lemma 2.1.3. *For any $\zeta, \zeta' \geq \|\hat{A}\theta^* - \hat{b}\|_M$ and $C_1, C_2 > 0$, it holds that*

$$\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + C_1\|\hat{\theta}_\lambda\|_1 + C_2 \leq \max\left\{\left(C_1 - \frac{\lambda}{2\zeta}\right)\left(\frac{\zeta'^2}{\lambda} + \|\theta^*\|_1\right), 0\right\} + \frac{\lambda}{2\zeta}\|\theta^*\|_1 + C_2 + \zeta.$$

Corollary 2.1.4. *If*

$$\hat{\lambda} = 2\Delta_A(\Delta_A\|\theta^*\|_1 + \Delta_b + \|A\theta^* - b\|_M),$$

then

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq 2\Delta_A\|\theta^*\|_1 + 2\Delta_b + \|A\theta^* - b\|_M.$$

This corollary says that there appropriate regularization can be used so that $L(\hat{\theta}_\lambda)$ is not much larger than$L(\theta^*)$. Unfortunately, we cannot choose $\hat{\lambda}$ as in Corollary 2.1.4, because it depends on the knowledge of A, $\|\theta^*\|_1$ and $L(\theta^*)$. Still, this oracle choice is a key quantity in the proofs of our model selection bounds.

2.1.2 Model selection

Model selection is a general term for a process of choosing the best model – an estimator, a classifier, a predictor – among different options. In this work, we will abuse nomenclature so that whenever we say model selection we will specifically refer to processes that choose $\hat{\lambda}$ from a set of candidate values Λ, such that $\hat{\theta}_{\hat{\lambda}}$ has certain properties. In other words, given some measure of performance U, the model selection will find the maximizer in Λ of $U(\hat{\theta}_\lambda)$, and we hope that knowing about $U(\hat{\theta}_{\hat{\lambda}})$ we will be able to prove bounds for $L(\hat{\theta}_{\hat{\lambda}})$.

Our model selection is done by picking

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda(a,b)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda' \|\hat{\theta}_\lambda\|_1,$$

where $\lambda' \geq 0$ is to be appropriately chosen by us, and $\Lambda(a, b)$ is an exponential grid of points on the interval $[a, B]$ ($0 < a \leq B$)

$$\Lambda(a, B) \doteq \left\{ a2^k : k \in \mathbb{N}, 0 \leq k \leq \left\lfloor \log_2 \frac{B}{a} \right\rfloor \right\} \cup \{B\},$$

for suitable a, B.

We will derive bounds as functions of a, and then, during the stochastic analysis, make convenient choices of a and λ'. The choice of B is made based on Assumption 2.1.5.

Assumption 2.1.5. *There exists a $B > 0$ such that, with probability one (w.p. 1), $\hat{\theta}_B = \mathbf{0}$.*

The KKT optimality conditions for the Lasso estimator (exercise 3.26 of Hastie et al. (2009)), imply that this is the case when, w.p. 1, $c = \|\hat{A}^\top M \hat{b}\|_\infty < \infty$, because $\hat{\theta}_\lambda = \mathbf{0}$ for any $\lambda \geq c$.

Under Assumption 2.1.5, $\hat{\theta}_\lambda = \mathbf{0}$ for every $\lambda \geq B$, so we will restrict ourselves to analyzing the model selection for candidates with $\lambda \in [a, B]$, and this allows us to prove the following lemma:

Lemma 2.1.6. *Under Assumption 2.1.5, if*

$$\hat{\lambda} \doteq \arg \min_{\lambda \in \Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda' \|\hat{\theta}_\lambda\|_1$$

and $\lambda' \geq \Delta_A$, then, for any $\zeta \geq \|\hat{A}\theta^* - \hat{b}\|_M$,

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq \max\left\{2\lambda', \frac{a}{2\zeta}\right\} \|\theta^*\|_1 + \Delta_b + \zeta.$$

Note that we do not need to split our sample into a training and a testing set. Instead, we use Lemma 2.1.1 to express $\|A\hat{\theta}_{\hat{\lambda}} - b\|_M$ in terms of $\|\hat{A}\hat{\theta}_{\hat{\lambda}} - b\|_M$, and then we use λ' to eliminate the term $\Delta_A \|\hat{\theta}_{\hat{\lambda}}\|_1$ which is not as easy to control as the other terms in the bound.

It is possible to see that we can already make a meaningful choice of a, based on a choice of ζ. We can take, for example, $\zeta = \Delta_A \|\theta^*\|_1 + \Delta_b + c$ and $a = 2c\lambda'$, with c being of the same scale of Δ_b and s.t. $2c\lambda' \leq B$. Then, provided that λ' dominates Δ_A but also concentrates around zero, we have a bound in the form we would like to have, with the exception of the uniformity in δ. We will instantiate a in the next section, when we start to use concentration inequalities that will allow us to choose c in a suitable fashion.

2.2 Stochastic analysis

In this section we provide extensions of the deterministic lemmas by assuming quantities such as Δ_A, Δ_b concentrate around zero. In this section as in the previous one, we first work on specific choices λ, and provide results for the model selection scheme afterwards.

2.2.1 Fixed choice of λ

In this section we provide a high-probability bound for $\|A\hat{\theta}_\lambda - b\|_M$ when $A\theta^* = b$, which we refer to as the realizable case, and a bound for $\mathbb{E}\left[\|A\hat{\theta}_\lambda - b\|_M\right]$ when $A\theta^* \neq b$, which we call the unrealizable case. The following lemma presents the first bound.

Lemma 2.2.1 (Uniform convergence bound for a fixed choice of λ, when $A\theta^* = b$)**.** *Assume that $A\theta^* = b$ and that there exist a positive constant c_1 and a decreasing function $S : (0,1) \to (0,\infty)$ such that, for any $0 < \delta < 1$, with probability at least $1-\delta$, the following hold simultaneously:*

$$\Delta_A \leq c_1 S\left(\frac{\delta}{2}\right), \tag{2.1}$$

$$\Delta_b \leq c_1 S\left(\frac{\delta}{2}\right). \tag{2.2}$$

If

$$\hat{\lambda} = c_1^2,$$

then, for any $0 < \delta < 1$, with probability at least $1-\delta$, it holds that

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq c_1 \cdot \max\left\{(\|\theta^*\|_1 + 1)^2 S\left(\frac{\delta}{2}\right)^3 + \frac{1}{2}S\left(\frac{\delta}{2}\right)(\|\theta^*\|_1 + 1),\right.$$
$$\left.\frac{1}{2S\left(\frac{\delta}{2}\right)}\frac{\|\theta^*\|_1}{\|\theta^*\|_1 + 1} + S\left(\frac{\delta}{2}\right)\right\} + c_1 S\left(\frac{\delta}{2}\right)\|\theta^*\|_1 + c_1 S\left(\frac{\delta}{2}\right).$$

Lemma 2.2.1 implies that for any $0 < \delta < \frac{1}{2e}$,

$$\|A\hat{\theta}_\lambda - b\|_M \leq c_1 2(\|\theta^*\| + 2)^2 S\left(\frac{\delta}{2}\right)^3$$

holds with probability at least $1-\delta$. This is a crude upper-bound that allows us to see how we exploited that $L(\theta^*) = 0$ in order to choose λ and control the term $\frac{\Delta_A}{\lambda}(\Delta_A\|\theta^*\|_1 + \Delta_b)^2$, which is the upper-bound we use for the term $\Delta_A\|\hat{\theta}_\lambda\|_1$.

The next lemma provides an upper-bound for $\mathbb{E}\left[\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M\right]$ with a fixed choice of λ.

Lemma 2.2.2 (Expectation bound for a fixed choice of λ)**.** *Assume that there exist a positive constant $c_1 < 1$ and a decreasing function $S : (0,1) \to (0,\infty)$ such that, for any $0 < \delta < 1$, with probability at least $1-\delta$,*

$$\Delta_A \leq c_1 S(\delta).$$

Moreover, assume that w.p. 1

$$\|\hat{A}\theta^* - \hat{b}\|_M \leq L,$$
$$\Delta_A \leq A_{\max},$$

and that there exists c_2 *such that*

$$\mathrm{Var}(\|\hat{A}\theta^* - \hat{b}\|_M) \leq c_2^2.$$

Then, choosing

$$\hat{\lambda} = 2(L + c_2)c_1 S\left(c_1^2\right)$$

implies that

$$\mathbb{E}\left[\|A\hat{\theta}_\lambda - b\|_M\right] \leq A_{\max}\left(\frac{L}{2S\left(c_1^2\right)} + c_1\|\theta^*\|_1\right)c_1 + \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c_2$$
$$+ 2\frac{L + c_2}{\mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c_2}\|\theta^*\|_1 c_1 S\left(c_1^2\right) + \mathbb{E}\left[\Delta_b\right].$$

The bound in Lemma 2.2.2 is important because its proof technique – tail integration – can be used to extend high-probability bounds, *e.g.*, that of Lemma 2.2.1, into expectation bounds. The strategy is simple: state the concentration assumption of the high-probability bound in the same form as in lemma 2.2.2, take the expectations on both sides, and finally choose δ – the size of the events where quantities do not concentrate – for a suitable rate.

In Section 2.3 we will see why it is harder to go in the opposite direction, *i.e.*, derive high-probability bounds based on expectation bounds such as that of Lemma 2.2.2. Still, it is possible to prove high-probability error bounds for the case in which $A\theta^* \neq b$, and this is done by using model selection.

Now, choosing λ as in Lemma 2.2.2 may not be possible in practice because even if we can assume that $\|\hat{A}\theta^* - \hat{b}\|_M \leq L$, often the values of L that are "available" depend on unknown quantities such as $\|\theta^*\|_1$. Fortunately, we can avoid having to select a good λ by using model selection, which we cover next.

2.2.2 Model Selection

In this section we will state the results of performing model selection using an exponential grid. The upper-bound of the grid interval will be B (from Assumption 2.1.5), and the value of a is chosen in order to optimize the bound. This upper-bounding strategy is based in the general approach to model selection adopted by Farahmand and Szepesvári (2011), as expressed in their Theorem 1.

The following two results are a high-probability bound on the error of the selected model, and an expectation bound on that error. Both are non-uniform in δ because λ' is a function of it.

Corollary 2.2.3 (Non-uniform high-probability performance bound for estimators obtained through model selection)**.** *Assume that there exist positive constants* $c_1, c_2 <$

1, *and decreasing functions* $S_1, S_2 : (0,1) \to (0,\infty)$ *such that, for any* $0 < \delta < 1$, *with probability at least* $1-\delta$, *the following hold simultaneously:*

$$\Delta_A \leq c_1 S_1\left(\frac{\delta}{2}\right),$$
$$\Delta_b \leq c_2 S_2\left(\frac{\delta}{2}\right).$$

Then, under Assumption 2.1.5, for any $0 < \delta < 1$, *if*

$$\lambda' = c_1 S_1\left(\frac{\delta}{2}\right),$$
$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda(2c_3\lambda', B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda'\|\hat{\theta}_\lambda\|_1,$$

where c_3 *is a constant s.t.* $0 < 2c_3\lambda' \leq B$, *then, with probability at least* $1-\delta$, *it holds that*

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq 3c_1 S_1\left(\frac{\delta}{2}\right)\|\theta^*\|_1 + 2c_2 S_2\left(\frac{\delta}{2}\right) + \|A\theta^* - b\|_M + c_3.$$

Corollary 2.2.4 (Expectation performance bound for estimators obtained through model selection)**.** *Assume that there exist positive constants* $c_1, c_2 < 1$, *and a decreasing function* $S : (0,1) \to (0,\infty)$ *such that, for any* $0 < \delta < 1$, *with probability at least* $1-\delta$, *the following holds:*

$$\Delta_A \leq c_1 S(\delta).$$

Also, assume that there exist constants $A_{\max}$ *and* $b_{\max}$ *s.t.*

$$\Delta_A \leq A_{\max},$$
$$\Delta_b \leq b_{\max},$$

w.p. 1. Under Assumption 2.1.5, if, for any $0 < \delta < 1$,

$$\lambda' = c_1 S\left(\frac{c_3 c_1^2}{2}\right),$$
$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda(2c_3\lambda', B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda'\|\hat{\theta}_\lambda\|_1,$$

where c_3 *is a constant s.t.* $0 < 2c_3\lambda' \leq B$, *then it holds that*

$$\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}\left(\frac{b_{\max}^2}{2S\left(\frac{c_3 c_1^2}{2}\right)}\right)c_1 + 2c_1 S\left(\frac{c_3 c_1^2}{2}\right)\|\theta^*\|_1$$
$$+ \mathbb{E}\left[\Delta_A\right]\|\theta^*\|_1 + 2\mathbb{E}\left[\Delta_b\right] + \|A\theta^* - b\|_M + c_3.$$

This concludes the generic stochastic analysis, and we will now proceed to a more specific study of our results.

2.3 Independent, identically-distributed sampling analysis

In this section, we extend our results to cases where $\hat{A}$ and $\hat{b}$ are averages of i.i.d. random variables.

Assumption 2.3.1. *Assume we have a sample (A_t, b_t) for observations $t = 1, \ldots, n$ drawn i.i.d. w.r.t. to a probability measure ρ, and let*

$$\hat{A} = \frac{1}{n}\sum_{t=1}^{n} A_t,$$
$$\hat{b} = \frac{1}{n}\sum_{t=1}^{n} b_t.$$

Hoeffding's inequality will provide the necessary concentration inequalities. In particular, Corollary 6.15 in Steinwart and Christmann (2008):

Theorem 2.3.2 (Hoeffding's inequality in Hilbert Spaces)**.** *Let H be a separable Hilbert space and $B > 0$. Furthermore, let $\xi_1, \ldots, \xi_n$ be independent H-valued random variables satisfying $\|\xi_i\| \leq B$ almost surely for all $i = 1, \ldots, n$. Then, for all $0 < \delta < 1$,*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}\left[\xi_i\right])\right\|_H \geq B\sqrt{\frac{2}{n}\ln\frac{1}{\delta}} + B\sqrt{\frac{1}{n}} + \frac{4B}{3n}\ln\frac{1}{\delta}\right) \leq \delta.$$

We make the following definition to be concise and to match the general form of the concentration assumptions in the lemmas of Section 2.2.

Definition 2.3.3. *Let*

$$T\left(n, \delta\right) \doteq \sqrt{\ln\frac{1}{\delta}} + \sqrt{\frac{1}{2}} + \sqrt{\frac{2}{n}} \cdot \frac{2}{3}\ln\frac{1}{\delta}.$$

It only remains to make some boundedness assumptions, which will guarantee the necessary conditions to apply Hoeffding's inequality.

Assumption 2.3.4. *Assume that there exist constants $F_{2,\infty}$, $F'_{2,\infty}$, R_∞ such that, w.p. 1,*

$$\|M^{\frac{1}{2}} A_t\|_F \leq F_{2,\infty} F'_{2,\infty},$$
$$\|M^{\frac{1}{2}} b_t\|_2 \leq F_{2,\infty} R_\infty.$$

Corollary 2.3.5. *Assumption 2.3.4 implies that the following hold w.p. 1:*

$$\Delta_A \leq 2F_{2,\infty}F'_{2,\infty},$$
$$\Delta_b \leq 2F_{2,\infty}R_\infty,$$
$$\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M \leq F_{2,\infty}R_\infty.$$

We are now ready to state the concentration inequalities we will use.

Corollary 2.3.6. *Consider Assumptions 2.3.4. From theorem 2.3.2 it follows that*

$$\mathbb{P}\left(\Delta_A > 2F_{2,\infty}F'_{2,\infty}T\left(n,\delta\right)\sqrt{\frac{2}{n}}\right) \leq \delta,$$
$$\mathbb{P}\left(\Delta_b > 2F_{2,\infty}R_{\infty}T\left(n,\delta\right)\sqrt{\frac{2}{n}}\right) \leq \delta,$$

where T is as in Definition 2.3.3.

These concentration inequalities allow us to apply Lemmas 2.2.1 and 2.2.2, and Corollary 2.1.4.

2.3.1 Fixed choice of λ

Corollary 2.3.7 shows us what rate we can get for an oracle choice of λ.

Corollary 2.3.7. *Under Assumptions 2.3.1 and 2.3.4, if*

$$\lambda = 2\Delta_A(\Delta_A\|\theta^*\|_1 + \Delta_b + \|A\theta^* - b\|_M),$$

then, for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\|A\hat{\theta}_\lambda - b\|_M \leq 4F_{2,\infty}F'_{2,\infty}\sqrt{\frac{2}{n}}T\left(n,\delta\right)\|\theta^*\|_1 + 4F_{2,\infty}R_{\infty}\sqrt{\frac{2}{n}}T\left(n,\delta\right) + \|A\theta^* - b\|_M.$$

The choice of λ in this case is a (random) function of $\hat{A}$, $\hat{b}$ and some unknown quantities (*e.g.*, $\|\theta^*\|_1$), *not a function of the range constants and δ*[1]. The choice of λ that optimizes the bound in Corollary 2.3.7 behaves as $O\left(\frac{1}{n}\right)$ when $A\theta^* = b$, and otherwise as $O\left(\sqrt{\frac{1}{n}}\right)$. This is not to say that these rates indicate appropriate choices of λ, because to claim that we would need lower bounds, but it does show what performance ensues from certain choices of λ.

The next two results are the extensions for fixed choices of λ that are based on quantities assumed to be known *a priori*. The first result follows from Lemma 2.2.1.

