Efficient Sequential Decision Making

by

Alan Malek

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter Bartlett, Chair
Associate Professor Pieter Abbeel
Professor Allan Sly

Spring 2017

Efficient Sequential Decision Making

Abstract

Efficient Sequential Decision Making

by

Alan Malek

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Peter Bartlett, Chair

This thesis studies three problems in sequential decision making across two different frameworks. The first framework we consider is online learning: for each round of a T round repeated game, the learner makes a prediction, the adversary observes this prediction and reveals the true outcome, and the learner suffers some loss based on the accuracy of the prediction. The learner's aim is to minimize the regret, which is defined to be the difference between the learner's cumulative loss and the cumulative loss of the best prediction strategy in some class. We study the minimax strategy, which guarantees the lowest regret against all possible adversary strategies. In general, computing the minimax strategy is exponential in T; we focus on two setting where efficient algorithms are possible.

The first is prediction under squared Euclidean loss. The learner predicts a point in $\mathbb{R}^d$ and the adversary is constrained to respond with a point in some compact set. The regret is with respect to the single best prediction in the set. We compute the minimax strategy and the value of the game for any compact set and show that the value is the product of a horizon-dependent constant and the squared radius of the smallest enclosing ball of the set. We also present the optimal strategy of the adversary for two important sets: ellipsoids and polytopes that intersect their smallest enclosing ball at all vertices. The minimax strategy can be cast as a simple shrinkage of the past data towards the center of this minimum enclosing ball, where the shrinkage factor can be efficiently computed before the start of the game. Noting that the value does not have any explicit dimension dependence, we then extend these results to Hilbert space, finding, once again, that the value is proportional to the squared radius of the smallest enclosing ball.

The second setting where we derive efficient minimax strategies is online linear regression. At the start of each round, the adversary chooses and reveals a vector of covariates. The regret is defined with respect to the best linear function of the covariates. We show that the minimax strategy is an easily computed linear predictor, provided that the adversary adheres to some natural constraints that prevent him from misrepresenting the scale of the problem. This strategy is horizon-independent: regardless of the length of the game, this strategy incurs no more regret than any strategy that has knowledge of the number of rounds. We also provide an interpretation of the minimax algorithm as

a follow-the-regularized-leader strategy with a data-dependent regularizer and obtain an explicit expression for the minimax regret.

We then turn to the second framework, reinforcement learning. More specifically, we consider the problem of controlling a Markov decision process (MDP) with a large state-space. Since it is intractable to compete with the optimal policy for large scale problems, we pursue the more modest goal of competing with a low-dimensional family of policies. Specifically, we restrict the variables of the dual linear program to lie in some low-dimensional subspace, and show that we can find a policy that performs almost as well as the best policy in this class. We derive separate results for the average cost and discounted cost cases. Most importantly, the complexity of our method depends on the size of the comparison class but not the size of the state-space. Preliminary experiments show the effectiveness of the proposed algorithms in a queuing application.

To my wife and family.

Contents

Acknowledgments

First and foremost, I would like to thank Peter Bartlett for all the mentorship, supervision, and guidance. I will spend the rest of my life honing my intuition, but it will never be as sharp as Peter's.

I would also like to thank my collaborators: Wouter Koolen, Yasin Abbasi, Wojciech Kotłowski, Mohammad Ghavamzadeh, Yinlam Chow, Sumeet Katariya, Manfred Warmuth, Eiji Takimoto, Victor Gabillon, Xi Chen, Fares Hedayati, and Suvrit Sra. All were instrumental in my research, and I always enjoyed our discussions. In particular, I would like to thank frequent collaborators Yasin Abbasi and Wouter Koolen for the many hours of fruitful discussion and many papers. I hope we can keep this up.

The support staff at Berkeley has been wonderful, and I'd like to especially thank Shirley Salanio and Angie Abbatecola for their help and always being patient with my paperwork.

I'd also like to thank Csaba Szepesvari, Jake Abernethy, and Manfred Warmuth for their hospitality and supporting my visits.

Internships have provided a nice perspective and I am grateful to my mentors Panagiotis Papadimitriou at Upwork and Mohammad Ghavamzadeh at Adobe Research for showing me glimpses of the world outside of academia. I'd also like to thank Nikos Vlassis and Branislav Kveton, also at Adobe, for many interesting conversations.

I wouldn't have remained sane without many good friends throughout the years; from board games to lifting in the gym to climbing to eating, I've been fortunate to have many friends encouraging me to spend some time relaxing. In no particular order, these include: Mason Liang, Matt Weiner, Paula Lipka, Alex Segal, Matt Tai, Alice Pao, Yusef Shafi, Shangliang Jiang, Walid Krichene, Alekh Agarwal, Nick Boyd, Yasin Abbasi, James McKeone, Kelvin So, Aaron Halquist, Kuang Xu, and Wouter Koolen.

Of course, my wife Yuanyuan deserves all my thanks for the endless encouragement and always telling me to follow my passions. She has been the best companion and provider of emotional support I could ask for, even when I took longer to finish that I should. My appreciation gratefully extends to Yuanyuan's family, Chia-Peng Pao, Wei-Yi Ku, Alice Pao and Matt Tai, who have been extremely supportive through the years.

And last but certainly not least, my parents, Yuko and George, for always looking out for my education and providing all the support they could. I couldn't have asked for a better family.

Chapter 1

Introduction

1.1 Motivation

The decades since the creation of the internet have seen an explosion in the rate of data generation and rapid movement towards the ubiquity of personal computers. With these developments, we have the ability for sequential interaction with systems of unprecedented size and complexity and an urgent need to understand large-scale sequential decision problems. This thesis studies two strategies to deal with this problem.

1. Online learning focuses on simple frameworks, often eschewing probabilistic assumptions, and develops simple algorithms with robust performance guarantees, even against non-stochastic data.

2. Reinforcement learning, on the other hand, attempts to model the environment as a complicated stochastic system and find an optimal policy to control it. We adapt this framework to the large-scale sequential setting by developing methods that search for approximately optimal solutions on low-complexity versions of the model.

1.2 Online Learning

Online learning describes the sequential decision problem as a repeated game between the learner and some adversary (e.g. nature). For every round $t = 1, \dots, T$,

1. the learner picks an action $a_t \in \mathcal{A}$

2. the adversary observes a_t and plays a response $x_t \in \mathcal{X}$

3. the learner suffers loss $\ell(a_t, x_t) \in [0, 1]$, and

4. the learner receives feedback (e.g. $\ell(\cdot, x_t)$ or $\ell(a_t, x_t)$).

In contrast to machine learning which often assumes stochastic data and generates sophisticated algorithms, online learning studies simple algorithms but proves performance bounds with minimal assumptions, often allowing the data to be adversarially generated as responses to the choices of the learner [12].

Since the data are adversarial, the cumulative loss can always be made large; instead, we minimize the *regret*, defined as

$$\mathcal{R} := \sum_{t=1}^{T} \ell(a_t, x_t) - L_T^*(x_1, \dots, x_T), \tag{1.1}$$

where $L_T^*(x_1, \dots, x_T)$ is the best loss of some reference class. If we take the class to be the set of fixed action, then $L_T^*(x_1, \dots, x_T) = \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, x_t)$. Obtaining regret that grows sublinearly with T implies that we are learning, as we can play almost as well as the best action in hindsight but without seeing the data sequence ahead of time.

Online learning is a useful tool in the context of large-scale sequential decision problems for several reasons. First, the algorithms are often simple with fast updates. They typically have a small memory footprint and only need to track a small number of parameters; specifically, data may be discarded once they are processed. Second, the guarantees are robust even against non-stochastic data. Third, online learning has been successful in games with partial information such as noisy feedback or only observing the loss for the taken action (e.g. multi-arm bandit problems).

The typical analysis for an online learning algorithm is to

- propose an algorithm,
- prove an upper bound on regret (e.g. using some potential function argument),
- prove a lower bound on regret (e.g. by counter example) for any algorithm, and
- hope that the two bounds meet.

This outline has been very successful and many algorithms have upper bounds that match the rate of the lower bound. In contrast, the minimax optimal algorithm provides a stronger notion of optimality: the algorithm must achieve the best possible regret against all sequences, not simply a regret that grows with the correct dependence on T.