Corollary 2.3.8. *Under Assumptions 2.3.1 and 2.3.4, if $A\theta^* = b$ and*

$$\hat{\lambda} = \frac{2}{n}F_{2,\infty}^2\max\{R_\infty, F'_{2,\infty}\}^2,$$

then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, it holds that

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq F_{2,\infty}\max\{R_\infty, F'_{2,\infty}\}\sqrt{\frac{2}{n}} \cdot \max\left\{(\|\theta^*\|_1 + 1)^2 T\left(n, \frac{\delta}{2}\right)^3 + \right.$$
$$\left. \frac{1}{2}T\left(n,\frac{\delta}{2}\right)(\|\theta^*\|_1 + 1), \frac{1}{2T\left(n,\frac{\delta}{2}\right)}\frac{\|\theta^*\|_1}{\|\theta^*\|_1 + 1} + T\left(n,\frac{\delta}{2}\right)\right\}$$
$$+ F_{2,\infty}\max\{R_\infty, F'_{2,\infty}\}\sqrt{\frac{2}{n}}T\left(n,\frac{\delta}{2}\right)(\|\theta^*\|_1 + 1),$$

where T is as in Definition 2.3.3.

[1] As discussed before, a choice of λ depending on δ would lead to non-uniformity of the bound, and require us to provide a confidence parameter prior to the computation of $\hat{\theta}_\lambda$.

This corollary shows that when $A\theta^* = b$, $\|A\hat{\theta}_{\hat{\lambda}} - b\|_M = O\left(\sqrt{\frac{1}{n}}\right)$, and that the chosen regularizer $\lambda = O\left(\frac{1}{n}\right)$. In terms of n, both the rate of convergence and the order of λ display the same behavior as the respective quantities for the oracle choice of the regularizer.

Corollary 2.3.8 suggests that the case in which $A\theta^* = b$ is easy enough so that we can perform as well as an oracle, modulo some constants, with a fixed choice of λ. We were unable to prove the same type of result for the non-realizable case [2], therefore we present a bound based on Lemma 2.2.2, showing that $\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] = O\left(\sqrt{\frac{\ln n}{n}} + \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right]\right)$.

Corollary 2.3.9. *Assume that there exists $L > 0$ s.t., w.p. 1,*

$$\|\hat{A}\theta^* - \hat{b}\|_M < L,$$

and let

$$c \doteq \sqrt{\frac{2}{n}} F_{2,\infty}(F'_{2,\infty}\|\theta^*\|_1 + R_\infty)\left(T\left(n, \frac{1}{n}\right) + 2\right),$$

where T is as in Definition 2.3.3. Under Assumptions 2.3.1 and 2.3.4, if $n > 2$ and

$$\hat{\lambda} = 2(L + c)\sqrt{\frac{2}{n}} F_{2,\infty} F'_{2,\infty} T\left(n, \frac{2}{n}\right),$$

then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, it holds that

$$\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}\left(\frac{L + c}{2F_{2,\infty}F'_{2,\infty}T\left(n, \frac{2}{n}\right)} + \sqrt{\frac{2}{n}}\|\theta^*\|_1\right)\sqrt{\frac{2}{n}} + \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c + 2\frac{L + c}{\mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c}\|\theta^*\|_1\sqrt{\frac{2}{n}} F_{2,\infty}F'_{2,\infty}T\left(n, \frac{2}{n}\right) + \mathbb{E}\left[\Delta_b\right].$$

where T is as in Definition 2.3.3.

The two previous corollaries would combine nicely if they used the same choice of λ. Unfortunately, this is not the case, and *a priori* we may not know whether or not $A\theta^* = b$, but, this can be worked around using model selection. As it is shown in the next section, we can make a suitable choice of λ whose associated model has some performance guarantees without us needing to know whether $A\theta^* = 0$.

2.3.2 Model selection

In order to be able to apply the model selection lemma (Lemma 2.1.6), we must first show that Assumption 2.1.5 is satisfied under Assumptions 2.3.4. Letting

$$B' = \inf_{\lambda \geq 0 : \hat{\theta}_\lambda = \mathbf{0}} \lambda,$$

[2] By starting from Lemma 2.1.3, it is possible to see that choosing a fixed λ when $A\theta^* \neq b$ forces terms such as $\|A\theta^* - b\|_M^2$ or $2\|A\theta^* - b\|_M$ to appear in the right-hand side.

from the Lasso optimality conditions (Hastie et al. (2009), exercise 3.26) it must hold that $\|\hat{A}^\top M\hat{b}\|_\infty = B'$, and so

$$\begin{aligned} B' &= \|\hat{A}^\top M\hat{b}\|_\infty \\ &\leq \|\hat{A}^\top M\hat{b}\|_2 \\ &\leq \|M^{\frac{1}{2}}\hat{A}\|_F \|M^{\frac{1}{2}}\hat{b}\|_2 \\ &\leq F_{2,\infty}^2 F'_{2,\infty} R_\infty \end{aligned}$$

with probability one. Therefore, with $B = F_{2,\infty}^2 F'_{2,\infty} R_\infty \geq B'$, Assumption 2.1.5 is satisfied, because $\hat{\theta}_\lambda = \mathbf{0}$ for any $\lambda \geq B$, and so Lemma 2.1.6 can be applied.

The next corollary shows that by using model selection we can obtain an $O\left(\sqrt{\frac{1}{n}}\right)$ rate of convergence for $\|A\hat{\theta}_{\hat{\lambda}} - b\|_M - \|A\theta^* - b\|_M$, that is non-uniform in δ, with high probability (w.h.p.). We also obtain an $O\left(\sqrt{\frac{\ln n}{n}}\right)$ rate in expectation. The rate is not as fast as that of Corollary 2.3.8, and lacks uniformity in δ, but it applies more generally.

Corollary 2.3.10. *Under Assumptions 2.3.1 and 2.3.4, for any $0 < \delta < 1$, if*

$$\begin{aligned} \lambda' &= F_{2,\infty} F'_{2,\infty} \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right), \\ a &= 4\sqrt{\frac{1}{n}}\lambda', \\ B &= F_{2,\infty}^2 F'_{2,\infty} R_\infty, \\ \hat{\lambda} &= \arg\min_{\lambda \in \Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda' \|\hat{\theta}_\lambda\|_1, \end{aligned}$$

then, for n large enough so that $a \leq B$, with probability at least $1 - \delta$, it holds that

$$\begin{aligned} \|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq\ & 3F_{2,\infty} F'_{2,\infty} \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right) \|\theta^*\|_1 + 2F_{2,\infty} R_\infty \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right) \\ & + \|A\theta^* - b\|_M + \sqrt{\frac{1}{n}}. \end{aligned}$$

Corollary 2.3.11. *Under Assumptions 2.3.1 and 2.3.4, for any $0 < \delta < 1$, if*

$$\begin{aligned} \lambda' &= F_{2,\infty} F'_{2,\infty} \sqrt{\frac{2}{n}} T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right), \\ a &= 4\sqrt{\frac{1}{n}}\lambda', \\ B &= F_{2,\infty}^2 F'_{2,\infty} R_\infty, \\ \hat{\lambda} &= \arg\min_{\lambda \in \Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda' \|\hat{\theta}_\lambda\|_1, \end{aligned}$$

then it holds that

$$\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}\left(\frac{b_{\max}^2}{2F_{2,\infty}F'_{2,\infty}\sqrt{2}T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right)}\right)\sqrt{\frac{1}{n}} + 2\sqrt{\frac{1}{n}}F_{2,\infty}F'_{2,\infty}\sqrt{2}T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right)\|\theta^*\|_1 + \mathbb{E}\left[\Delta_A\right]\|\theta^*\|_1 + 2\mathbb{E}\left[\Delta_b\right] + \|A\theta^* - b\|_M + \sqrt{\frac{1}{n}}.$$

Note also that for this model selection scheme we only need to evaluate $\|\hat{A}\theta - \hat{b}\|_M$ for $\left\lceil \ln \frac{B}{a} \right\rceil = O\left(\ln n\right)$ models. This has convenient consequences to computation, which will be detailed in Appendix A.

2.4 Extensions and generalizations

Our results allow simple and useful extensions, and we believe they can also be generalized in interesting ways for different losses and penalties. This section is meant to provide summaries of how to make such extensions, and a discussion on the generalizations.

2.4.1 Extensions

The following list shows the steps needed to apply our results for specific scenarios.

- **Define M, A, b, $\hat{A}$, $\hat{b}$.** These are the key quantities upon which the deterministic results are based;
- **State the concentration inequalities.** They are related to how the sample is obtained, and they are necessary to apply the stochastic results.
- **Establish the correspondence between quantities of the concentration inequalities.** Map the constants in the chosen inequalities into those required by the lemmas in Section 2.2.
- **Ensure that Assumption 2.1.5 holds.** This is only necessary to extend the model selection results.

For scenarios with i.i.d. observations, most of the previous steps have been taken care of in Section 2.3, so that only two steps are necessary:

- **define A_t and b_t;**
- **ensure that Assumptions 2.3.4 hold.**

2.4.2 Generalizations to other solution candidates

It is important to point out that if we assume that with probability at least $1 - \delta$, for $0 < \delta < 1$, $\Delta_A \leq c_1 \S \frac{\delta}{2}$ and $\Delta_b \leq c_2 S'\left(\frac{\delta}{2}\right)$ hold simultaneously (as assumed in Corollary 2.2.3), and if we define

$$\hat{\theta} \in \arg\min_{\theta} \|\hat{A}\theta - \hat{b}\|_M + c_1 \S \frac{\delta}{2} \|\theta\|_1,$$

then it is easy to see that, without the need to make assumptions about the consistency of the system $A\theta = b$ (*i.e.*, it may be the case that $A\theta^* \neq b$), the following holds

$$\|A\hat{\theta} - b\|_M \leq 2c_1 \S \frac{\delta}{2} \|\theta^*\|_1 + 2c_2 S'\left(\frac{\delta}{2}\right) + \|A\theta^* - b\|_M.$$

This is a non-uniform result with the same form as the result we would get by applying concentration assumptions to Corollary 2.1.4 would be.

Therefore, studying the excess loss of $\hat{\theta}$ is easier, and computing it is easy, because $\|\hat{A}\theta - \hat{b}\|_M + c_1 \S \frac{\delta}{2} \|\theta\|_1$ is a convex function of θ. Still, it remains to be seen whether the gap in computational complexity between computing $\hat{\theta}_{\hat{\lambda}}$ and $\hat{\theta}$ is small enough that it is worth using $\hat{\theta}$ in practice, rather than $\hat{\theta}_{\hat{\lambda}}$ chosen by model selection.

2.4.3 Generalizations to other penalties

It is not hard to see that similar results should also hold for Ridge regression Hastie et al. (2009) and the elastic net Zou and Hastie (2005). For the Ridge regression estimator,

$$\hat{\theta}_\beta \in \arg\min_{\theta} \|\hat{A}\theta - \hat{b}\|_M^2 + \beta \|\theta\|_2^2,$$

a result corresponding to Corollary 2.1.4 would be

$$\|A\hat{\theta}_\beta\|_M \leq \Delta_A \max\left\{1 + \|\theta^*\|_2^2, 2\|\theta^*\|_2^2\right\} + 2\Delta_b + \|A\theta^* - b\|_M,$$

with $\beta = 2\Delta_A \|A\theta^* - b\|_M$. The max term occurs because we decompose $\|(A - \hat{A})\hat{\theta}_\beta\|_M \leq \Delta_A \|\hat{\theta}_\beta\|_2$, and the upper-bound $\|\hat{\theta}_\beta\|_2$.

Likewise, for the Elastic net estimator with $0 \leq \alpha \leq 1$,

$$\hat{\theta}_{\alpha,\beta,\lambda} \in \arg\min_{\theta} \|\hat{A}\theta - \hat{b}\|_M^2 + \alpha\lambda\|\theta\|_1 + (1 - \alpha)\beta\|\theta\|_2^2,$$

the correspondent of Corollary 2.1.4 would be an α-convex combination of the results for Lasso and Ridge regression:

$$\|A\hat{\theta}_\beta\|_M \leq \alpha 2\Delta_A \|\theta^*\|_1 + (1 - \alpha)\Delta_A \max\left\{1 + \|\theta^*\|_2^2, 2\|\theta^*\|_2^2\right\} + 2\Delta_b + \|A\theta^* - b\|_M,$$

with $\beta = \lambda = 2\Delta_A \|A\theta^* - b\|_M$.

The modification of the results to other types of penalties $p(\cdot)$ requires three considerations. First, if the penalty is a random function, then these dependencies need to be taken into account during the stochastic analysis. Second, the Δ quantities will need to be properly redefined: for example, Δ_A must be changed so that

$\|(\hat{A} - A)\theta\|_M \leq \Delta_A p(\theta)$. This is needed because the proofs are constructed so that we can cancel out factors of $p(\hat{\theta}_\lambda)$ or $p(\hat{\theta}_{\hat{\lambda}})$ using some function of $p(\theta^*)$. In fact, this is the third consideration: to make sure that $p(\hat{\theta}_\lambda)$ is either bounded (w.p. 1, or w.h.p.), or that we can somehow replace it in the bound for a factor of $p(\theta^*)$. The errors $\|A\hat{\theta}_\lambda - b\|_M$ and $\|A\hat{\theta}_{\hat{\lambda}} - b\|_M$ will ultimately be given in terms of $p(\theta^*)$.

2.4.4 Generalizations to other losses

We believe that extending the results to other losses is also possible. Our suggestions in this case are speculative, but we believe they can provide insight about our proof techniques and the potential generalizations.

The general form of the loss we are evaluating is $h(\hat{A}\theta - \hat{b})$ and the loss we are minimizing is $h'(\hat{A}\theta - \hat{b})$, with $h, h' : \mathbb{R} \to \mathbb{R}$. In our case we have used $h(\cdot) = \|\cdot\|_M$ and $h'(\cdot) = \|\cdot\|_M^2$.

We have heavily relied on the fact that triangle inequalities are defined for the ℓ_2-norm, and on the ease of relating it to its square, through linear lower-bounds, or, more explicitly, linear upper-bounds on the square-root function.

A super-additive h should be enough to reproduce the triangle-inequality steps, and h' would have to be chosen taking into consideration its relationship to h, as well as algorithmic aspects, *e.g.*, ease to minimize. A suitable replacement for Lemma 2.1.3 to associate h and h' would be necessary, and this would likely impact the form and proof of all subsequent results.

As in generalizations to other penalty functions $p(\cdot)$, the Δ quantities must be properly redefined so that they can be split into $\Delta p(\theta)$, *e.g.*, $h((\hat{A} - A)\theta) \leq \Delta_A p(\theta)$

Of course, considering that the ultimate purpose of studying $\|A\theta - b\|_M$ is its relationship to other error metrics on specific scenarios, reworking the lemmas for other h and h' would only make sense if $h(A\theta - b)$ was itself meaningful.

2.4.5 Oracle inequalities

It would be interesting to have oracle inequalities, *e.g.*, a bound of the form

$$\|A\hat{\theta}_\lambda - b\|_M \leq 2\Delta_b + \inf_{\theta \in \mathbb{R}^d} \left(2\Delta_A \|\theta\|_1 + \|A\theta - b\|_M\right).$$

These are interesting generalizations of our bounds, and we believe they might not be hard to derive because the techniques we use to relate $\|A\hat{\theta}_\lambda - b\|_M$ to quantities pertaining to θ^* may also, in principle, be used to do the same for quantities relating to any other $\theta \in \mathbb{R}^d$. We have exploited that $\|A\theta^* - b\|_M = 0$ under the assumption that $A\theta^* = b$, so in this case there might be some trouble to obtain bounds that have the form we are after (with $1 \cdot L(\theta)$, w.r.t. to the θ that minimizes the term whose infimum is taken).

Chapter 3

Applications to regression

The theme of this chapter is linear regression. After formulating the problem as a linear estimation problem and stating our results in Section 3.1, we see, in Section 3.2, the consequences of our results derived in Section 2.3, under the light of related work in the literature.

3.1 Formulation

In linear regression, we assume there exist a measurable input space $\mathcal{X}$, a probability measure ρ and a measurable function $g : \mathcal{X} \to \mathbb{R}$ such that $g(X) = \mathbb{E}[Y|X]$, where X and Y are jointly distributed according to ρ.

We will limit ourselves to linear approximation, so we are given a feature extractor $\phi : \mathcal{X} \to \mathbb{R}^d$ and we wish to find an element of

$$\arg\min_{f \in \mathcal{F}} \mathbb{E}\left[(f(X) - Y)^2\right],$$

where

$$\mathcal{F} \doteq \left\{ f_\theta | \theta \in \mathbb{R}^d \right\},$$
$$f_\theta(\cdot) \doteq \phi(\cdot)^\top \theta.$$

That is, we want to find the orthogonal projection of g, with respect to the euclidean norm weighted according to ρ, onto $\mathcal{F}$, the linear space induced by ϕ. Let us call this orthogonal projection $g_\mathcal{F}$.

Let

$$\theta^* \doteq \arg\min_{\theta \in \Theta^*} \|\theta\|_1,$$

where

$$\Theta^* \doteq \{\theta : f_\theta = g_\mathcal{F}\}.$$

Even though $g_\mathcal{F}$ is unique, Θ^* may not be a singleton, so we choose θ^* as its most convenient element for our analysis.

We evaluate the performance of an estimator θ by

$$\mathbb{E}\left[(f_\theta(X) - Y)^2\right] = \theta^\top \mathbb{E}\left[\phi(X)\phi(X)^\top\right]\theta - 2\theta^\top \mathbb{E}\left[\phi(X)Y\right] + \mathbb{E}\left[Y^\top Y\right].$$

Letting

$$A \doteq C \doteq \mathbb{E}\left[\phi(X)\phi(X)^\top\right],$$
$$b \doteq \mathbb{E}\left[\phi(X)Y\right],$$

and provided that $C \succ 0$, we have that

$$\mathbb{E}\left[(f_\theta(X) - Y)^2\right] = \theta^\top C\theta - 2\theta^\top b + \mathbb{E}\left[Y^\top Y\right],$$

and by the definition of θ^*, it holds that

$$\mathbb{E}\left[\phi(X)\phi(X)^\top\right]\theta^* = \mathbb{E}\left[\phi(X)Y\right],$$
$$A\theta^* = C\theta^* = b.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[(f_\theta(X) - Y)^2\right] &= \theta^\top C\theta - 2\theta^\top b + bC^{-1}b + \mathbb{E}\left[Y^\top Y\right] - bC^{-1}b \\
&= \|A\theta - b\|^2_{C^{-1}} + \mathbb{E}\left[Y^\top Y\right] - 2bC^{-1}b + bC^{-1}b \\
&= \|A\theta - b\|^2_{C^{-1}} + \mathbb{E}\left[Y^\top Y\right] - 2b\theta^* + \theta^{*\top} C\theta^* \\
&= \|A\theta - b\|^2_{C^{-1}} + \mathbb{E}\left[(g_{\mathcal{F}}(X) - Y)^2\right].
\end{aligned}$$

The second term is an approximation error that can only be reduced by changing $\mathcal{F}$ appropriately, which is out of the scope of this work.

So, finding θ that minimizes $\|A\theta - b\|^2_{C^{-1}}$ is a sensible way of approximating g, however, we are only provided with a sample (x_t, y_t), for $t = 1, \ldots, n$ drawn w.r.t. the measure ρ. We we will assume that these observations are drawn i.i.d., and, with a few definitions, we will see that this regression problem can be cast as a linear estimation problem to which our results apply, and for this we need to do as indicated in Section 2.4.1: define A_t and b_t, and ensure that Assumption 2.3.4 holds. We will also define corresponding quantities related to C, because they are of interest to our analysis.

For conciseness, let $\varphi_t = \phi(x_t)$ for $t = 1, \ldots, n$. Based on φ_t and y_t, for $t = 1, \ldots, n$, we can construct A_t, b_t and C_t for $t = 1, \ldots, n$ such that

$$A_t \doteq C_t \doteq \varphi_t\varphi_t^\top,$$
$$b_t \doteq \varphi_t y_t.$$

Also, let

$$\hat{C} \doteq \frac{1}{n}\sum_{t=1}^{n} C_t.$$

With this, our definitions for regression are in accordance with Assumptions 2.3.1 and the problem statement in Section 1.1. Even though we would like to use $M = C^{-1}$, C is not known to us, so we will work with $M \succ 0$, and in Section 3.2 we will investigate the consequences of different choices of M. Now it only remains to ensure that Assumption 2.3.4 is satisfied.

Assumption 3.1.1. *Assume that there exist positive constants $F_{2,\infty}, F'_{2,\infty}, F_{\infty,\infty}, R_\infty$ such that, with probability one,*

$$\begin{aligned}
&\|M^{\frac{1}{2}}\varphi_t\|_2 \le F_{2,\infty},\\
&\|\varphi_t\|_2 \le F'_{2,\infty}\\
&\|M^{\frac{1}{2}}\varphi_t\|_\infty \le F_{\infty,\infty}\\
&|y_t| \le R_\infty
\end{aligned}$$

These, along with the definition of A_t and b_t, imply that

$$\|M^{\frac{1}{2}}A_t\|_F = \|M^{\frac{1}{2}}\varphi_t\varphi_t^\top\|_F \le F_{2,\infty}F'_{2,\infty}$$

almost surely.

The statement above can be proved using the formulation of the Frobenius norm as a trace, and then the rotation property of the trace operator. By the Cauchy-Schwartz inequality,

$$\|M^{\frac{1}{2}}b_t\|_F = \sup_{\varphi_t, y_t} \|M^{\frac{1}{2}}\varphi_t y_t\|_2 \le F_{2,\infty}R_\infty$$

almost surely. These constructions allow the results of Section 2.3 to be applied to our regression scenario. Because we have that $A\theta^* = b$, we can use Corollary 2.3.8:

Corollary 2.3.8. *Under Assumptions 2.3.1 and 2.3.4, if $A\theta^* = b$ and*

$$\hat{\lambda} = \frac{2}{n}F_{2,\infty}^2 \max\{R_\infty, F'_{2,\infty}\}^2,$$

then, for any $0 < \delta < 1$, with probability at least $1-\delta$, it holds that

$$\begin{aligned}
\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \le F_{2,\infty}\max\{R_\infty, F'_{2,\infty}\}\sqrt{\frac{2}{n}} \cdot \max\Bigg\{ &(\|\theta^*\|_1+1)^2 T\left(n, \frac{\delta}{2}\right)^3 + \\
&\frac{1}{2}T\left(n,\frac{\delta}{2}\right)(\|\theta^*\|_1+1), \frac{1}{2T\left(n,\frac{\delta}{2}\right)}\frac{\|\theta^*\|_1}{\|\theta^*\|_1+1} + T\left(n,\frac{\delta}{2}\right)\Bigg\}\\
&+ F_{2,\infty}\max\{R_\infty, F'_{2,\infty}\}\sqrt{\frac{2}{n}}T\left(n,\frac{\delta}{2}\right)(\|\theta^*\|_1+1),
\end{aligned}$$

where T is as in Definition 2.3.3.

This bound is uniform, and has the structure we were aiming for:

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \le O\left((\|\theta^*\|_1+1)^2\sqrt{\frac{1}{n}}\right).$$

In the next section we interpret this result, and study its implications to regression, for different choices of M.

3.2 Consequences and related work

The first evident consequence of Corollary 2.3.8 is that if we choose $M = C^{-1}$, Corollary 2.3.8 and $A\theta^* = b$ imply that, in any regression problem satisfying our conditions,

$$\|A\hat{\theta}_\lambda - b\|_M^2 = O\left(\frac{1}{n}\right)$$

with high probability. Since $\hat{\theta}_\lambda$ is random, in order to analyze $\mathbb{E}\left[(f_{\hat{\theta}_\lambda}(X) - g(X))^2\right]$ we need to take an expectation bound of $\|A\hat{\theta}_\lambda - b\|_M^2$. By tail integration of the bound in Corollary 2.3.8, we can see that

$$\mathbb{E}\left[(f_{\hat{\theta}_\lambda}(X) - g(X))^2\right] = \mathbb{E}\left[(g(X) - g_{\mathcal{F}}(X))^2\right] + O\left(\frac{\ln n}{n}\right).$$

The result of Corollary 2.3.8 with $M = C^{-1}$ is similar to Theorem 11.3 of Györfi et al. (2002), which reads:

Theorem 3.2.1 (Theorem 11.3 of Györfi et al. (2002))**.** *Assume*

$$\sigma^2 = \sup_{x \in \mathcal{X}} \mathrm{Var}(Y|X = x) < \infty$$

and

$$\|g\|_\infty = \sup_{x \in \mathcal{X}} |g(x)| \leq L$$

for some $L \geq 0$. Let $\mathcal{F}_n$ be a linear vector space of functions $f : \mathbb{R}^d \to \mathbb{R}$. Let K_n be the vector space dimension of $\mathcal{F}_n$. Define the estimate m_n by

$$m_n(\cdot) = \max\left\{-L, \min\left\{\tilde{m}_n(\cdot), L\right\}\right\},$$
$$\tilde{m}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} (f(\phi(x_t)) - y_t)^2.$$

Then

$$\mathbb{E}\left[(m_n(\phi(X)) - g(X))^2\right] \leq c \max\left\{L^2, \sigma^2\right\} \frac{K_n \ln n}{n} + 8 \inf_{f \in \mathcal{F}_n} \mathbb{E}\left[(f(\phi(X)) - g(X))^2\right].$$

The assumptions in Theorem 3.2.1 are more general than ours, because the responses Y can be unbounded, provided that they have a bounded second moment, and the input X can also be unbounded as long as $g(X)$ is bounded almost surely. There is an explicit dependence on the dimension K_n of $\mathcal{F}_n$, whereas in our case this dependence is implicit in the range constants.

Now, truncation is an important concept used in the theorem to dismiss boundedness assumptions on X, and it would be convenient for us to use it as well. Unfortunately, it is not clear how to define a suitable notion of truncation for our problem, and one that is compatible with the concentration inequalities we use.

Of course, only an oracle can choose $M = C^{-1}$. For any θ, loose bounds for $\|A\theta - b\|_{C^{-1}}$ can be obtained by taking $M = \mathbf{I}_d$ and then using the fact that

$$\begin{aligned}\|A\theta - b\|_{C^{-1}} &\leq \sqrt{\lambda_{\max}(C^{-1})}\|A\theta - b\|_2 \\ &= \sqrt{\frac{1}{\lambda_{\min}(C)}}\|A\theta - b\|_2.\end{aligned}$$

Remark 3.2.2 (about the looseness of bounds that use minimum eigenvalues). We know that $\lambda_{\min}(C)$, the minimum eigenvalue of C, is strictly positive, but it can be arbitrarily close to zero. The resulting bounds are loose in the sense that there may be a large, though constant, gap between $\sqrt{\frac{1}{\lambda_{\min}(C)}}\sup_{\varphi_t}\|\varphi_t\|_2$ and $\sup_{\varphi_t}\|\varphi_t\|_M$.

We believe there are other approaches that are worth investigating, with a potential to better relate $\|A\theta - b\|_{C^{-1}}$ and $\|A\theta - b\|_M$. We would need to estimate C^{-1}; it is not as straightforward as inverting an estimate or a shifted estimate of C, because doing so yields a biased estimator of C^{-1}.

Remark 3.2.3 (about using estimates of C^{-1} and other random matrices as M). We may simplify [1] our problem by using an estimate $\hat{C}''$ of C that is independent of $\hat{A}$, $\hat{b}$, $\hat{A}'$ and $\hat{b}'$, and then show, for some $D \succ 0$, that $\mathbb{E}\left[(\hat{C}'' + D)^{-1}\right] \approx C^{-1}$.

Alternatively, one might demonstrate that for fixed $v \in \mathbb{R}^d$ $\|v\|_{\hat{C}''^{+}}$ concentrates around its expectation, by working on the subspace where the Moore-penrose pseudo-inverse function ($^+$) is convex, and use Jensen's inequality to see that

$$\mathbb{E}\left[\|v\|_{\hat{C}''^{+}}\right] \leq \|v\|_{C^{-1}}.$$

This approach is particularly suitable for semi-supervised learning scenarios, where we can obtain many observations of the covariance matrix C. In supervised learning, on the other hand, this would mean separating part of the sample for estimation of C and discarding the corresponding responses.

In fact, using a random M requires many details to be dealt with, *e.g.*, dependencies, concentration inequalities and the range assumptions. These, in particular, must hold w.p. 1, so one may need to re-express them in terms of non-random quantities, or use appropriate conditioning throughout the derivations.

Section 6.12 of Bühlmann and Geer (2011) discusses compatibility conditions that may be used to relate $\hat{C}^{+}$, or, more generally, some random $\hat{M}$, to C^{-1}. A compatibility condition (with parameter c) holds when $\|\hat{M} - C^{-1}\|_{\infty,\infty} \leq c'$, so that, with a few other assumptions on the sparsity of $\hat{\theta}_{\hat{\lambda}}$, $\left|\frac{\|A\hat{\theta}_{\hat{\lambda}} - b\|_{\hat{M}}}{\|A\hat{\theta}_{\hat{\lambda}} - b\|_{C^{-1}}} - 1\right| = O\left(\frac{c'}{c^2}\right)$. Note that we can recover a compatibility condition by having a bound based on $\lambda_{\min}(C)$, as we did before to have performance bounds for $M = \mathbf{I}_d$.

Hsu et al. (2011) analyze $\mathbb{E}\left[f_{\hat{\theta}}(X) - g_{\mathcal{F}}(X))^2\right]$ where $\hat{\theta}$ is a least-squares estimator. In Lemma 3, they exploit the structure of $\hat{\theta}$ to bound

$$\|\hat{\theta} - \theta^*\|_C^2 \leq \|C^{\frac{1}{2}}\hat{C}^{-1}C^{\frac{1}{2}}\|_F\mathbb{E}\left[g(X) - g_{\mathcal{F}}(X))^2\right],$$

[1] We chose to simplify the problem by using an independent estimate of C because using $\hat{C}$ or $\hat{C}'$ introduces the necessity of dealing with dependencies.

assuming that $\hat{C} \succ 0$. It may be possible to devise a similar bound for $\hat{\theta}_{\hat{\lambda}'}$, or to use a similar type of separation to obtain a more suitable definition of A and b for regression, that dismisses our need to choose M close to C^{-1}.

Finally, it is important to mention the work of Bartlett et al. (2009). They study the performance of a Lasso (ℓ_1-regularizer least-squares) estimator, and how to choose λ so as to bound $\mathbb{E}\left[f_{\hat{\theta}}(X) - g_{\mathcal{F}}(X))^2\right]$. In fact, they provide oracle inequalities, but they choose λ based on a choice of δ, and require n to be large enough for the bounds to hold, *i.e.*, the smaller the confidence parameter δ, the larger the smallest n to which the respective bound applies. Our results for regression are not oracle inequalities, but they are uniform in δ.

As for the dependence on dimensionality, the bounds by Bartlett et al. (2009) have a factor of $\sqrt{\ln d}$, the impact of dimensionality on our results is more indirect, as discussed ahead.

Remark 3.2.4 (about the impact of d on the bounds). All the corollaries derived in this section do not have an explicit dependence on the number of features d. Evidently, the dimensionality will have its impact on $F_{2,\infty}$ and $F'_{2,\infty}$, so the behavior of these quantities will ultimately dictate how large d can be. For example, generating d Haar wavelet (Wasserman (2006); Sweldens and Schröder (1996)) features would maintain $F_{2,\infty}, F'_{2,\infty} = O(\ln d)$. For certain combinations of input spaces and feature extractors, however, $F_{2,\infty}$ and $F'_{2,\infty}$ can be as large as $\Theta(\sqrt{d})$ [2], so the choice of d is also dependent on the choice.

[2] Regret lower bounds in online learning, such as those in Gerchinovitz and Yu (2011), can be used to help devise such combination.

Chapter 4

Applications to reinforcement learning

In this chapter we extend our lemmas into results for on- and off-policy value prediction in reinforcement learning (RL). First, in Section 4.1, we present the necessary RL context to pose the value prediction problem, which we cast it as a linear estimation problem to which our results apply. Section 4.2 is about the consequences of the results in Section 2.3 for the value prediction problem in RL. Finally, in Section 4.3, we state the related results existing in the literature.

4.1 Formulation

The formulation in this section is derived from Chapter 1 in the book of Szepesvári (2010), our notation being slightly different at parts.

A Markov decision process (MDP) $\mathcal{M}$ is a quintuple $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. $\mathcal{X}$ is a (non-empty) state space and $\mathcal{A}$ is a (non-empty) action space. $\mathcal{P}$ is a transition probability kernel assigning a probability measure $P(\cdot|X, E) : \mathcal{X} \to \mathbb{R}$ to each element of the state-action space $\mathcal{X} \times \mathcal{A}$ such that the next state X' w.r.t. to a state action pair (x, a) is distributed according to $P(\cdot|X = x, E = a)$. The immediate reward kernel $\mathcal{R}$ induces an immediate reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that $\mathbb{E}[r(X, E)|X = x, E = a]$ is the expected immediate reward received when in action a is taken in state x. $0 \leq \gamma < 1$ is a discount factor.

When $\gamma = 0$, we recover a regression scenario with the notion of states, which is more general than the linear regression covered in Chapter 3.

A *deterministic policy* $\pi : \mathcal{X} \to \mathcal{A}$ maps states into actions, and induces a value function $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that, for every $x \in \mathcal{X}, a \in \mathcal{A}$,

$$Q^\pi(x, a) = \mathcal{T}^\pi Q^\pi(x, a),$$

where the Bellman operator $\mathcal{T}^\pi$ is such that

$$\mathcal{T}^\pi f(x, a) = \mathbb{E}\left[r(X, E) + \gamma f(X', \pi(X'))|X = x, E = a\right]$$

holds for any $f : \mathcal{X} \times \mathcal{A} \to \mathcal{R}$, for all $x \in \mathcal{X}, a \in \mathcal{A}$. In the above definition and in the remainder of this chapter, X' denotes a next-state random variable with distribution $P(\cdot|X, E)$.

We will tackle the (potentially) uncountably large size of the state-action space by representing the state-action pairs through *finite* d-dimensional feature vectors, obtained through a feature extractor $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$.