The Minimax Algorithm

The minimax value of a game is defined to be

$$V := \min_{a_1 \in \mathcal{A}} \max_{x_1 \in \mathcal{X}} \cdots \min_{a_T \in \mathcal{A}} \max_{x_T \in \mathcal{X}} \sum_{t=1}^{T} \ell(a_t, x_t) - L_T^*(x_1, \ldots, x_T).$$

The value can be though of as regret when both players play optimal responses to the other player's actions. A strategy that plays the $\operatorname{argmin} a_t$ in the above expression is the *minimax algorithm*; it is precisely the optimal response to the adversary. It is optimal in the sense that, for any other strategy, there is an x_t sequence that causes that strategy to suffer more regret than the minimax strategy would suffer on the x_t sequence.

The value of the game and optimal strategies can be calculated by *backwards induction*. We can write

$$\begin{aligned} V &= \min_{a_1 \in \mathcal{A}} \max_{x_1 \in \mathcal{X}} \cdots \min_{a_T \in \mathcal{A}} \max_{x_T \in \mathcal{X}} \sum_{t=1}^{T} \ell(a_t, x_t) - L_T^*(x_1, \ldots, x_T) \\ &= \min_{a_1 \in \mathcal{A}} \max_{x_1 \in \mathcal{X}} \ell(a_1, x_1) + \min_{a_2 \in \mathcal{A}} \max_{x_2 \in \mathcal{X}} \ell(a_2, x_2) + \cdots + \min_{a_T \in \mathcal{A}} \max_{x_T \in \mathcal{X}} \ell(a_T, x_T) - L_T^*(x_1, \ldots, x_T). \end{aligned}$$

This form suggests that we study the recursion with base case

$$V_T(x_1, \ldots, x_T) \;=\; - L_T^*(x_1, \ldots, x_T)$$

and induction step

$$V_{t-1}(x_1, \ldots, x_{t-1}) \;=\; \min_{a_t} \max_{x_t} \ell(a_t, x_t) + V(x_1, \ldots, x_t).$$

This recursion yields the value, i.e. $V = V_0()$, and it conveniently encodes the temporal structure of the problem. That is, the minimizer for a_t depends on $x_1, \ldots, x_{t-1}$ only, which is exactly the game history the learner would have access to when choosing a_t. Similarly, the maximizer for x_t can depend on $x_1, \ldots, x_{t-1}$ as well as a_t. Performing this series of saddle point problems is known as *backwards induction*. Backwards induction requires that the learner account for the current loss, $\ell(a_t, x_t)$, as well as the future regret which encoded in $V(x_1, \ldots, x_t)$.

Advantages of the minimax algorithm include robustness and a very strong notion of optimality: the optimal constant, not just the optimal rate, is achieved. For regimes where the game length and the dimensions are comparable, the difference in constants could be significant. From a theoretical standpoint, minimax algorithms capture the inherent complexity of the problem. Instead of an arbitrary regularization (e.g. ridge regression), the structure of the problem provides the regularization. For example, for the game with square loss and a Euclidean ball constraint on nature's responses, the minimax algorithm plays a carefully shrunk empirical mean where the shrinking is determined by the game parameters (the ball radius and the current round). As a consequence, there are no parameters to tune.

Efficient Minimax Algorithms

The largest drawback of the minimax algorithm is that its computational complexity is often exponential in T. At a high level, the value-to-go function must be computed for every possible history sequence $x_1, \ldots, x_t$, which is typically exponential in t. From a more mechanical perspective, each backwards induction step tends to add complexity. Even if $L_T^*(x_1, \ldots, x_T)$ has some structure, it is unlikely that $\min_{a_T} \max_{x_T} \ell(a_T, x_T) + V(x_1, \ldots, x_T)$ will maintain it.

There are relatively few examples where one can perform the minimax analysis efficiently. For example, consider log loss, first discussed in [45]. While the minimax algorithm, Normalized Maximum Likelihood, is well known [12], it is only efficient in two cases: the multinomial case where fast Fourier transforms may be exploited [29], and very particular exponential families that cause NML to be a Bayesian strategy [26], [5]. The minimax optimal strategy is known also for: (i) the ball game with $\boldsymbol{W} = \boldsymbol{I}$ [50] (our generalization to Mahalanobis $\boldsymbol{W} \neq \boldsymbol{I}$ results in fundamentally different strategies), (ii) the ball game with $\boldsymbol{W} = \boldsymbol{I}$ and a constraint on the player's deviation from the current empirical minimizer [2] (for which the optimal strategy is Follow-the-Leader), (iii) Lipschitz-bounded convex loss functions [2], (iv) experts with an L^* bound [3], (v) static experts with absolute loss [13], and (vi) time series prediction against a smoothness-penalized comparator [31]. While this list may not be exhaustive, hopefully the reader is convinced that tractable minimax algorithms are rare.