One goal in value prediction is to compute a good approximation of Q^π that is a linear function of the features. We wish to find $\Pi_{\mathcal{F},\rho} Q^\pi$, where

$$\begin{aligned}\Pi_{\mathcal{F},\rho} f' &\doteq \arg\inf_{f\in\mathcal{F}} \mathbb{E}\left[(f(X,E) - f'(X,E))^2\right],\\ \mathcal{F} &\doteq \left\{Q_\theta : \theta \in \mathbb{R}^d\right\},\\ Q_\theta(x,a) &\doteq \phi(x,a)^\top \theta.\end{aligned}$$

This corresponds to finding Q_θ for (any) $\theta \in \Theta^*$, where

$$\Theta^* \doteq \{\theta : Q_\theta = \Pi_{\mathcal{F},\rho} Q^\pi\}.$$

For convenience, let

$$\theta^* \doteq \arg\min_{\theta\in\Theta^*} \|\theta\|_1.$$

Caveat lector: *The function Q_{θ^*} is not to be mistaken for Q^*, which is such that*

$$Q^*(x,a) = \mathbb{E}\left[r(X,E) + \gamma \sup_{a'} Q^*(X',a'))|X = x, E = a\right],$$

and induces at least one deterministic policy π^* such that $Q^{\pi^*} = Q^*$ (Sutton and Barto (1998); Szepesvári (2010)). Rather, Q_{θ^*} is simply the orthogonal projection of Q^π upon $\mathcal{F}$. For any two functions $f, f' : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, let

$$\|f - f'\|_\rho^2 \doteq \mathbb{E}\left[(f(X,E) - f'(X,E))^2\right],$$

where $(X,E) \sim \rho$. For any $\theta \in \mathbb{R}^d$, we can write

$$\begin{aligned}\mathbb{E}\left[(Q^\pi(X,E) - Q_\theta(X,E))^2\right] &= \|Q^\pi - Q_\theta\|_\rho^2\\ &= \|Q^\pi - \Pi_{\mathcal{F},\rho} Q^\pi\|_\rho^2 + \|\Pi_{\mathcal{F},\rho} Q^\pi - Q_\theta\|_\rho^2.\end{aligned}$$

The only way to reduce $\|Q^\pi - \Pi_{\mathcal{F},\rho} Q^\pi\|_\rho^2$ is to redefine $\mathcal{F}$. This is a problem on its own, which we will not address, so we will restrict ourselves to finding a θ that minimizes $\|\Pi_{\mathcal{F},\rho} Q^\pi - Q_\theta\|_\rho^2$. We have that

$$\begin{aligned}\|\Pi_{\mathcal{F},\rho} Q^\pi - Q_\theta\|_\rho &= \|\Pi_{\mathcal{F},\rho}(\mathcal{T}^\pi Q^\pi - Q_\theta)\|_\rho\\ &\le \|\Pi_{\mathcal{F},\rho}(\mathcal{T}^\pi Q^\pi - \mathcal{T}^\pi Q_\theta)\|_\rho + \|\Pi_{\mathcal{F},\rho}(\mathcal{T}^\pi Q_\theta - Q_\theta)\|_\rho.\end{aligned}$$

The term $\|\Pi_{\mathcal{F},\rho}(\mathcal{T}^\pi Q_\theta - Q_\theta)\|_\rho$ is called the projected Bellman error (PBE, Antos et al. (2008)). By applying the definitions of $\Pi_{\mathcal{F},\rho}$ and $\mathcal{T}^\pi$, it is easy to show that

$$\|\Pi_{\mathcal{F},\rho}(\mathcal{T}^\pi Q_\theta - Q_\theta)\|_\rho^2 = \|A\theta - b\|_{C^{-1}}^2,$$

where

$$\begin{aligned}A &\doteq \mathbb{E}\left[\phi(X,E)(\phi(X,E) - \gamma\phi(X',\pi(X')))^\top\right],\\ b &\doteq \mathbb{E}\left[\phi(X,E) r(X,E)\right],\\ C &\doteq \mathbb{E}\left[\phi(X,E)\phi(X,E)^\top\right],\end{aligned}$$

provided that $C \succ 0$, which we assume to be the case.

It is possible to approximate Q_{θ^*} through the least-squares temporal difference learning (LSTD) Bradtke and Barto (1996); Szepesvári (2010). In particular, when $A\theta^* = b$ it yields the same set of minimizers as the PBE, because

$$A\theta^* = b \iff \|A\theta^* - b\|_2^2 = 0 \iff \|A\theta^* - b\|_{C^{-1}}^2 = 0.$$

In our case, we have very little knowledge about ρ, or we want to make only a few assumptions about it. That is to say that we do not have access to A, b and C directly, but we are given transition observations $((X_t, E_t), r_t, X_{t+1})$, for $t = 1, \dots, n+m$, where (X_t, E_t) are drawn from ρ, and r_t and X_{t+1} are drawn from the reward and next-state distributions associated to (X_t, E_t). We then construct a sample $(\varphi_t, r_t, \varphi'_{t+1})$ for $t = 1, \dots, n+m$ where $\varphi_t = \phi(X_t, E_t)$ and $\varphi_t = \phi(X_{t+1}, \pi(X_{t+1}))$. The distribution ρ could be, for example, induced by running a policy π' in the MDP. π', often referred to as the behavior policy, may be different from the evaluation policy π (Sutton and Barto (1998)). When $\pi = \pi'$, the policy evaluation is called on-policy, otherwise it is called off-policy.

In order to formulate the value prediction problem as a linear estimation problem, we need to define E_t and b_t, and ensure that Assumptions 2.3.4 hold, as described in Section 2.4.1, so that the results in Section 2.3 apply. First of all, assume that the observations are drawn i.i.d. from ρ, and let

$$\begin{aligned} A_t &\doteq \varphi_t(\varphi_t - \gamma\varphi'_{t+1})^\top, \\ b_t &\doteq \varphi_t r_t. \end{aligned}$$

Because we will need to make considerations about C, let

$$\begin{aligned} C_t &\doteq \varphi_t \varphi_t^\top, \\ \hat{C} &\doteq \frac{1}{n}\sum_{t=1}^{n} C_t. \end{aligned}$$

Furthermore,

Assumption 4.1.1. *Assume that there exist positive constants $F_{2,\infty}, F'_{2,\infty}, F'_{\infty,\infty}, R_\infty$ such that, with probability one,*

$$\begin{aligned} &\|M^{\frac{1}{2}}\varphi_t\|_2 \le F_{2,\infty}, \\ &\|\varphi_t - \gamma\varphi'_{t+1}\|_2 \doteq F'_{2,\infty}, \\ &\|\varphi_t - \gamma\varphi'_{t+1}\|_\infty \doteq F'_{\infty,\infty}, \\ &|r_t| \doteq R_\infty. \end{aligned}$$

As observed in Chapter 3 after Assumption 3.1.1, Assumption 4.1.1 implies that, with probability one,

$$\|M^{\frac{1}{2}}A_t\|_F = \|M^{\frac{1}{2}}\varphi_t(\varphi_t - \gamma\varphi'_{t+1})^\top\|_F \le F_{2,\infty}F'_{2,\infty}.$$

and likewise,

$$\|M^{\frac{1}{2}}b_t\|_F = \|M^{\frac{1}{2}}\varphi_t r_t\|_2 \le F_{2,\infty}R_\infty.$$

All corollaries in Section 2.3 apply to this scenario, but we will only restate Corollaries 2.3.8, 2.2.3 and 2.2.4, which are of particular interest. The result of Corollary 2.3.9 can be safely replaced by that of Corollary 2.2.4.

The first result applies to the on-policy case, in which $\|A\theta^* - b\|_{C^{-1}} = 0$:

Corollary 2.3.8. *Under Assumptions 2.3.1 and 2.3.4, if $A\theta^* = b$ and*

$$\hat{\lambda} = \frac{2}{n} F_{2,\infty}^2 \max\{R_\infty, F'_{2,\infty}\}^2,$$

then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, it holds that

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq F_{2,\infty} \max\{R_\infty, F'_{2,\infty}\} \sqrt{\frac{2}{n}} \cdot \max \Bigg\{ &(\|\theta^*\|_1 + 1)^2 T\left(n, \frac{\delta}{2}\right)^3 + \\ &\frac{1}{2} T\left(n, \frac{\delta}{2}\right)(\|\theta^*\|_1 + 1), \frac{1}{2T\left(n, \frac{\delta}{2}\right)} \frac{\|\theta^*\|_1}{\|\theta^*\|_1 + 1} + T\left(n, \frac{\delta}{2}\right) \Bigg\} \\ &+ F_{2,\infty} \max\{R_\infty, F'_{2,\infty}\} \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right)(\|\theta^*\|_1 + 1),\end{aligned}$$

where T is as in Definition 2.3.3.

Corollary 2.3.8 allows us to provide a bound for the projected Bellman error in on-policy scenarios. In this case, with the (unrealistic) choice of $M = C^{-1}$, we obtain, for any $0 < \delta < 1$,

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_{C^{-1}}^2 = O\left(\frac{(\|\theta^*\|_1 + 1)^4}{n}\right)$$

holds with probability at least $1 - \delta$. If we choose $M = \mathbf{I}$ and use the technique illustrated in Section 3.2, we can show that for any $0 < \delta < 1$,

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_{C^{-1}}^2 = O\left(\frac{(\|\theta^*\|_1 + 1)^4}{n\lambda_{\min}(C)}\right)$$

holds with probability at least $1 - \delta$. Note that the effect of the dimensionality is embedded in $F_{2,\infty}$, R_∞ and $F'_{2,\infty}$ (as discussed in Remark 3.2.4, which also applies to this case).

The next two results hold for both on- and off-policy scenarios:

Corollary 2.3.10. *Under Assumptions 2.3.1 and 2.3.4, for any $0 < \delta < 1$, if*

$$\begin{aligned}\lambda' &= F_{2,\infty} F'_{2,\infty} \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right), \\ a &= 4\sqrt{\frac{1}{n}}\lambda', \\ B &= F_{2,\infty}^2 F'_{2,\infty} R_\infty, \\ \hat{\lambda} &= \arg \min_{\lambda \in \Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda' \|\hat{\theta}_\lambda\|_1,\end{aligned}$$

then, for n large enough so that $a \leq B$, with probability at least $1-\delta$, it holds that

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq 3F_{2,\infty}F'_{2,\infty}\sqrt{\frac{2}{n}}T\left(n, \frac{\delta}{2}\right)\|\theta^*\|_1 + 2F_{2,\infty}R_\infty\sqrt{\frac{2}{n}}T\left(n, \frac{\delta}{2}\right) + \|A\theta^* - b\|_M + \sqrt{\frac{1}{n}}.$$

Corollary 2.3.11. *Under Assumptions 2.3.1 and 2.3.4, for any $0 < \delta < 1$, if*

$$\begin{aligned}
\lambda' &= F_{2,\infty}F'_{2,\infty}\sqrt{\frac{2}{n}}T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right), \\
a &= 4\sqrt{\frac{1}{n}}\lambda', \\
B &= F^2_{2,\infty}F'_{2,\infty}R_\infty, \\
\hat{\lambda} &= \arg\min_{\lambda \in \Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda'\|\hat{\theta}_\lambda\|_1,
\end{aligned}$$

then it holds that

$$\begin{aligned}
\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}&\left(\frac{b^2_{\max}}{2F_{2,\infty}F'_{2,\infty}\sqrt{2}T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right)}\right)\sqrt{\frac{1}{n}} \\
&+ 2\sqrt{\frac{1}{n}}F_{2,\infty}F'_{2,\infty}\sqrt{2}T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right)\|\theta^*\|_1 \\
&+ \mathbb{E}\left[\Delta_A\right]\|\theta^*\|_1 + 2\mathbb{E}\left[\Delta_b\right] + \|A\theta^* - b\|_M + \sqrt{\frac{1}{n}}.
\end{aligned}$$

We can see that choosing $M = C^{-1}$, we obtain, for $0 < \delta < 1$ used to choose λ',

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|^2_{C^{-1}} = O\left(\frac{(\|\theta^*\|_1 + 1)}{n}\right) + \left(1 + O\left(\sqrt{\frac{1}{n}}\right)\right)\|A\theta^* - b\|^2_{C^{-1}}$$

holds with probability at least $1-\delta$. If we choose $M = \mathbf{I}$ and use the technique illustrated in Section 3.2, we can show that for $0 < \delta < 1$ used to choose λ',

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|^2_{C^{-1}} = O\left(\frac{(\|\theta^*\|_1 + 1)}{n\lambda_{\min}(C)}\right) + \left(\frac{\lambda_{\max}(C)}{\lambda_{\min}(C)} + O\left(\sqrt{\frac{\lambda_{\max}(C)}{n\lambda_{\min}(C)}}\right)\right)\|A\theta^* - b\|^2_{C^{-1}}$$

holds with probability at least $1-\delta$. These bounds have a "slow" term *and* the conditioning number of C multiplying the projected Bellman error of θ^*, which is not a desirable form for the result. However, the unsquared projected Bellman error has the form we would like, *i.e.*,

$$O\left(\sqrt{\frac{1}{n}}\right) + \|A\theta^* - b\|_{C^{-1}}$$

with the exception of the inverse smallest eigenvalue of C multiplying the excess loss, and the conditioning number of C multiplying $\|A\theta^* - b\|_{C^{-1}}$ if we use the eigenvalue upper-bounding.

The remarks about the expectation bound of Corollary 2.3.11 are analogous to those about the results of Corollary 2.3.10.

Now, we are ready to discuss the implications of these corollaries to reinforcement learning.

4.2 Consequences

In the previous section, we have argued why minimizing the PBE is important, and that we wish to minimize it as a means of minimizing $\|Q^\pi - Q_\theta\|_\rho$. In this section, we will see what our results can tell us about the PBE of certain estimators. The discussion will be limited to $\hat{\theta}_{\hat{\lambda}'}$, as the observations follow similarly for the other estimators we have studied.

As in the regression scenario, taking M as C^{-1} yields bounds for the PBE. However, this is an oracle choice that cannot be made in practice. One alternative is to do as prescribed in Section 3.2: choose $M = \mathbf{I}_d$, which corresponds to performing ℓ_1-regularized LSTD, and then use that

$$\|A\theta - b\|_{C^{-1}} \leq \sqrt{\frac{1}{\lambda_{\min}(C)}}\|A\theta - b\|_2.$$

Since it may be the case that $\|A\theta^* - b\|_2 > 0$, we will also need to use that

$$\|A\theta - b\|_2 \leq \sqrt{\lambda_{\max}(C)}\|A\theta - b\|_{C^{-1}},$$

but the resulting bounds may be loose, for the reason expressed in Remark 3.2.2.

It is natural to ask whether θ minimizing $\|A\theta - b\|_M$ for other M will enjoy better estimation error than those computed through LSTD. This question is motivated by the only other rates of convergence for LSTD, given by Lazaric et al. (2010), and the claim in Antos et al. (2008) that there are more suitable losses to minimize than $\|\hat{A}\theta - \hat{b}\|_2^2$.

In our analysis, the choice $M = C^{-1}$, though impractical, yields better estimators than $M = \mathbf{I}_d$. This hints at a positive answer for the question stated before, but it is not definite because our results apply to ℓ_1-regularized LSTD, and in principle we have not proved that bounding $\|A\theta - b\|_{C^{-1}}^2$ for LSTD estimators will necessarily require a factor of $\frac{1}{\lambda_{\min}(C)}$ to appear in the bound. Two alternatives to performing LSTD are to use $M = \hat{C}^{-1}$ or to have $M = \hat{C}'^{-1}$ (where $\hat{C}'$ is independent of all the other random variables).

Remark 3.2.3 is also valid for this RL case, but there is an additional issue: we may need to upper-bound $\|A\theta^* - b\|_M$ by $\|A\theta^* - b\|_{C^{-1}}$ and some other vanishing terms. This is not a problem in regression because in that case $\|A\theta^* - b\|_{C^{-1}} = 0$, but in off-policy value prediction it may be the case that $\|A\theta^* - b\|_{C^{-1}} > 0$.

One possibility is to introduce the factor of $\sqrt{\lambda_{\max}(C)}$, which is cannot be much larger than the other quantities we are using. Another possibility is to generalize the results of Pittenger (1990) to matrices; the important result is stated in this text as Lemma B.2.1 and applies to scalars. The idea would be to have some bound of $\|A\theta^* - b\|_{(\hat{C}'+D)^{-1}}$, where $D \succ 0$, in terms of $\|A\theta^* - b\|_{C^{-1}}$ plus some vanishing term.

The role of dimensionality in this scenario is the same as in regression, as expressed in Remark 3.2.4, with the exception that $F'_{2,\infty}$ may be different in RL.

4.3 Related work

In this section, we delve into related studies about value function approximation in RL. The existing finite-time performance bounds are recent, and they also pertain to minimizing $\|Q^\pi - Q_\theta\|_\rho$, but as we will see our approach is somewhat different from the the other existing ones.

The goal of this section is to show the bounds proved and the types of estimators studied elsewhere. This exposition will be enough to contrast our results and our approach with those in this subfield of RL.

The works of Lazaric et al. (2010); Ghavamzadeh et al. (2010, 2011) provide risk-like bounds for variations of LSTD. It is important to emphasize that the primary quantities they investigate are different from the ones we do, but there is a common ground where comparison of results is possible, where techniques they use may help us expand our results, and vice-versa. In particular, they are also interested in bounding $\|Q^\pi - Q_{\hat{\theta}}\|_\rho^2$ and

$$\|Q^\pi - Q_{\hat{\theta}}\|_n^2 \doteq \frac{1}{n}\sum_{t=1}^{n}(Q^\pi(X_t, E_t) - Q_{\hat{\theta}}(X_t, E_t))^2,$$

with $\hat{\theta}$ being computed differently for each study, by the respective variations of LSTD. The data used to compute the estimators is a Markov chain

$$\mathcal{S}_n = (X_1, r_1, X_2), \ldots, (X_n, r_n, X_{n+1}),$$

where r_t is the reward at time-step t, and ρ is a stationary distribution given by a β-mixing process. Ghavamzadeh et al. (2011), sparsity oracle inequalities are provided, but not the generalization bounds. We will refrain from discussing the sparsity oracle inequalities because parameter identification is out of the scope of this work; see, for example, Candes and Tao (2007) for interesting results and additional references. We will also omit the discussion of the generalization bounds, but, next, we will present in detail the other results.

Lazaric et al. (2010) use the empirical projection operator

$$\widehat{\Pi}_{\mathcal{F},n} f' \doteq \arg\min_{f\in\mathcal{F}} \|f' - f\|_n,$$

and define the pathwise Bellman $\widehat{\mathcal{T}}^\pi$ operator

$$\widehat{\mathcal{T}}^\pi f(X_t) \doteq \begin{cases} r_t + \gamma f(X_{t+1}) \text{ , if } 1 \leq t < n, \\ r_t \text{ , if } t = n, \end{cases}$$

in order to compute the pathwise LSTD

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \|f - \widehat{\Pi}_{\mathcal{F},n}\widehat{\mathcal{T}}^\pi f\|_2^2.$$

They prove that $\widehat{\Pi}_{\mathcal{F},n}\widehat{\mathcal{T}}^\pi$ is a γ-contraction w.r.t. to $\|\cdot\|_n$, so Banach's fixed-point theorem implies that $f = \widehat{\Pi}_{\mathcal{F},n}\widehat{\mathcal{T}}^\pi f$ for some f. By the definition of pathwise LSTD, $\hat{f}$ is that fixed-point.