1.3 Reinforcement Learning

One powerful tool for reasoning in complex systems is reinforcement learning. However, a lack of computational efficiency prevents a more wide-spread adoption. While recent success such as Google DeepMind demonstrate RL's efficacy, theoretical understanding of optimality guarantees, especially for large scales, is lacking.

Reinforcement learning assumes that the world is a Markov decision process, abbreviated MDP, and tries to learn the MDP while simultaneously performing well. An MDP is parameterized by

1. a discrete state space $\mathcal{X} = \{1, 2, \ldots, X\}$,
2. a discrete action space $\mathcal{A} = \{1, 2, \ldots, A\}$,
3. transition dynamics $P : \mathcal{X} \times \mathcal{A} \to \triangle_{\mathcal{X}}$ that describes the distribution of the next state given a current state and action, and
4. loss function $\ell : \mathcal{X} \times \mathcal{A} \to [0, 1]$ that provides the cost of taking an action in a given state.

For time step t, the learner observes the state x_t, chooses an action a_t, and receives loss $\ell(x_t, a_t)$. The next state is then generated stochastically: $x_{t-1} \sim P(x_t, a_t)$. The (fully observed) state encapsulates all the persistent information of the environment, and the influence of the agent is captured through the transition distribution, which is a function of the current state and the current action. The state evolves in a Markov fashion: x_t is conditionally independent of the past given x_{t-1} and a_{t-1}. A policy π provides a distribution of actions for every state, and the usual goal is to find a good policy.

We focus on an important subproblem to reinforcement learning: planning. The MDP planning problem assumes perfect knowledge of the MDP and tries to find the optimal policy. Even though the planning problem decouples the optimization problem (finding π) from the learning problem (estimating the parameters of the MDP), it is still challenging when $\mathcal{X}$ or $\mathcal{A}$ are large.

Classical techniques such as value iteration [7] or policy iteration [27] solve the planning problem but scale quadratically with X. We restrict the optimization space to arrive at an approximately optimal policy but with complexity scaling with the size of the restriction instead of X. Exploiting the fact that the optimal policy can be written as the solution to a linear program [36], the *approximate linear programming* line of work reduces the computational complexity of the planning problem by approximating solutions to the linear program by only considering solutions on some low-dimensional subspace. The first theoretical guarantees in [19, 17] bounded the suboptimality of the greedy policy corresponding to the best value function representation. Later papers, such as [15, 51], improved the computational burden by e.g. sampling constraints. Unfortunately, these algorithms all require that the optimal policy be near the low-dimensional subspace.

1.4 Contributions

Two chapters of this thesis are devoted to efficient minimax algorithms. In Chapter 2, we study the prediction game. Given some compact set $\mathcal{X} \in \mathbb{R}^d$ and a game length T, each round $t = 1, \ldots, T$ of the prediction game consists of:

1. the learner predicts $\boldsymbol{a}_t \in \mathbb{R}^d$,
2. the adversary observes $\boldsymbol{a}_t$ and responds with $\boldsymbol{x}_t \in \mathcal{X}$, and
3. the learner observes $\boldsymbol{x}_t$ and receives loss $\|\boldsymbol{x}_t - \boldsymbol{a}_t\|_2^2$.

The regret is with respect to $L_T^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) = \min_{\boldsymbol{a} \in \mathcal{X}} \sum_{t=1}^T \|\boldsymbol{x}_t - \boldsymbol{a}\|_2^2$. We derive the minimax algorithm, show it is efficient, and calculate the value. We also extend the analysis to sets $\mathcal{X}$ in Hilbert spaces. This work is a broad generalization of the previous published work [30].

Chapter 3 studies the minimax algorithm for the classic problem of linear regression. For each round $t = 1, \ldots, T$, the game protocol is:

1. the adversary chooses a covariate $\boldsymbol{x}_t \in \mathcal{X}_t \subset \mathbb{R}^d$,
2. the learner predicts a label $\hat{y}_t \in \mathbb{R}$,
3. the adversary observes $\hat{y}_t$ and responds with $y_t \in \mathcal{Y}_t \subset \mathbb{R}$, and
4. the learner observes y_t and receives loss $(\hat{y}_t - y_t)^2$,

where $\mathcal{X}_t$ and $\mathcal{Y}_t$ are some constraints we will specify later (and will be carefully chosen to ensure that the minimax algorithm is computationally efficient). The linear regression is encoded into the regret term. Specifically, we define regret with respect to

$$L_T^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, y_1, \ldots, y_T) = \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^T \|\theta^\intercal \boldsymbol{x}_t - y_t\|_2^2,$$

which is the best linear predictor in hindsight. This protocol emphasizes that the structure of minimax analysis can be entirely derived from the regret term. We shall see that, despite the learner's prediction being unconstrained, the minimax strategy (under certain restrictions on the adversary) is a linear function similar to ridge regression but with a data-dependent regularization. The first part of this chapter was previously published as [6].

Chapter 4 switches to the second topic: planning in large state-space Markov decision problems. Our method builds off the approximate linear programming literature but takes the novel approach of approximating in the dual space of the linear program. The dual LP can be interpreted as optimizing stationary distributions over state action pairs with dual objective equal to the expected loss. We show that approximately solving a low-dimensional LP suffices to provide a policy that is near optimal with respect to the best in the class, even if the class is far from the optimal policy. Thus, the performance bound is non-trivial even if the subspace does not contain a near optimal policy. This work is an extension of the previous published paper [35].

Finally, Chapter 5 summarizes the work and discusses further directions.

Chapter 2

Minimax Squared Loss Prediction

2.1 Introduction

We are interested in general strategies for sequential prediction and decision making (a.k.a. online learning) that improve their performance with experience. We model interaction with the environment as a repeated game where the learner and environment take turns playing actions with full knowledge of the past. We formalize the learning task as a problem of minimizing regret, the difference between the learner's performance and the performance of the best strategy from some fixed reference set. In many cases, we have efficient algorithms with an upper bound on the regret that meets the game-theoretic lower bound (up to a constant factor). In a few special cases, we have the exact minimax strategy, meaning that we understand the learning problem at all levels of detail. In even fewer cases we can also efficiently execute the minimax strategy. These cases serve as exemplars to guide our thinking about learning algorithms.

This chapter explores efficiently computable minimax strategies for the squared Euclidean loss. Our setup, described in Figure 2.1, is as follows. Given a game length T and a set $\mathcal{X} \subset \mathbb{R}^d$, for each round $t = 1, \dots, T$, the learner plays $\boldsymbol{a}_t \in \mathbb{R}^d$, the adversary plays $\boldsymbol{x}_t \in \mathcal{X}$, and then the learner is penalized using the squared Euclidean distance $\|\boldsymbol{a} - \boldsymbol{x}\|^2$. After a sequence of T such interactions, we compare the loss of the learner to the loss of the best fixed prediction $\boldsymbol{a}^* \in \mathbb{R}^d$. In all our examples, this best fixed action in hindsight is the mean outcome $\boldsymbol{a}^* = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t$.

Given: T, $\mathcal{X}$.
For $t = 1, 2, \dots, T$,

- learner chooses prediction $\boldsymbol{a}_t \in \mathbb{R}^d$,
- adversary chooses outcome $\boldsymbol{x}_t \in \mathcal{X}$, and
- learner incurs loss $\|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2$.

Figure 2.1: Protocol

We use regret, the difference between the loss of the learner and the loss of $\boldsymbol{a}^*$, to evaluate performance. The *minimax regret* for the T-round game, also known as the value of the game, is given by

$$V := \inf_{\boldsymbol{a}_1} \sup_{\boldsymbol{x}_1} \cdots \inf_{\boldsymbol{a}_T} \sup_{\boldsymbol{x}_T} \sum_{t=1}^T \|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2 - \inf_{\boldsymbol{a}} \sum_{t=1}^T \|\boldsymbol{a} - \boldsymbol{x}_t\|^2 \tag{2.1}$$