They use these results to show that

$$\|Q^\pi - Q_{\hat{\theta}}\|_n \leq \frac{1}{1-\gamma}\left(\|Q^\pi - \widehat{\Pi}_{\mathcal{F},n}Q^\pi\|_n + \|\widehat{\Pi}_{\mathcal{F},n}(Q^\pi - \widehat{\mathcal{T}}^\pi Q^\pi)\|_n\right),$$

and relate this bound to $\|Q^\pi - Q_{\hat{\theta}}\|_\rho$ by applying concentration inequalities for their β-mixing process.

In addition, they show that, with probability at least $1-\delta$,

$$\|\widehat{\Pi}_{\mathcal{F},n}(Q^\pi - \widehat{\mathcal{T}}^\pi Q^\pi)\|_n = O\left(\sqrt{\frac{d\ln\frac{d}{\delta}}{\nu_n n}}\right),$$

where ν_n is the minimum *strictly positive* eigenvalue of $\hat{C}$. We must point out that $\|\widehat{\Pi}_{\mathcal{F},n}(Q^\pi - \widehat{\mathcal{T}}^\pi Q^\pi)\|_n$ is very much different from $\|A\theta^* - b\|_{C^{-1}}$. While the latter quantity can be larger than zero,

$$\begin{aligned}\lim_{n\to\infty}\|\widehat{\Pi}_{\mathcal{F},n}(Q^\pi - \widehat{\mathcal{T}}^\pi Q^\pi)\|_n &= \|\Pi_{\mathcal{F},\rho}(Q^\pi - \mathcal{T}^\pi Q^\pi)\|_\rho \\ &= 0,\end{aligned}$$

because $Q^\pi = \mathcal{T}^\pi Q^\pi$.

Even though ν_n is strictly positive, it can be arbitrarily close to zero, so the authors assume n is large enough so that ν_n is larger than some quantity with probability at least $1-\delta$. However, there is a caveat in doing this: for any fixed n there exists a minimum strictly positive confidence with which the inequalities hold, and in this sense they are not uniform in δ. We believe it is possible to use regularization to guarantee that ν_n is large enough for any $\delta > 0$. For example, performing the projection with ℓ_2 regularization weighed by $\frac{\beta}{2}$ would increase ν_n by β.

Pathwise LSTD is only defined for on-policy scenarios. Although extending the results to state-action pairs is straightforward, it is not clear how $\widehat{\mathcal{T}}^\pi$ can be redefined to accommodate multiple chains, or an off-policy sample (which would be the case of multiple length-1 chains), in a way that it is still a contraction, in particular because they exploit that the observations are chained together in their original proof.

Ghavamzadeh et al. (2010) study pathwise LSTD combined with random projections (Dasgupta (2000)) to perform dimensionality reduction. They show that, with probability at least $1-\delta$,

$$\begin{aligned}\|Q^\pi - \widehat{\Pi}_{\mathcal{G},n}Q^\pi\|_n \leq \frac{1}{\sqrt{1-\gamma^2}}&\left(\|v - \widehat{\Pi}_{\mathcal{F},n}Q^\pi\|_n + O\left(\sqrt{\frac{\ln\frac{n}{\delta}}{d}}m\left(\widehat{\Pi}_{\mathcal{F},n}Q^\pi\right)\right)\right) \\ &+ O\left(\sqrt{\frac{d}{\nu_n}}\left(\sqrt{\frac{\ln\frac{d}{\delta}}{n}} + \frac{1}{n}\right)\right),\end{aligned}$$

where $\mathcal{G}$ is the low-dimensional space, of dimension $d \geq 15\ln\frac{8n}{\delta}$, and $\mathcal{F}$ is the high-dimensional space, with dimension D. ν_n is defined w.r.t. the covariance matrix in the smaller space, and its appearance is a consequence of the result in Lazaric et al. (2010). The need to have $\nu_n > c > 0$, for some constant c, and $d \geq 15\ln\frac{8n}{\delta}$ imply

that the bound is non-uniform in δ, because d has to be chosen based on δ, and if δ is sufficiently small then $15 \ln \frac{8n}{\delta} > D$.

The term $m(Q_\theta)$, defined as [1]

$$m(Q_\theta) = \|\theta\|_2 \sup_{\varphi_t} \|\varphi_t\|_2,$$

is influenced by D and by the way the features are constructed (cf. Remark 3.2.4). It may be $O(\log D)$ for some feature extractors, but $O\left(\sqrt{D}\right)$ for others.

The authors show that with

$$d = O\left(m\left(\widehat{\Pi}_{\mathcal{F},n} Q^\pi\right) \sqrt{n\nu_n}\right),$$

we have

$$\|Q^\pi - \widehat{\Pi}_{\mathcal{G},n} Q^\pi\|_n = O\left(\sqrt{\ln n}\left(\frac{1}{n\nu_n}\right)^{\frac{1}{4}}\right),$$

where the we have hidden the confidence parameter δ of the bound, range constants, γ, and $m\left(\widehat{\Pi}_{\mathcal{F},n} Q^\pi\right)$. We point out that by choosing

$$d = \Theta\left(m\left(\widehat{\Pi}_{\mathcal{F},n} Q^\pi\right) \nu_n \sqrt{n}\right),$$

we can eliminate the effect of $\left(\frac{1}{\nu_n}\right)^{\frac{1}{4}}$, which can be arbitrarily large if δ is small enough. There will be a factor of $\sqrt{\nu_n}$ multiplying one of the terms, but this factor cannot be arbitrarily large.

The real caveat is that we must ensure that $d \geq 15 \ln \frac{8n}{\delta}$, but either of the choices may not respect that, if ν_n is too small. Moreover, computing this eigenvalue will affect the cost of the whole procedure, as will computing $m\left(\widehat{\Pi}_{\mathcal{F},n} Q^\pi\right)$, because of D.

This error rate is slower than that of pathwise LSTD in terms of n, but if D is large enough, even $D = \sqrt{n}$, the random projections still allow good performance to be proved, whereas the bounds for plain pathwise LSTD become vacuous.

In Ghavamzadeh et al. (2011), pathwise LSTD is combined with ℓ_1-norm regularization, which called Lasso-TD. They define

$$\widehat{\Pi}_{\mathcal{F},\rho,\lambda} f = f_{\hat{\theta}} \ : \ \hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|f_\theta - f\|_n^2 + \lambda \|\theta\|_1,$$

where $f_\theta(\cdot) = \phi(\cdot)^\top \theta$. This operator is shown to be a non-expansion, so $\widehat{\Pi}_{\mathcal{F},\rho,\lambda} \widehat{\mathcal{T}}^\pi$ is a γ-contraction w.r.t. to $\|\cdot\|_n$, and the authors claim that finding its fixed point is equivalent to computing the fixed point of

$$\hat{\theta} = \arg\min_\theta \|\widehat{\Pi}_{\mathcal{F},\rho,\lambda} \widehat{\mathcal{T}}^\pi f_{\hat{\theta}} - f_\theta\|_n^2.$$

They also claim that this is equivalent to what LARS-TD Kolter and Ng (2009) and LC-TD Johns et al. (2010) compute. That is to say that these two methods and

[1] If there is more than one θ yielding Q_θ, choose the one with smallest ℓ_2-norm for $m(Q_\theta)$.

Lasso-TD first enforce the KKT optimality conditions of Lasso Hastie et al. (2009), and then apply the fixed-point equality for the optimizer. This approach is ill-defined for off-policy scenarios, because in those cases the fixed point of $\Pi_{\mathcal{F},\rho,\lambda}\mathcal{T}^\pi$ may not exist.

Their results imply [2] that, with probability at least $1-\delta$,

$$\|Q^\pi - Q_{\hat{\theta}}\|_n \leq \frac{1}{1-\gamma}\|Q^\pi - Q_{\theta^*}\|_n + O\left(\sqrt{\|\theta\|_1}\left(\left(\frac{\ln\frac{d}{\delta}}{n}\right)^{\frac{1}{4}} + \sqrt{\frac{1}{n}}\right)\right).$$

As the authors state, this rate is an improvement over the previous results for pathwise LSTD, because the dependence on ν_n was eliminated. The rate is slower in terms of n only in some cases, because in the Lasso-TD bounds the dependence on the dimensionality is $O(\ln d)$, whereas for pathwise LSTD it is $O\left(\sqrt{d\ln d}\right)$. Thus, their bounds convey that regularization is appropriate for $d = \Omega\left(\left(\frac{1}{n}\right)^{\frac{1}{4}}\right)$

The choice of λ for these bounds is also dependent on δ, but also for this case it should not be hard to provide bounds that are uniform on the confidence parameter.

As in the three works discussed before, we are also interested finding estimators with small $\|Q^\pi - Q_\theta\|_\rho$. Our approach is different because we chose to minimize the projected Bellman error and then relate this quantity to $\|Q^\pi - Q_\theta\|_\rho$.

There is an important point, however: the Lasso-TD results have no dependency on ν_n, whereas our result with the choice $M = \mathbf{I}_d$ depends on $\lambda_{\min}(C)$. Therefore their estimator has better (though non-uniform) guarantees for smaller n, but for large enough n LSTD with ℓ_1 penalty, *i.e.*, our procedure with $M = \mathbf{I}_d$, has tighter error bounds than Lasso-TD.

[2] The original result takes the infimum over all θ of the right-hand side.

Chapter 5

Conclusion

In this dissertation we showed results for ℓ_1-regularized linear estimation. The special cases of regression and reinforcement learning have been studied by Bartlett et al. (2009) and Ghavamzadeh et al. (2011), respectively. We took a more general approach to obtain results for both cases, and we showed that the quantities we bound are in fact related to important error measures in those scenarios. For our regression formulation, which is different from the one used by Bartlett et al. (2009), we have obtained performance bounds that are uniform in δ.

Furthermore, we separated the deterministic analysis from the stochastic one. This separation allows other extensions to be made, *e.g.*, performance bounds when different sampling processes are used.

5.1 Future work

We would like to test the potential of our analysis and results, by investigating how they can be further generalized and what other special cases deserve special attention. We can state the future work in terms of questions we wish to investigate, in order of priority:

1. How can we choose M to improve our results? Can we extend our results to apply for random M?
2. Is there a suitable notion of estimate truncation for our linear estimation problem?
3. What kinds of penalties can our results be extended for?
4. What types of losses can our results be extended for?
5. Can we provide meaningful bounds for $M = \hat{C}^+$?
6. Should we use other estimators of C^{-1} as M, rather than $\hat{C}^+$?
7. Are there other problems we can shed light upon by extending our results?
8. How far can we go with proving performance lower-bounds?

There are, as well, some other questions of secondary importance to this work that are nonetheless interesting:

- What is the maximum number of steps performed by LARS Efron et al. (2004) (as a function of n and d)?
- How can we generalize the results of Pittenger (1990) for matrices?

Bibliography

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Bartlett, P. L., Mendelson, S., and Neeman, J. (2009). ℓ_1-regularized linear regression: Persistence and oracle inequalities. Technical report, UC Berkeley.

Bonnans, J. and Shapiro, A. (2000). *Perturbation analysis of optimization problems*. Springer series in operations research. Springer.

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

Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1):33–57.

Bühlmann, P. and Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351.

Dasgupta, S. (2000). Experiments with random projection. *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, pages 143–151.

Doukhan, P. (1994). *Mixing: properties and examples*. Lecture notes in statistics. Springer.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Farahmand, A.-M. and Szepesvári, C. (2011). Model selection in reinforcement learning. *Machine Learning*, pages 1–34. 10.1007/s10994-011-5254-7.

Gerchinovitz, S. and Yu, J. Y. (2011). Adaptive and optimal online linear regression on ℓ_1-balls. *ArXiv preprint arXiv:1105.4042*.

Ghavamzadeh, M., Lazaric, A., Maillard, O.-A., and Munos, R. (2010). LSTD with random projections. *Advances in Neural Information Processing Systems*, 23:721–729.

Ghavamzadeh, M., Lazaric, A., Munos, R., and Hoffman, M. (2011). Finite-sample analysis of Lasso-TD. *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1177–1184.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition.

Hsu, D., Kakade, S. M., and Zhang, T. (2011). An analysis of random design linear regression. *ArXiv preprint at arXiv:1106.2363v1*, math.ST.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, pages 361–379.

Johns, J., Painter-Wakefield, C., and Parr, R. (2010). Linear complementarity for regularized policy evaluation and improvement. *Advances in Neural Information Processing Systems*, 23:1009–1017.

Kolter, J. Z. and Ng, A. (2009). Regularization and feature selection in least-squares temporal difference learning. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 521–528.

Lazaric, A., Ghavamzadeh, M., and Munos, R. (2010). Finite-sample analysis of LSTD. *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 615–622.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep*, 76:2007.

Pittenger, A. O. (1990). Sharp mean-variance bounds for Jensen-type inequalities. *Statistics & Probability Letters*, 10(2):91–94.

Rosasco, L. (2006). *Regularization Approaches in Learning Theory*. PhD thesis, DISI, Università di Genova.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer, 1st edition.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press.

Sweldens, W. and Schröder, P. (1996). Building your own wavelets at home. In *Wavelets in Computer Graphics, ACM SIGGRAPH Course Notes*. ACM.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Morgan & Claypool.

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

Wasserman, L. (2006). *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320.

Appendix A

Algorithmic considerations

It is useful to discuss algorithmic aspects of our linear estimation problem, because the underlying minimization can be done efficiently and model selection can be incorporated into the procedure at marginal cost.

A.1 Computing the Lasso

There are at least three different ways of computing the Lasso, that is, finding

$$\arg\min_{\theta} \frac{1}{2}\|\hat{A}\theta - \hat{b}\|_M + \lambda\|\theta\|_1.$$

We have included the factor $\frac{1}{2}$ multiplying the loss for convenience when dealing with its gradient. In this chapter, we will abuse the notation to denote a minimizer of this Lasso problem by $\hat{\theta}_\lambda$, whereas $\hat{\theta}_{2\lambda}$ would be the correct notation. Also, recall that $\hat{A} \in \mathbb{R}^{q\times r}$, $\hat{b} \in \mathbb{R}^r$, and $M \in \mathbb{R}^{q\times r}$.

One can use Nesterov's fast gradient method Nesterov (2007), which is easy to implement and only requires careful and analytical derivations of some quantities described in the paper. An alternative is to use a coordinate descent method, since there is a closed-form solution for the Lasso if one is optimizing w.r.t. to a single coordinate of θ (Hastie et al. (2009)).

Both of these methods require provision of stopping conditions in terms of approximation quality, for instance, estimating rates of convergence or computing duality gaps. These conditions, some of which are described in Boyd and Vandenberghe (2004), may not be good enough, or may be too expensive to be computed at every iteration.

A third option, the least-angle regression shrinkage (LARS) algorithm (Hastie et al. (2009), Efron et al. (2004)), does not require a stopping condition of such nature, and allows us to compute $\hat{\theta}_\lambda$ for various λ with the cost of computing $\hat{\theta}_\lambda$ for the smallest such λ. The variation of LARS used to compute the Lasso is presented in Table A.1, and is based mostly on the presentation by Hastie et al. (2009). We must introduce some necessary terminology and notation before discussing LARS.

First, our notation for sub-vectors and sub-matrices: for a given vector $v \in \mathbb{R}^r$, v_i denotes the value of its i-th coordinate, and for a matrix $D \in \mathbb{R}^{q\times r}$, D_i denotes its i-th column. Likewise, for a set $\mathcal{I}$ whose elements are integers in $[1, r]$, $v_\mathcal{I}$ is given by the coordinates of v indexed by the elements in $\mathcal{I}$ (sorted by index), and $D_\mathcal{I}$ is

Procedure LARS
Input: $\hat{A}$,$\hat{b}$,M,λ
Output: $\hat{\theta}_\lambda$
$\hat{\theta}_\eta \leftarrow \mathbf{0}$
$\lambda' \leftarrow \|\nabla\mathcal{L}(\hat{\theta}_\eta)\|_\infty$
$j \leftarrow$ an arbitrary i s.t. $|\nabla\mathcal{L}(\hat{\theta}_\eta)_i| = \lambda'$
$\mathcal{J} \leftarrow \{j\}$
while $\eta > \lambda$ **do**
  $\hat{\theta} \leftarrow \mathrm{LS}(\hat{A}, \hat{b}, M, \mathcal{J})$
  $c(\alpha) \leftarrow \nabla\mathcal{L}(\hat{\theta}_\eta) + \alpha\nabla\mathcal{L}(\hat{\theta} - \hat{\theta}_\eta)$, for $0 \leq \alpha \leq 1$
  $\alpha_1 \leftarrow \min\{\alpha : \alpha \in (0,1], \exists i \notin \mathcal{J}, j \in \mathcal{J} \text{ s.t. } |c(\alpha)_i| = |c(\alpha)_j|\} \cup \{\infty\}$
  $\alpha_2 \leftarrow \min\left\{\alpha : \alpha \in (0,1], \exists j \in \mathcal{J} \text{ s.t. } ((1-\alpha)\hat{\theta}_\eta + \alpha\hat{\theta})_j = 0\right\} \cup \{\infty\}$
  $\alpha_3 \leftarrow \min\{\alpha : \alpha \in (0,1], |c(\alpha)_j| = \lambda \; \forall j \in \mathcal{J}\} \cup \{\infty\}$
  $\alpha' \leftarrow \min\{\alpha_1, \alpha_2, \alpha_3\}$
  $\hat{\theta}_\eta \leftarrow (1-\alpha')\hat{\theta}_\eta + \alpha'\hat{\theta}$
  $\lambda' \leftarrow \|\nabla\mathcal{L}(\hat{\theta}_\eta)\|_\infty$
  if $\alpha_1 < \alpha_2, \alpha_3$ **then**
    $j \leftarrow$ an arbitrary $i \notin \mathcal{J}$ s.t. $|\nabla\mathcal{L}(\hat{\theta}_\eta)_i| = \lambda'$
    $\mathcal{J} \leftarrow \mathcal{J} \cup \{j\}$
  else if $\alpha_2 < \alpha_3$ **then**
    $\mathcal{J} \leftarrow \left\{j : \hat{\theta}_\eta \neq 0\right\}$
  endif
end
$\hat{\theta}_\lambda \leftarrow \hat{\theta}_\eta$

Table A.1: The LARS algorithm for computing the Lasso, Efron et al. (2004).

the matrix given by the columns of D indexed by the elements in $\mathcal{I}$ (also sorted by index). This indexing precedes transposition, *i.e.*, $D_{\mathcal{I}}^\top = (D_{\mathcal{I}})^\top$.

Let $\mathcal{L}(\theta) \doteq \frac{1}{2}\|\hat{A}\theta - \hat{b}\|_M^2$, and LS be such that if $\hat{\theta} = \mathrm{LS}(X, y, D, \mathcal{I})$ then

$$\hat{\theta}_{\mathcal{I}} = (X_{\mathcal{I}}^\top D X_{\mathcal{I}})^+ D X_{\mathcal{I}}^\top y,$$
$$\hat{\theta}_{\bar{\mathcal{I}}} = \mathbf{0},$$

where $+$ denotes the Moore-Penrose pseudoinverse.

The algorithm works by computing $\hat{\theta}_\eta$ for decreasing η until the desired estimator is obtained, *i.e.*, until $\eta = \lambda$. The computation is done by ensuring that the KKT conditions of the Lasso for $\hat{\theta}_\eta$ are respected.

The way $\hat{\theta}_\eta$ is modified from one iteration to the other is simple: it is a linear interpolation of $\hat{\theta}_\eta$ and $\hat{\theta}$, with parameter α'. This ensures that $|\nabla\mathcal{L}(\hat{\theta}_\eta)_j| = \lambda$ for all $j \in \mathcal{J}$. At every iteration, the value of $c(\alpha')$ before $\hat{\theta}_\eta$ is updated is the correlation vector of the "new estimator", and it is equivalent to $\nabla\mathcal{L}(\hat{\theta}_\eta)$ after $\hat{\theta}_\eta$ is updated. α' is chosen as the smallest value in $(0, 1]$ that will yield a new estimator satisfying one of the following:

1. some coordinate not in the current active set will have the same correlation

as the coordinates in that set,

2. the value of some coordinate in the current active set will become zero,

3. the correlation of the new estimator will be λ.

In the first case we include one of the satisfying coordinates in the active set, in the second case we remove from the active set all coordinates that reach zero, and in the third case we stop the algorithm and return $\hat{\theta}_\eta$, which is $\hat{\theta}_\lambda$. Also, note at least one of the cases will happen at every iteration, and case one happens when $\alpha' = \alpha_1$, and so on. Ties are broken by having case 3 be the most preceding, then case 2.

As for the complexity of LARS, if we are given $M^{\frac{1}{2}}$ beforehand, the per-iteration cost is dominated by $O\left(qr + \|\hat{\theta}_\lambda\|_0^3\right)$ – at every iteration a matrix of dimension at most $\|\hat{\theta}_\lambda\|_0 \times \|\hat{\theta}_\lambda\|_0$ has to be computed for us to have $\hat{\theta}$. If M is not diagonal, there is also an initial $\Theta(qr^2)$ cost of computing $M^{\frac{1}{2}}\hat{A}$ and $M^{\frac{1}{2}}\hat{b}$. The per-iteration cost of using M explicitly at every iteration instead of using $M^{\frac{1}{2}}\hat{A}$ and $M^{\frac{1}{2}}\hat{b}$ is $O\left(q^2r + \|\hat{\theta}_\lambda\|_0^3\right)$ if M is not diagonal. Note that the cost for diagonal M is simply $O\left(qr + \|\hat{\theta}_\lambda\|_0^3\right)$, and that providing $M^{\frac{1}{2}}$ might require performing an SVD beforehand. Finally, LARS will perform at least $\|\hat{\theta}_\lambda\|_0$ iterations, but, to the best of our knowledge, the effective number of iterations, or even whether this number is polynomial on q and r, has never been shown.

A.2 Performing model selection

Lemma A.2.1, which is based on exercise 3.27(d) of Hastie et al. (2009), shows that $\hat{\theta}_\lambda$ is a piecewise linear function of λ.

Lemma A.2.1. *For any $\lambda \in [\lambda_0, \lambda_1)$ for which the active set $\mathcal{J} = \{j : \hat{\theta}_\lambda(j) \neq 0\}$ does not change, it holds that*

$$(\hat{\theta}_\lambda)_{\mathcal{J}} = (\hat{\theta}_{\lambda_0})_{\mathcal{J}} - (\lambda - \lambda_0)\left(\hat{A}_{\mathcal{J}}^\top M \hat{A}_{\mathcal{J}}\right)^+ \operatorname{sign}(\hat{\theta}_{\lambda_0})_{\mathcal{J}}$$
$$(\hat{\theta}_\lambda)_{\bar{\mathcal{J}}} = \mathbf{0}.$$

Proof. The second equality is true by definition of the active set, so we are left with proving the first one. Since the active set does not change, we can consider the Lasso problem with $\hat{A}_{\mathcal{J}}$, $\hat{b}$, M, and $\lambda \in [\lambda_0, \lambda_1)$ as inputs:

$$\min_\theta \frac{1}{2}\|\hat{A}_{\mathcal{J}}\theta_{\mathcal{J}} - \hat{b}\|_M^2 + \lambda\|\theta_{\mathcal{J}}\|_1.$$

We know that at least at least one optimizer $\hat{\theta}_\lambda$ has $(\hat{\theta}_\lambda)_{\mathcal{J}} \neq \mathbf{0}$, so

$$\begin{aligned}
\hat{A}_{\mathcal{J}}^\top M(\hat{A}_{\mathcal{J}}(\hat{\theta}_\lambda)_{\mathcal{J}} - \hat{b}) + \lambda\operatorname{sign}((\hat{\theta}_\lambda)_{\mathcal{J}}) &= 0 \\
(\hat{A}_{\mathcal{J}}^\top M \hat{A}_{\mathcal{J}})(\hat{\theta}_\lambda)_{\mathcal{J}} &= \hat{A}_{\mathcal{J}}^\top M\hat{b} - \lambda\operatorname{sign}((\hat{\theta}_\lambda)_{\mathcal{J}}) \\
(\hat{\theta}_\lambda)_{\mathcal{J}} &= \left(\hat{A}_{\mathcal{J}}^\top M \hat{A}_{\mathcal{J}}\right)^+ \left(\hat{A}_{\mathcal{J}}^\top M\hat{b} - \lambda\operatorname{sign}((\hat{\theta}_\lambda)_{\mathcal{J}})\right),
\end{aligned}$$

This reasoning holds for any $\lambda \in [\lambda_0, \lambda_1)$, and $\operatorname{sign}(\hat{\theta}_\lambda)$ is the same for every λ in that interval, because $\mathcal{J}$ is fixed. Therefore, the lemma follows by using the equation above to factor $(\hat{\theta}_{\lambda_0})_\mathcal{J}$ out in the same equation applied to $(\hat{\theta}_\lambda)_\mathcal{J}$.

□

So, combining model selection and LARS is simple: as we compute $\hat{\theta}_a$, we also compute $\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M^2 + \lambda'\|\hat{\theta}_\lambda\|_1$ for all $\lambda \in \Lambda(a, B)$. We first calculate it for $\hat{\theta}_B$, *i.e.*, $\|\hat{b}'\|_M^2$. Then, at each iteration, immediately after we compute α' and whenever there are $\alpha \in (0, \alpha']$ such that $\hat{\theta}_\lambda = (1-\alpha)\hat{\theta}_\eta + \alpha\hat{\theta}$ for $\lambda \in \Lambda(a, B)$, we evaluate $\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda'\|\hat{\theta}_\lambda\|_1$ for all such $\hat{\theta}_\lambda$ (there may be more than one). Finally, instead of returning $\hat{\theta}_a$, we return the model with the minimum validation error, $\hat{\theta}_{\hat{\lambda}}$.

The overall cost incurred by the model selection is that of evaluating the λ'-penalized losses for each model (plus a marginal cost of a few extra comparisons). Given that $\Lambda(a, B)$ has $O(\ln n)$ elements, and that computing $\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M^2$ is dominated by computing $\hat{\theta}$ at the iteration in which this error is evaluated, the model selection will not, asymptotically, affect the complexity of computing $\hat{\theta}_a$[1].

[1]Except when LARS executes $o(\ln n)$ iterations, but even in this case the extra cost is small.

Appendix B

Proofs

In this chapter we present the proofs for the results in Chapter 2, and some discussion on issues pertaining to these proofs, such as remarks on the techniques used.

B.1 Deterministic analysis

B.1.1 Fixed choice of λ

To prove our first lemma, Lemma 2.1.1, we need the following result:

Lemma B.1.1. *For $Q \in \mathbb{R}^{k,l}$ and $w \in \mathbb{R}^l$, we have*

$$\|Qw\|_2 \leq \|Q\|_F \|w\|_2 \leq \|Q\|_F \|w\|_1.$$

Proof of Lemma B.1.1. Denoting the i-th *row* of Q by Q_i, we can use the Cauchy-Schwarz inequality to show that

$$\begin{aligned} \|Qw\|_2^2 &= \sum_{i=1}^{k} \langle Q_i, w \rangle^2 \\ &\leq \sum_{i=1}^{k} \|Q_i\|_2^2 \|w\|_2^2 \\ &= \|Q\|_F^2 \|w\|_2^2. \end{aligned}$$

The lemma follows by taking the square root of both sides, and by using the fact that $\|w\|_1 \geq \|w\|_2$. □

Lemma 2.1.1. *For any $\theta \in \mathbb{R}^d$,*

$$\left| \|A\theta - b\|_M - \|\hat{A}\theta - \hat{b}\|_M \right| \leq \Delta_A \|\theta\|_1 + \Delta_b.$$

Proof of Lemma 2.1.1. The lemma is a consequence of triangle inequality, remark B.1.1 and the definition of Δ_A and Δ_b. The two statements follow by chaining

$$\begin{aligned} \|(\hat{A} - A)\theta - (\hat{b} - b)\|_M &\leq \|(\hat{A} - A)\theta\|_M + \|(\hat{b} - b)\|_M \\ &\leq \|M^{\frac{1}{2}}(\hat{A} - A)\|_F \|\theta\|_1 + \|(\hat{b} - b)\|_M \quad \text{(Lemma B.1.1)} \\ &= \Delta_A \|\theta\|_1 + \Delta_b, \end{aligned}$$

with

$$\|(\hat{A}-A)\theta-(\hat{b}-b)\|_M \geq |\|\hat{A}\theta-\hat{b}\|_M - \|A\theta-b\|_M|,$$

and then taking, respectively, either the positive or the negative of the term inside the absolute value.

□

Lemma 2.1.2. *We have*

$$\|\hat{A}\hat{\theta}_\lambda-\hat{b}\|_M^2 \leq \|\hat{A}\theta^*-\hat{b}\|_M^2 + \lambda(\|\theta^*\|_1 - \|\hat{\theta}_\lambda\|_1).$$

Proof of Lemma 2.1.2. The lemma follows from the definition of $\hat{\theta}_\lambda$, which implies that

$$\|\hat{A}\hat{\theta}_\lambda-\hat{b}\|_M^2 + \lambda\|\hat{\theta}_\lambda\|_1 \leq \|\hat{A}\theta-\hat{b}\|_M^2 + \lambda\|\theta\|_1,$$

for any θ, particularly for θ^*. □

Lemma 2.1.3. *For any $\zeta, \zeta' \geq \|\hat{A}\theta^*-\hat{b}\|_M$ and $C_1, C_2 > 0$, it holds that*

$$\|\hat{A}\hat{\theta}_\lambda-\hat{b}\|_M + C_1\|\hat{\theta}_\lambda\|_1 + C_2 \leq \max\left\{\left(C_1-\frac{\lambda}{2\zeta}\right)\left(\frac{\zeta'^2}{\lambda}+\|\theta^*\|_1\right), 0\right\} + \frac{\lambda}{2\zeta}\|\theta^*\|_1 + C_2 + \zeta.$$

Proof of Lemma 2.1.3. Lemma 2.1.2 and a linear upper-bound on the square-root function, imply that for any $\zeta \geq \|\hat{A}\theta^*-\hat{b}\|_M$

$$\begin{aligned}\|\hat{A}\hat{\theta}_\lambda-\hat{b}\|_M &\leq \left(\|\hat{A}\theta^*-\hat{b}\|_M^2 + \lambda(\|\theta^*\|_1 - \|\hat{\theta}_\lambda\|_1)\right)^{\frac{1}{2}} \\ &\leq \left(\zeta^2 + \lambda(\|\theta^*\|_1 - \|\hat{\theta}_\lambda\|_1)\right)^{\frac{1}{2}} \\ &\leq \zeta + \frac{\lambda}{2\zeta}(\|\theta^*\|_1 - \|\hat{\theta}_\lambda\|_1).\end{aligned}$$

Therefore,

$$\|\hat{A}\hat{\theta}_\lambda-\hat{b}\|_M + C_1\|\hat{\theta}_\lambda\|_1 + C_2 \leq \zeta + \frac{\lambda}{2\zeta}(\|\theta^*\|_1 - \|\hat{\theta}_\lambda\|_1) + C_1\|\hat{\theta}_\lambda\|_1 + C_2.$$

If we group the terms multiplying $\|\hat{\theta}_\lambda\|_1$, we obtain $\left(C_1 - \frac{\lambda}{2\zeta}\right)\|\hat{\theta}_\lambda\|_1$. Then we use Lemma 2.1.2 to see that for any $\zeta' \geq \|\hat{A}\theta^*-\hat{b}\|_M$, $\|\hat{\theta}_\lambda\|_1 \leq \frac{\zeta'^2}{\lambda} + \|\theta^*\|_1$. Chaining these inequalities together, we get

$$\left(C_1-\frac{\lambda}{2\zeta}\right)\|\hat{\theta}_\lambda\|_1 \leq \max\left\{\left(C_1-\frac{\lambda}{2\zeta}\right)\left(\frac{\zeta'^2}{\lambda}+\|\theta^*\|_1\right), 0\right\}.$$

Hence, the statement follows. □

Corollary 2.1.4. *If*

$$\hat{\lambda} = 2\Delta_A(\Delta_A\|\theta^*\|_1 + \Delta_b + \|A\theta^* - b\|_M),$$

then

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq 2\Delta_A\|\theta^*\|_1 + 2\Delta_b + \|A\theta^* - b\|_M.$$

Proof of Corollary 2.1.4. We will use Lemma 2.1.3 by applying a specific choice of $\hat{\lambda}$.

From Lemma 2.1.1, we see that the condition of Lemma 2.1.3 is met with $C_1 = \Delta_A$ and $C_2 = \Delta_b$. We apply this lemma with $\zeta = \zeta' = \Delta_A\|\theta^*\|_1 + \Delta_b + \|A\theta^* - b\|_M \geq \|\hat{A}\theta^* - \hat{b}\|_M$ to obtain

$$\begin{aligned}\|A\hat{\theta}_\lambda - b\|_M &\leq \max\left\{\frac{C_1\zeta^2}{\lambda} - \frac{\zeta}{2} + C_1\|\theta^*\|_1 - \frac{\lambda}{2\zeta}\|\theta^*\|_1, 0\right\} + \frac{\lambda}{2\zeta}\|\theta^*\|_1 + C_2 + \zeta \\ &= \max\left\{\frac{\Delta_A\zeta^2}{\lambda} + \frac{\zeta}{2} + \Delta_A\|\theta^*\|_1, \frac{\lambda}{2\zeta}\|\theta^*\|_1 + \zeta\right\} + \Delta_b.\end{aligned}$$

The two terms in the $\max$ are equal when $\lambda = 2\Delta_A\zeta$, and with $\hat{\lambda} = 2\Delta_A\zeta$

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq \Delta_A\|\theta^*\|_1 + \Delta_b + \zeta,$$

and the result follows. □

B.1.2 Model selection

Lemma 2.1.6. *Under Assumption 2.1.5, if*

$$\hat{\lambda} \doteq \arg\min_{\lambda\in\Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda'\|\hat{\theta}_\lambda\|_1$$

and $\lambda' \geq \Delta_A$, then, for any $\zeta \geq \|\hat{A}\theta^* - \hat{b}\|_M$,

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq \max\left\{2\lambda', \frac{a}{2\zeta}\right\}\|\theta^*\|_1 + \Delta_b + \zeta.$$

Proof of Lemma 2.1.6. Let $\lambda_o \geq 0$ be a particular, but not yet specified, real number. The proof is divided in three cases: $\lambda_o \in [a, B]$, $\lambda_o \in (0, a)$ and $\lambda_o \in (B, \infty)$. The case in which $\lambda_o = 0$ will be seen to be absolutely trivial. First, we will show how to relate $\|A\hat{\theta}_{\hat{\lambda}} - b\|_M$ and $\|A\hat{\theta}_\lambda - b\|_M$ for any $\lambda \in \Lambda(a, B)$.