where the $\boldsymbol{a}_t$ range over actions in $\mathbb{R}^d$ and the $\boldsymbol{x}_t$ range over outcomes in $\mathcal{X}$. (Note that the value depends on $\mathcal{X}$ and T, but we omit the dependence from the notation.) The *minimax strategy* chooses the $\boldsymbol{a}_t$, given all past outcomes $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{t-1}$, to achieve this regret, and the *maximin strategy* (of the adversary) chooses $\boldsymbol{x}_t$ given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{t-1}$ and $\boldsymbol{a}_t$. Equation (2.1) can equivalently be thought of as

$$V = \inf_{\substack{\text{Player}\\\text{Strategies}}} \sup_{\substack{\text{Environment}\\\text{Strategies}}} \sum_{t=1}^{T} \|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2 - \inf_{\boldsymbol{a}} \sum_{t=1}^{T} \|\boldsymbol{a} - \boldsymbol{x}_t\|^2$$

with the causal dependencies of $\boldsymbol{a}_t$ and $\boldsymbol{x}_t$ on the past made explicit. Hence, the minimax value is the best response to the worst case.

The contributions of the paper can be succinctly summarized as identifying constraints on $\mathcal{X}$ that allow efficient computation of (2.1) and the minimax strategy. We proceed by first solving two special cases in explicit detail, then generalizing.

Special Cases

Our goal is to derive the minimax strategy for a general compact $\mathcal{X} \in \mathbb{R}^d$. Our previous work [30] explicitly computed the minimax strategy in two special cases: the Brier game, where the action and outcome spaces are the probability simplex with K outcomes, and the Mahalanobis loss game, where the action and outcome spaces are the Euclidean norm ball. The former is traditionally popular in meteorology [10] and the latter has connections with online Gaussian density estimation [50].

We generalize these results to polytopes in general position (Section 2.3) and to arbitrary ellipsoids (Section 2.4). Recall that a collection of points in $\mathbb{R}^d$ is in general position if, for any $k \leq d+1$, no $(k-2)$-dimensional flat (i.e., a subspace with an offset) contains k of the points.

Definition 1 (simplex). *For a collection of $k \leq d+1$ points in general position, $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k \in \mathbb{R}^d$, we use $\boldsymbol{Z}$ to denote both the matrix with columns $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k$ as well as the collection of points. We define the $\boldsymbol{Z}$-*simplex *as the convex hull of these points,*

$$\triangle_{\boldsymbol{Z}} := \left\{ \sum_{i=1}^{k} p_k \boldsymbol{z}_k : p_1, \dots, p_k \geq 0, \mathbf{1}^\intercal \boldsymbol{p} = 1 \right\}. \tag{2.2}$$

We say that $\triangle_{\boldsymbol{Z}}$ is of size k. Use $\triangle_k$ to denote the probability simplex over k actions,

$$\triangle_k := \{\boldsymbol{p} : p_1, \dots, p_k \geq 0, \mathbf{1}^\intercal \boldsymbol{p} = 1\}.$$

It is easy to see that $\triangle_{\boldsymbol{Z}}$ is exactly the image of $\triangle_k$ through the linear map $\boldsymbol{Z}$.

In particular, the $k = d+1$ case corresponds to the standard definition of a simplex that is affinely independent in $\mathbb{R}^d$; that is, the polytope cannot be translated to lie in some low dimensional subspace. Also, the notion of general position is equivalent to requiring that

$$\operatorname{Rank}\left(\begin{bmatrix} \boldsymbol{Z} \\ \mathbf{1}^\intercal \end{bmatrix} \right) = k.$$

For a positive semi-definite matrix $\boldsymbol{W}$, we define the Mahalanobis norm as

$$\|\boldsymbol{x}\|_{\boldsymbol{W}}^2 := \boldsymbol{x}^\intercal \boldsymbol{W}^\dagger \boldsymbol{x}, \tag{2.3}$$

where $\boldsymbol{W}^{\dagger}$ is the Moore-Penrose Pseudo inverse. Without loss of generality, we assume that $\boldsymbol{W}$ is symmetric.

The unit ball in this norm is an ellipsoid.

Definition 2. *A positive semi-definite, symmetric $\boldsymbol{W}$ and center $\boldsymbol{z} \in \mathbb{R}^d$ define the ellipsoid*

$$B(\boldsymbol{z}, \boldsymbol{W}) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x} - \boldsymbol{z}\|_W \leq 1\} \tag{2.4}$$

with centered case $\bigcirc_{\boldsymbol{W}} \coloneqq B(0, \boldsymbol{W})$.