Lemma 2.1.1, the definition of $\hat{\theta}_{\hat{\lambda}}$ and $\lambda \geq \Delta_A$ imply that, for any $\lambda \in \Lambda(a, B)$,

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M &\leq \Delta_A\|\hat{\theta}_{\hat{\lambda}}\|_1 + \Delta_b + \|\hat{A}\hat{\theta}_{\hat{\lambda}} - \hat{b}\|_M \\ &\leq \Delta_A\|\hat{\theta}_{\hat{\lambda}}\|_1 + \Delta_b + \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda'(\|\hat{\theta}_\lambda\|_1 - \|\hat{\theta}_{\hat{\lambda}}\|_1) \\ &\leq \lambda'\|\hat{\theta}_\lambda\|_1 + \Delta_b + \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M.\end{aligned}$$

Now we will cover the proofs for the three cases. First, suppose $\lambda_o \in [a, B]$. Because of the grid scheme, there exists $\lambda_p \in \Lambda(a, B)$ s.t. $\frac{1}{2}\lambda_o \leq \lambda_p \leq \lambda_o$, so if we apply Lemma 2.1.3 with $C_1 = \lambda'$, $C_2 = \Delta_b$, and

$$\zeta = \zeta' \geq \|\hat{A}\theta^* - \hat{b}\|_M,$$

we obtain

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M &\leq \lambda'\|\hat{\theta}_{\lambda_p}\|_1 + \Delta_b + \|\hat{A}\hat{\theta}_{\lambda_p} - \hat{b}\|_M \\ &\leq \max\left\{\lambda'\left(\frac{\zeta^2}{\lambda_p} + \|\theta^*\|_1\right) - \frac{\zeta}{2} - \frac{\lambda_p}{2\zeta}\|\theta^*\|_1, 0\right\} + \frac{\lambda_p}{2\zeta}\|\theta^*\|_1 + \Delta_b + \zeta \\ &\leq \max\left\{\lambda'\left(\frac{\zeta^2}{\lambda_p} + \|\theta^*\|_1\right) - \frac{\zeta}{2}, \frac{\lambda_p}{2\zeta}\|\theta^*\|_1\right\} + \Delta_b + \zeta.\end{aligned}$$

We want the factor of $\|A\theta^* - b\|_M$ to be at most one in the bound. Thus we will use that $\lambda_p \in \left[\frac{1}{2}\lambda_o, \lambda_o\right]$ and choose $\lambda_o = 4\lambda'\zeta$ to ensure this. To conclude the proof for the first case,

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M &\leq \max\left\{\lambda'\left(\frac{2\zeta^2}{\lambda_o} + \|\theta^*\|_1\right) - \frac{\zeta}{2}, \frac{\lambda_o}{2\zeta}\|\theta^*\|_1\right\} + \Delta_b + \zeta \\ &\leq \max\left\{\lambda'\|\theta^*\|_1, 2\lambda'\|\theta^*\|_1\right\} + \Delta_b + \zeta \\ &= 2\lambda'\|\theta^*\|_1 + \Delta_b + \zeta.\end{aligned}$$

Now suppose that $\lambda_o \in (0, a)$, *i.e.*, $a \geq 4\lambda'\zeta$. Then we have that

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M &\leq \lambda'\|\hat{\theta}_a\|_1 + \Delta_b + \|\hat{A}\hat{\theta}_a - \hat{b}\|_M \\ &\leq \max\left\{\lambda'\left(\frac{2\zeta^2}{a} + \|\theta^*\|_1\right) - \frac{\zeta}{2}, \frac{a}{2\zeta}\|\theta^*\|_1\right\} + \Delta_b + \zeta \\ &\leq \max\left\{\lambda'\left(\frac{2\zeta^2}{\lambda_o} + \|\theta^*\|_1\right) - \frac{\zeta}{2}, \frac{a}{2\zeta}\|\theta^*\|_1\right\} + \Delta_b + \zeta \\ &\leq \max\left\{\lambda'\|\theta^*\|_1, \frac{a}{2\zeta}\|\theta^*\|_1\right\} + \Delta_b + \zeta.\end{aligned}$$

For the third case, when $\lambda_o > B$, we have that $\|\hat{\theta}_B\|_1 = \|\hat{\theta}_{\lambda_o}\|_1 = 0$, and

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M &\leq \Delta_b + \|\hat{A}\hat{\theta}_B - \hat{b}\|_M + \frac{B}{2\zeta}\|\theta^*\|_1 \\ &\leq \Delta_b + \|\hat{A}\hat{\theta}_B - \hat{b}\|_M + \frac{\lambda_o}{2\zeta}\|\theta^*\|_1 \\ &\leq \Delta_b + \zeta + 2\lambda'\|\theta^*\|,\end{aligned}$$

where we have used Lemma 2.1.2 a linear upper-bound in the square-root function, as in the proof of Lemma 2.1.3.

The lemma follows by taking the maximum over the bounds for each case.

□

B.2 Stochastic analysis

B.2.1 Fixed choice of λ

Lemma 2.2.1 (Uniform convergence bound for a fixed choice of λ, when $A\theta^* = b$)**.** *Assume that $A\theta^* = b$ and that there exist a positive constant c_1 and a decreasing function*

$S : (0,1) \to (0,\infty)$ *such that, for any* $0 < \delta < 1$*, with probability at least* $1-\delta$*, the following hold simultaneously:*

$$\Delta_A \leq c_1 S\left(\frac{\delta}{2}\right), \tag{2.1}$$

$$\Delta_b \leq c_1 S\left(\frac{\delta}{2}\right). \tag{2.2}$$

If

$$\hat{\lambda} = c_1^2,$$

then, for any $0 < \delta < 1$*, with probability at least* $1-\delta$*, it holds that*

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq c_1 \cdot \max\left\{(\|\theta^*\|_1+1)^2 S\left(\frac{\delta}{2}\right)^3 + \frac{1}{2}S\left(\frac{\delta}{2}\right)(\|\theta^*\|_1+1), \right.$$
$$\left. \frac{1}{2S\left(\frac{\delta}{2}\right)}\frac{\|\theta^*\|_1}{\|\theta^*\|_1+1} + S\left(\frac{\delta}{2}\right)\right\} + c_1 S\left(\frac{\delta}{2}\right)\|\theta^*\|_1 + c_1 S\left(\frac{\delta}{2}\right).$$

Proof of Lemma 2.2.1. Consider the proof of Lemma 2.1.3 with

$$\zeta = \zeta' = c_1 S\left(\frac{\delta}{2}\right)(\|\theta^*\|_1+1),$$

$C_1 = \Delta_A$ and $C_2 = \Delta_b$ [1]. Then

$$\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M \leq \max\left\{\frac{\Delta_A}{\lambda}(\|\theta^*\|_1+1)^2 c_1^2 S\left(\frac{\delta}{2}\right)^2 + \Delta_A\|\theta^*\|_1 \right.$$
$$+ \frac{1}{2}c_1 S\left(\frac{\delta}{2}\right)(\|\theta^*\|_1+1), \frac{\lambda}{2c_1 S\left(\frac{\delta}{2}\right)}\frac{\|\theta^*\|_1}{\|\theta^*\|_1+1}$$
$$\left. + c_1 S\left(\frac{\delta}{2}\right)(\|\theta^*\|_1+1)\right\} + \Delta_b$$

holds with probability at least $1-\delta$. For the particular value $\hat{\lambda} = c_1^2$,

$$\|\hat{A}\hat{\theta}_{\hat{\lambda}} - \hat{b}\|_M \leq \max\left\{\Delta_A(\|\theta^*\|_1+1)^2 S\left(\frac{\delta}{2}\right)^2 + \Delta_A\|\theta^*\|_1 \right.$$
$$+ \frac{1}{2}c_1 S\left(\frac{\delta}{2}\right)(\|\theta^*\|_1+1), \frac{c_1}{2S\left(\frac{\delta}{2}\right)}\frac{\|\theta^*\|_1}{\|\theta^*\|_1+1}$$
$$\left. + c_1 S\left(\frac{\delta}{2}\right)(\|\theta^*\|_1+1)\right\} + \Delta_b,$$

and the lemma follows by upper-bounding Δ_A and Δ_b with $c_1 S\left(\frac{\delta}{2}\right)$. □

[1] There is no particular relationship among C_1,C_2 and c_1, other than the relationships given by what they are defined as, *i.e.*, the similarity in the names C_1 and c_1 does not imply anything.

To prove Lemma 2.2.2, a result by Pittenger (1990) will be used to control $\mathbb{E}\left[\frac{1}{\zeta}\right]$, where $\zeta \geq \|\hat{A}\theta^* - \hat{b}\|_M$. More generally, the result can be used to bound inverse (shifted) moments of non-negative random variable in terms of its mean and variance. Lemma B.2.1 extends theorem 1 and example 3.3 in Pittenger (1990) for the case of a positive random variable, and provides a bound for the inverse shifted moment in terms of an upper-bound on the variable's variance (instead of the variance itself, as originally in Pittenger (1990)).

Lemma B.2.1. *Let $X \in [0; \infty)$ be a random variable such that $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) \leq \sigma^2$. Also, let $c > 0$ be some constant. Then*

$$\mathbb{E}\left[\frac{1}{c+X}\right] \leq \frac{1}{c} \cdot \frac{c\mu + \sigma^2}{c\mu + \mu^2 + \sigma^2}.$$

Moreover, if $c = \sigma$, then

$$\mathbb{E}\left[\frac{1}{c+X}\right] \leq \frac{2}{c+\mu}.$$

Proof. To prove the first statement, we use a change of measure and Jensen's inequality. First, note that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{X+c} - \frac{1}{c}\right)\frac{1}{\mu}\right] &= \mathbb{E}\left[\left(\frac{1}{X+c} - \frac{1}{c}\right)\frac{1}{\mu}\mathbb{I}_{\{X>0\}}\right] \\ &= \mathbb{E}\left[\left(\frac{1}{X+c} - \frac{1}{c}\right)\frac{1}{X}\frac{X}{\mu}\mathbb{I}_{\{X>0\}}\right].\end{aligned}$$

Assume ν is a measure s.t.

$$\begin{aligned}\mathbb{E}_\nu[X] &= \mathbb{E}\left[\frac{X^2}{\mu}\mathbb{I}_{\{X>0\}}\right] \\ &= \mathbb{E}\left[\frac{X^2}{\mu}\right] \\ &= \frac{\mathrm{Var}(X)}{\mu} + \mu \\ &\leq \frac{\sigma^2}{\mu} + \mu,\end{aligned}$$

where we have used that $\mathbb{E}\left[\frac{X^2}{\mu}\mathbb{I}_{\{X=0\}}\right] = 0$. Then

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{X+c} - \frac{1}{c}\right)\frac{1}{X}\frac{X}{\mu}\mathbb{I}_{\{X>0\}}\right] &= \mathbb{E}_\nu\left[\left(\frac{1}{X+c} - \frac{1}{c}\right)\frac{1}{X}\right] \\ &= \mathbb{E}_\nu\left[-\frac{1}{c(X+c)}\right],\end{aligned}$$

and Jensen's inequality tells us that

$$\begin{aligned}\mathbb{E}_\nu\left[-\frac{1}{c(X+c)}\right] &\leq -\frac{1}{c\left(\mathbb{E}_\nu[X] + c\right)} \\ &\leq -\frac{1}{c\left(\frac{\sigma^2}{\mu} + \mu + c\right)}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{X+c}\right] &\leq \frac{1}{c} - \frac{\mu}{c\left(\frac{\sigma^2}{\mu}+\mu+c\right)} \\ &= \frac{1}{c}\left(1 - \frac{\mu^2}{\sigma^2+\mu^2+c\mu}\right) \\ &= \frac{1}{c}\left(\frac{\sigma^2+c\mu}{\sigma^2+\mu^2+c\mu}\right).\end{aligned}$$

The second statement follows from simple algebra, after we use that $c = \sigma$:

$$\begin{aligned}\frac{1}{c}\left(\frac{\sigma^2+c\mu}{\sigma^2+\mu^2+c\mu}\right) &= \frac{1}{c}\left(\frac{c^2+c\mu}{c^2+\mu^2+c\mu}\right) \\ &= \frac{c+\mu}{\frac{1}{2}(c+\mu)^2+\frac{1}{2}\left(c^2+\mu^2\right)} \\ &\leq \frac{2}{c+\mu}.\end{aligned}$$

□

Lemma 2.2.2 (Expectation bound for a fixed choice of λ)**.** *Assume that there exist a positive constant $c_1 < 1$ and a decreasing function $S : (0,1) \to (0,\infty)$ such that, for any $0 < \delta < 1$, with probability at least $1-\delta$,*

$$\Delta_A \leq c_1 S\left(\delta\right).$$

Moreover, assume that w.p. 1

$$\begin{aligned}\|\hat{A}\theta^* - \hat{b}\|_M &\leq L, \\ \Delta_A &\leq A_{\max},\end{aligned}$$

and that there exists c_2 such that

$$\mathrm{Var}(\|\hat{A}\theta^* - \hat{b}\|_M) \leq c_2^2.$$

Then, choosing

$$\hat{\lambda} = 2(L+c_2)c_1 S\left(c_1^2\right)$$

implies that

$$\begin{aligned}\mathbb{E}\left[\|A\hat{\theta}_\lambda - b\|_M\right] \leq A_{\max}\left(\frac{L}{2S\left(c_1^2\right)} + c_1\|\theta^*\|_1\right)c_1 + \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c_2 \\ + 2\frac{L+c_2}{\mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c_2}\|\theta^*\|_1 c_1 S\left(c_1^2\right) + \mathbb{E}\left[\Delta_b\right].\end{aligned}$$

Proof of Lemma 2.2.2. We know from Lemmas 2.1.1 and 2.1.3 that for any $L \geq \zeta \geq \|\hat{A}\theta^* - \hat{b}\|_M$

$$\begin{aligned}
\|A\hat{\theta}_\lambda - b\|_M &\leq \max\left\{\left(\Delta_A - \frac{\lambda}{2\zeta}\right)\left(\frac{\zeta^2}{\lambda} + \|\theta^*\|_1\right), 0\right\} + \zeta + \frac{\lambda}{2\zeta}\|\theta^*\|_1 + \Delta_b \\
&= \left(\Delta_A - \frac{\lambda}{2\zeta}\right)\left(\frac{\zeta^2}{\lambda} + \|\theta^*\|_1\right)\mathbb{I}_{\{\lambda < 2\zeta\Delta_A\}} + \zeta + \frac{\lambda}{2\zeta}\|\theta^*\|_1 + \Delta_b \\
&\leq A_{\max}\left(\frac{(L+c_2)^2}{\lambda} + \|\theta^*\|_1\right)\mathbb{I}_{\{\lambda < 2\zeta\Delta_A\}} + \zeta + \frac{\lambda}{2\zeta}\|\theta^*\|_1 + \Delta_b.
\end{aligned}$$

We will choose $\zeta = \|\hat{A}\theta^* - \hat{b}\|_M + c_2$ and take the expectation of both sides of the bound, which results in

$$\begin{aligned}
\mathbb{E}\left[\|A\hat{\theta}_\lambda - b\|_M\right] \leq A_{\max}&\left(\frac{(L+c_2)^2}{\lambda} + \|\theta^*\|_1\right)\mathbb{P}\left(\lambda < 2\zeta\Delta_A\right) \\
&+ \mathbb{E}\left[\zeta\right] + \mathbb{E}\left[\frac{\lambda}{2\zeta}\right]\|\theta^*\|_1 + \mathbb{E}\left[\Delta_b\right].
\end{aligned}$$

The concentration assumption and $\|\hat{A}\theta^* - \hat{b}\|_M \leq L$ imply that

$$\begin{aligned}
\mathbb{P}\left(2\Delta_A\zeta > 2(L+c_2)c_1 S\left(\delta\right)\right) &\leq \mathbb{P}\left(\Delta_A > c_1 S\left(\delta\right)\right), \\
&\leq \delta
\end{aligned}$$

thus if we pick $\hat{\lambda} = 2(L+c_2)c_1 S\left(c_1^2\right)$, we get

$$\begin{aligned}
\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] &\leq A_{\max}\left(\frac{(L+c_2)^2}{2(L+c_2)c_1 S\left(c_1^2\right)} + \|\theta^*\|_1\right)c_1^2 \\
&\quad + \mathbb{E}\left[\zeta\right] + \mathbb{E}\left[\frac{1}{\zeta}\right](L+c_2)\|\theta^*\|_1 c_1 S\left(c_1^2\right) + \mathbb{E}\left[\Delta_b\right] \\
&\leq A_{\max}\left(\frac{L+c_2}{2S\left(c_1^2\right)} + c_1\|\theta^*\|_1\right)c_1 \\
&\quad + \mathbb{E}\left[\zeta\right] + \mathbb{E}\left[\frac{1}{\zeta}\right](L+c_2)\|\theta^*\|_1 c_1 S\left(c_1^2\right) + \mathbb{E}\left[\Delta_b\right].
\end{aligned}$$

To bound $\mathbb{E}\left[\frac{1}{\zeta}\right]$, we use theorem B.2.1. The assumptions that $\mathrm{Var}(\|\hat{A}\theta^* - \hat{b}\|_M) \leq c_2^2$ and that $\zeta > 0$ come into play, so that we can apply Lemma B.2.1 with $c = c_2$ and $X = \zeta - c_2$:

$$\mathbb{E}\left[\frac{1}{\zeta}\right] \leq \frac{2}{\mathbb{E}\left[\zeta\right]}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}&\left(\frac{L+c_2}{2S\left(c_1^2\right)} + c_1\|\theta^*\|_1\right)c_1 + \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c_2 \\
&+ 2\frac{L+c_2}{\mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c_2}\|\theta^*\|_1 c_1 S\left(c_1^2\right) + \mathbb{E}\left[\Delta_b\right].
\end{aligned}$$

□

B.2.2 Model selection

Corollary 2.2.3 (Non-uniform high-probability performance bound for estimators obtained through model selection)**.** *Assume that there exist positive constants $c_1, c_2 < 1$, and decreasing functions $S_1, S_2 : (0,1) \to (0,\infty)$ such that, for any $0 < \delta < 1$, with probability at least $1-\delta$, the following hold simultaneously:*

$$\Delta_A \leq c_1 S_1\left(\frac{\delta}{2}\right),$$
$$\Delta_b \leq c_2 S_2\left(\frac{\delta}{2}\right).$$

Then, under Assumption 2.1.5, for any $0 < \delta < 1$, if

$$\lambda' = c_1 S_1\left(\frac{\delta}{2}\right),$$
$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda(2c_3\lambda', B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda'\|\hat{\theta}_\lambda\|_1,$$

where c_3 is a constant s.t. $0 < 2c_3\lambda' \leq B$, then, with probability at least $1-\delta$, it holds that

$$\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq 3c_1 S_1\left(\frac{\delta}{2}\right)\|\theta^*\|_1 + 2c_2 S_2\left(\frac{\delta}{2}\right) + \|A\theta^* - b\|_M + c_3.$$

Proof of Corollary 2.2.3. We use Lemma 2.1.6 with

$$\lambda' = c_1 S\left(\frac{\delta}{2}\right),$$
$$\zeta = \Delta_A\|\theta^*\|_1 + \Delta_b + \|A\theta^* - b\|_M + c_3,$$
$$a = 4c_3\lambda'.$$

The concentration assumptions imply that the following hold (jointly) with probability at least $1-\delta$:

$$\Delta_A \leq \lambda' = c_1 S\left(\frac{\delta}{2}\right),$$
$$\Delta_b \leq c_2 S\left(\frac{\delta}{2}\right).$$

Therefore, because

$$\frac{a}{2\zeta} \leq \frac{4c_3\lambda'}{2c_3} = 2\lambda',$$

if follows that, with probability at least $1-\delta$,

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M &\leq 2\lambda'\|\theta^*\|_1 + \Delta_A\|\theta^*\|_1 + 2\Delta_b + \|A\theta^* - b\|_M + c_3 \\ &\leq 3c_1 S\left(\frac{\delta}{2}\right)\|\theta^*\|_1 + 2c_2 S\left(\frac{\delta}{2}\right) + \|A\theta^* - b\|_M + c_3.\end{aligned}$$

□

Proof of Corollary 2.2.4. The statement can be proved by using tail integration, Lemma 2.1.6 and Corollary 2.2.3. First, taking

$$\zeta = \Delta_A \|\theta^*\|_1 + \Delta_b + \|A\theta^* - b\|_M + c_3,$$

notice that having $\|A\theta^* - b\|_M \leq \|b\|_M$, $\|M^{\frac{1}{2}}\hat{A}\|_F \leq A_{\max}$ and $\|M^{\frac{1}{2}}\hat{b}\|_2 \leq b_{\max}$ w.p. 1 implies that

$$\zeta \leq A_{\max}\|\theta^*\|_1 + 2b_{\max} + c_3$$

w.p. 1. In Lemma 2.1.6, if $\lambda' < \Delta_A$, then the term $(\Delta_A - \lambda')\|\hat{\theta}_{\hat{\lambda}}\|_1$ is positive, but otherwise the result is the same:

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq \mathbb{I}_{\{\lambda' < \Delta_A\}}(\Delta_A - \lambda')\|\hat{\theta}_{\hat{\lambda}}\|_1 + 2\lambda'\|\theta^*\|_1 \\ + \Delta_A\|\theta^*\|_1 + 2\Delta_b + \|A\theta^* - b\|_M + c_3. \end{aligned} \tag{B.1}$$

Now, the optimality conditions of $\hat{\theta}_{\hat{\lambda}}$ allow us to bound $\|\hat{\theta}_{\hat{\lambda}}\|_1$ in terms of the loss of the zero-vector:

$$\begin{aligned}\mathbb{I}_{\{\lambda' < \Delta_A\}}(\Delta_A - \lambda')\|\hat{\theta}_{\hat{\lambda}}\|_1 &\leq \mathbb{I}_{\{\lambda' < \Delta_A\}}(\Delta_A - \lambda')\left(\frac{\|\hat{b}\|_M^2}{\hat{\lambda}}\right) \\ &\leq \mathbb{I}_{\{\lambda' < \Delta_A\}} A_{\max}\left(\frac{b_{\max}^2}{a}\right),\end{aligned}$$

which holds w.p. 1. So, if we consider the bound above and take the expectation on both sides of (B.1) we obtain

$$\begin{aligned}\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}\left(\frac{b_{\max}^2}{a}\right)\mathbb{P}\left(\lambda' < \Delta_A\right) + 2\lambda'\|\theta^*\|_1 \\ + \mathbb{E}\left[\Delta_A\right]\|\theta^*\|_1 + 2\mathbb{E}\left[\Delta_b\right] + \|A\theta^* - b\|_M + c_3.\end{aligned}$$

Choosing

$$\begin{aligned}\lambda' &= c_1 S\left(\frac{\delta}{2}\right), \\ a &= 4c_3\lambda'\end{aligned}$$

implies, by the concentration assumption, that $\mathbb{P}\left(\lambda' < \Delta_A\right) \leq \delta$, and therefore

$$\begin{aligned}\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}\left(\frac{b_{\max}^2}{2c_3c_1S\left(\frac{\delta}{2}\right)}\right)\delta + 2c_1S\left(\frac{\delta}{2}\right)\|\theta^*\|_1 \\ + \mathbb{E}\left[\Delta_A\right]\|\theta^*\|_1 + 2\mathbb{E}\left[\Delta_b\right] + \|A\theta^* - b\|_M + c_3.\end{aligned}$$

The result follows by taking $\delta = c_3c_1^2$.

□

B.2.3 Independent, identically-distributed sampling analysis

Corollary 2.3.5. *Assumption 2.3.4 implies that the following hold w.p. 1:*

$$\begin{aligned}
&\Delta_A \leq 2F_{2,\infty}F'_{2,\infty},\\
&\Delta_b \leq 2F_{2,\infty}R_\infty,\\
&\|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M \leq F_{2,\infty}R_\infty.
\end{aligned}$$

Proof of Corollary 2.3.5. For the first statement,

$$\begin{aligned}
\sup_{\hat{A}} \Delta_A &= \sup_{\hat{A}} \|M^{\frac{1}{2}}(\hat{A} - A)\|_F\\
&\leq 2\sup_{\hat{A}} \|M^{\frac{1}{2}}\hat{A}\|_F\\
&\leq 2\sup_{\hat{A}} \frac{1}{n}\sum_{t=1}^{n} \|M^{\frac{1}{2}}A_t\|_F\\
&\leq 2F_{2,\infty}F'_{2,\infty},
\end{aligned}$$

and the statement on the range of Δ'_A follows analogously, because $\hat{A}$ and $\hat{A}'$ are assumed to be sampled from the same distribution. For the second and fourth statements, since $\hat{b}$ and $\hat{b}'$ are sampled from the same distribution,

$$\begin{aligned}
\sup_{\hat{b}} \Delta'_b &= \sup_{\hat{b}} \Delta_b\\
&= \sup_{\hat{b}} \|M^{\frac{1}{2}}(\hat{b} - b)\|_2\\
&\leq 2\sup_{\hat{b}} \|M^{\frac{1}{2}}\hat{b}\|_2\\
&\leq 2\sup_{\hat{b}} \frac{1}{n}\sum_{t=1}^{n} \|M^{\frac{1}{2}}b_t\|_2\\
&\leq 2F_{2,\infty}R_\infty.
\end{aligned}$$

□

Corollary 2.3.6. *Consider Assumptions 2.3.4. From theorem 2.3.2 it follows that*

$$\begin{aligned}
&\mathbb{P}\left(\Delta_A > 2F_{2,\infty}F'_{2,\infty}T(n,\delta)\sqrt{\frac{2}{n}}\right) \leq \delta,\\
&\mathbb{P}\left(\Delta_b > 2F_{2,\infty}R_\infty T(n,\delta)\sqrt{\frac{2}{n}}\right) \leq \delta,
\end{aligned}$$

where T is as in Definition 2.3.3.

Proof of Corollary 2.3.6. The statements are a sheer application of Hoeffding's inequality with the ranges as established by Corollary 2.3.5 □

Corollary 2.3.8. *Under Assumptions 2.3.1 and 2.3.4, if $A\theta^* = b$ and*

$$\hat{\lambda} = \frac{2}{n} F_{2,\infty}^2 \max\{R_\infty, F'_{2,\infty}\}^2,$$

then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, it holds that

$$\begin{aligned}\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq F_{2,\infty} \max\{R_\infty, F'_{2,\infty}\} \sqrt{\frac{2}{n}} \cdot \max \Bigg\{ &(\|\theta^*\|_1 + 1)^2 T\left(n, \frac{\delta}{2}\right)^3 + \\ &\frac{1}{2} T\left(n, \frac{\delta}{2}\right)(\|\theta^*\|_1 + 1), \frac{1}{2T\left(n, \frac{\delta}{2}\right)} \frac{\|\theta^*\|_1}{\|\theta^*\|_1 + 1} + T\left(n, \frac{\delta}{2}\right) \Bigg\} \\ &+ F_{2,\infty} \max\{R_\infty, F'_{2,\infty}\} \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right)(\|\theta^*\|_1 + 1),\end{aligned}$$

where T is as in Definition 2.3.3.

Proof of Corollary 2.3.8. The result follows from applying Lemma 2.2.1 with

$$\begin{aligned} c_1 &\doteq F_{2,\infty} \max\{R_\infty, F'_{2,\infty}\} \sqrt{\frac{2}{n}}, \\ c_2 &\doteq n. \end{aligned}$$

□

To prove Corollary 2.3.9, we will need a bound on $\mathrm{Var}(\|\hat{A}\theta^* - \hat{b}\|_M)$, which is provided by the following Lemma.

Lemma B.2.2. *Under Assumptions 2.3.1 and 2.3.4, we have that*

$$\mathrm{Var}(\|\hat{A}\theta^* - \hat{b}\|_M) \leq \frac{2}{n} F_{2,\infty}^2 (F'_{2,\infty}\|\theta^*\|_1 + R_\infty)^2 \left(T\left(n, \frac{1}{n}\right) + 2\right)^2,$$

where T is as in Definition 2.3.3.

Proof of lemma B.2.2. First, observe that

$$\begin{aligned} \mathrm{Var}(\|\hat{A}\theta^* - \hat{b}\|_M) &= \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M^2\right] - \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right]^2 \\ &\leq \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M^2\right] - \|\mathbb{E}\left[\hat{A}\theta^* - \hat{b}\right]\|_M^2 \\ &\leq \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b} - (A\theta^* - b)\|_M^2\right]. \end{aligned}$$

We will proceed to bound the term above. Assumption 2.3.4 establishes boundedness conditions on A_t and b_t, so

$$\begin{aligned} \mathbb{E}\left[\|A_t\theta^* - b_t\|_M^2\right] &\leq F_{2,\infty}^2 (F'_{2,\infty}\|\theta^*\|_1 + R_\infty)^2 \\ &\doteq L^2, \end{aligned}$$

moreover, Hoeffding's inequality implies that, for any $0 < \delta < 1$,

$$\mathbb{P}\left(\|\hat{A}\theta^* - \hat{b} - (A\theta^* - b)\|_M^2 \leq L^2 \frac{2}{n} T(n, \delta)^2\right) \geq 1 - \delta.$$

Therefore, by tail integration

$$\mathbb{E}\left[\|\hat{A}\theta^* - \hat{b} - (A\theta^* - b)\|_M^2\right] \leq (1-\delta)L^2\frac{2}{n}T\left(n,\delta\right)^2 + \delta \cdot 4L^2,$$

and if we pick $\delta = \frac{1}{n}$, we get

$$\mathbb{E}\left[\|\hat{A}\theta^* - \hat{b} - (A\theta^* - b)\|_M^2\right] \leq L^2\frac{2}{n}T\left(n,\frac{1}{n}\right)^2 + L^2\frac{4}{n}.$$

□

Corollary 2.3.9. *Assume that there exists $L > 0$ s.t., w.p. 1,*

$$\|\hat{A}\theta^* - \hat{b}\|_M < L,$$

and let

$$c \doteq \sqrt{\frac{2}{n}}F_{2,\infty}(F'_{2,\infty}\|\theta^*\|_1 + R_\infty)\left(T\left(n,\frac{1}{n}\right) + 2\right),$$

where T is as in Definition 2.3.3. Under Assumptions 2.3.1 and 2.3.4, if $n > 2$ and

$$\hat{\lambda} = 2(L+c)\sqrt{\frac{2}{n}}F_{2,\infty}F'_{2,\infty}T\left(n,\frac{2}{n}\right),$$

then, for any $0 < \delta < 1$, with probability at least $1-\delta$, it holds that

$$\begin{aligned}\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq A_{\max}&\left(\frac{L+c}{2F_{2,\infty}F'_{2,\infty}T\left(n,\frac{2}{n}\right)} + \sqrt{\frac{2}{n}}\|\theta^*\|_1\right)\sqrt{\frac{2}{n}} + \mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] \\ &+ c + 2\frac{L+c}{\mathbb{E}\left[\|\hat{A}\theta^* - \hat{b}\|_M\right] + c}\|\theta^*\|_1\sqrt{\frac{2}{n}}F_{2,\infty}F'_{2,\infty}T\left(n,\frac{2}{n}\right) + \mathbb{E}\left[\Delta_b\right].\end{aligned}$$

where T is as in Definition 2.3.3.

Proof of Corollary 2.3.9. This corollary follows from lemma 2.2.2 under Assumptions 2.3.1 and 2.3.4, if we take

$$\begin{aligned}L &\doteq F_{2,\infty}(F'_{2,\infty}\|\theta^*\|_1 + R_\infty), \\ c_1 &\doteq \sqrt{\frac{2}{n}}, \\ c_2 &\doteq F_{2,\infty}F'_{2,\infty}, \\ c_3 &\doteq n, \\ c_4 &\doteq \frac{2}{n}F_{2,\infty}^2(F'_{2,\infty}\|\theta^*\|_1 + R_\infty)^2\left(T\left(n,\frac{1}{n}\right) + 2\right),\end{aligned}$$

where T is as defined in Equation 2.3.3, Section 2.3, and Lemma B.2.2 is used to ensure that $\mathrm{Var}(\|\hat{A}\theta^* - \hat{b}\|_M) \leq c_4$.

□

Corollary 2.3.10. *Under Assumptions 2.3.1 and 2.3.4, for any $0 < \delta < 1$, if*

$$\begin{aligned}
\lambda' &= F_{2,\infty} F'_{2,\infty} \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right), \\
a &= 4\sqrt{\frac{1}{n}}\lambda', \\
B &= F_{2,\infty}^2 F'_{2,\infty} R_\infty, \\
\hat{\lambda} &= \arg \min_{\lambda \in \Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda' \|\hat{\theta}_\lambda\|_1,
\end{aligned}$$

then, for n large enough so that $a \leq B$, with probability at least $1 - \delta$, it holds that

$$\begin{aligned}
\|A\hat{\theta}_{\hat{\lambda}} - b\|_M \leq\ & 3F_{2,\infty} F'_{2,\infty} \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right) \|\theta^*\|_1 + 2F_{2,\infty} R_\infty \sqrt{\frac{2}{n}} T\left(n, \frac{\delta}{2}\right) \\
&+ \|A\theta^* - b\|_M + \sqrt{\frac{1}{n}}.
\end{aligned}$$

Proof of Corollary 2.3.10. The result follows from Corollary 2.2.3 with

$$\begin{aligned}
c_1 = c_2 = c_3 &\doteq \sqrt{\frac{1}{n}}, \\
S_1(\delta) &= F_{2,\infty} F'_{2,\infty} \sqrt{2} T(n, \delta), \\
S_2(\delta) &= F_{2,\infty} R_\infty \sqrt{2} T(n, \delta).
\end{aligned}$$

□

Corollary 2.3.11. *Under Assumptions 2.3.1 and 2.3.4, for any $0 < \delta < 1$, if*

$$\begin{aligned}
\lambda' &= F_{2,\infty} F'_{2,\infty} \sqrt{\frac{2}{n}} T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right), \\
a &= 4\sqrt{\frac{1}{n}}\lambda', \\
B &= F_{2,\infty}^2 F'_{2,\infty} R_\infty, \\
\hat{\lambda} &= \arg \min_{\lambda \in \Lambda(a,B)} \|\hat{A}\hat{\theta}_\lambda - \hat{b}\|_M + \lambda' \|\hat{\theta}_\lambda\|_1,
\end{aligned}$$

then it holds that

$$\begin{aligned}
\mathbb{E}\left[\|A\hat{\theta}_{\hat{\lambda}} - b\|_M\right] \leq\ & A_{\max} \left(\frac{b_{\max}^2}{2F_{2,\infty} F'_{2,\infty} \sqrt{2} T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right)} \right) \sqrt{\frac{1}{n}} \\
&+ 2\sqrt{\frac{1}{n}} F_{2,\infty} F'_{2,\infty} \sqrt{2} T\left(n, \frac{1}{2n^{\frac{3}{2}}}\right) \|\theta^*\|_1 \\
&+ \mathbb{E}[\Delta_A] \|\theta^*\|_1 + 2\mathbb{E}[\Delta_b] + \|A\theta^* - b\|_M + \sqrt{\frac{1}{n}}.
\end{aligned}$$

Proof of Corollary 2.3.11. The result follows from Corollary 2.2.4 with

$$
\begin{aligned}
c_1 = c_2 = c_3 &\doteq \sqrt{\frac{1}{n}}, \\
S_1(\delta) &= F_{2,\infty} F'_{2,\infty} \sqrt{2} T(n, \delta), \\
S_2(\delta) &= F_{2,\infty} R_\infty \sqrt{2} T(n, \delta).
\end{aligned}
$$

□

Note the simple correspondence between non-uniform high-probability bounds and expectation bounds. In our results, we use the non-uniform parameter choices to ensure that the term $\Delta_A \|\hat{\theta}_{\hat{\lambda}}\|_1$ (or $\Delta_A \|\hat{\theta}_\lambda\|_1$ in Lemma 2.2.2) is eliminated. We do not want this term to appear in the bound because the upper-bounds we use for the ℓ_1 norm of the $\hat{\theta}_\lambda$ estimators grow with the inverse of λ. In the high-probability bounds, this term is effectively eliminated, and in the expectation bounds it is not a problem for this term to grow with the inverse of λ, because we can choose δ small enough so that it is dominated by the other terms. In both cases, we use a suitable choice of λ' to obtain the desired result.