Note that the ellipsoid $\bigcirc_{\boldsymbol{W}}$ is the image of $\bigcirc \coloneqq \bigcirc_{\boldsymbol{I}}$ under the linear map $\boldsymbol{W}^{\frac{1}{2}\dagger}$. Instead of Euclidean loss on Mahalanobis balls, we could have equivalently played with Mahalanobis losses on Euclidean balls by transforming $\boldsymbol{x}$ to $W^{\frac{1}{2}\dagger}\boldsymbol{x}$; we shall later see that affine transformations do not change the regret.

Bridging the Gap

One of the main results of this chapter is that ellipsoid games and certain simplex games have the same minimax strategies and the same regret. The key geometric quantity to study is the smallest enclosing ball.

Definition 3 (smallest enclosing ball)**.** *For a compact set $\mathcal{X}$, the point $\boldsymbol{c}_{\mathcal{X}} \in \mathcal{X}$ and scalar $\rho_{\mathcal{X}} \in \mathbb{R}$ are defined to be the center and radius of the Euclidean ball with the minimum radius that contains $\mathcal{X}$, that is, $(\boldsymbol{c}_{\mathcal{X}}, \rho_{\mathcal{X}})$ is the solution to the optimization problem*

$$\begin{aligned} \min_{\boldsymbol{c},\rho} \quad & \rho \\ \textit{s.t.} \quad & \mathcal{X} \subseteq B(\boldsymbol{c}, \rho), \end{aligned}$$

where $B(\boldsymbol{c}, \rho)$ denotes the Euclidean ball,

$$B(\boldsymbol{c}, \rho) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d \mid \|x - c\| \leq \rho\}.$$

We shall show that the regret for any compact set is the same as the regret played on the smallest enclosing ball. Intuitively, the adversary only chooses points maximally far apart and hence plays on the intersection of the boundary of $\mathcal{X}$ and the smallest enclosing ball. The proof is by a sandwich argument: every compact set contains a simplex (possibly in a lower dimensional subspace), is contained in a Euclidean ball, and the regret of the games played on these two sets match. More precisely, we prove the following theorem.

Theorem 4. *Let $\mathcal{X}$ be a compact set and let $\boldsymbol{c}_{\mathcal{X}}$ and $\rho_{\mathcal{X}}$ be the center and radius of the smallest Euclidean ball that contains $\mathcal{X}$. For $n = 2, \ldots, T$, recursively define the coefficients $\alpha_T = 1/T$ and $\alpha_{n-1} = \alpha_n + \alpha_n^2$. Then the squared loss game has value*

$$V = \rho_{\mathcal{X}}^2 \sum_{n=1}^{T} \alpha_n, \tag{2.5}$$

which is achieved by the minimax strategy

$$\boldsymbol{a}_n = \boldsymbol{c}_{\mathcal{X}} + \alpha_n \sum_{t=1}^{n-1} (\boldsymbol{x}_t - \boldsymbol{c}_{\mathcal{X}}) \,. \tag{MM}$$

Hilbert Space

One may notice that the value of the game on a compact set $\mathcal{X}$ only depends on the radius of the smallest enclosing ball and some function of the game length; in particular, there is no dimension dependence. This invites the question, can we play in an infinite dimensional space with the same regret and, if so, is (MM) the minimax strategy? We will show that the answer to both questions is yes. We show that (MM) achieves the same regret (2.5) in Hilbert space, and show that this is optimal using a limiting argument from the finite case, by constructing a sequence of finite subsets with a minimum enclosing ball whose radius approaches that of $\mathcal{X}$.

Related Work

Repeated games with minimax strategies are frequently studied [12] and, in online learning, minimax analysis has been applied to a variety of losses and repeated games; however, computationally feasible algorithms are the exception, not the rule. For example, the minimax algorithm for log loss, Normalized Maximum Likelihood (NML), is well known [45], but it generally requires computation that is exponential in the time horizon, as one needs to aggregate over all data sequences. To our knowledge, there are two exceptions where efficient NML forecasters are possible: the multinomial case where fast Fourier transforms may be exploited [29], and very particular exponential families that cause NML to be a Bayesian strategy [26, 5]. The minimax optimal strategy is known also for: (i) the ball game with $\boldsymbol{W} = \boldsymbol{I}$ [50] (our generalization to Mahalanobis balls with $\boldsymbol{W} \neq \boldsymbol{I}$ results in fundamentally different strategies), (ii) the ball game with $\boldsymbol{W} = \boldsymbol{I}$ and a constraint on the player's deviation from the current empirical minimizer [2] (for which the optimal strategy is Follow-the-Leader), (iii) Lipschitz-bounded convex loss functions [2], (iv) experts with an L^* bound [3], (v) static experts with absolute loss [13], (vi) unconstrained linear optimization [37], and (vii) linear regression [6]. This close-to-exhaustive list demonstrates the rarity of tractable minimax algorithms.

The Ball game was considered previously by [50], who identify the minimax algorithm for the special case $\boldsymbol{W} = \boldsymbol{I}$. The generalization to $\boldsymbol{W} \neq \boldsymbol{I}$ results in fundamentally different strategies.

Outline

The paper is organized as follows. Section 2.2 begins with describing the minimax framework and introducing the tool of backwards induction as a means of computing the minimax strategy. We then apply this tool and fully solve the simplex and ellipsoid games in Sections 2.3 and 2.4, presenting the value, the minimax strategy, and the optimal strategy of the adversary. We then generalize to an arbitrary compact set in Section 2.5 by first proving several necessary properties of the smallest enclosing ball for convex sets and then deriving the minimax strategy and value using a sandwich argument. Section 2.6 then shows that the regret and minimax strategy remain essentially unchanged in the infinite dimensional case. We conclude in Section 2.7.

2.2 Value-to-go

In this section, we introduce the value-to-go function and some common elements of the ball and Brier games. All games will be specified by a game length T and a constraint on the adversary's action space $\mathcal{X} \subset \mathbb{R}^d$. For simplicity, we omit this dependence from our notation. For some observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the *value-to-go* for the remaining $T - n$ rounds is given by

$$V(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \;:=\; \inf_{\boldsymbol{a}_{n+1}} \sup_{\boldsymbol{x}_{n+1}} \cdots \inf_{\boldsymbol{a}_T} \sup_{\boldsymbol{x}_T} \sum_{t=n+1}^{T} \|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2 - \inf_{\boldsymbol{a}} \sum_{t=1}^{T} \|\boldsymbol{a} - \boldsymbol{x}_t\|^2,$$

where $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_T$ are constrained to be in $\mathcal{X}$ but $\boldsymbol{a}_{n_1}, \ldots, \boldsymbol{a}_T$ are unconstrained (although we will later see that the learner will never play outside of the convex hull of $\mathcal{X}$). The value-to-go at $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ encapsulates the effective future regret if both players play optimally starting with a history $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

By definition, the minimax regret (2.1) is $V = V(\epsilon)$ where ϵ is the empty sequence, and an easy induction shows that the value-to-go satisfies the recurrence

$$V(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \;=\; \begin{cases} -\inf_{\boldsymbol{a}} \sum_{t=1}^{T} \|\boldsymbol{a} - \boldsymbol{x}_t\|^2 & \text{if } n = T, \\ \inf_{\boldsymbol{a}_{n+1}} \sup_{\boldsymbol{x}_{n+1}} \|\boldsymbol{a}_{n+1} - \boldsymbol{x}_{n+1}\|^2 + V(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n+1}) & \text{if } n < T. \end{cases} \tag{2.6}$$

Our analysis for the two games proceeds in a similar manner. For some past history of plays $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ of length n, we summarize the state by $\boldsymbol{s} = \sum_{t=1}^{n} \boldsymbol{x}_t$ and $\sigma^2 = \sum_{t=1}^{n} \boldsymbol{x}_t^\intercal \boldsymbol{x}_t$. As we will see, the value-to-go after n of T rounds can be written as $V(\boldsymbol{s}, \sigma^2, n)$; i.e. it only depends on the past plays through $\boldsymbol{s}$ and σ^2. More surprisingly, for each n, the value-to-go $V(\boldsymbol{s}, \sigma^2, n)$ is a quadratic function of $\boldsymbol{s}$ and a linear function of σ^2 (under certain conditions on $\mathcal{X}$).

It is easy to see that the terminal value $V(\boldsymbol{s}, \sigma^2, T)$ is quadratic in the state; this is simply a consequence of the form of the loss of the best action in hindsight, as described in Lemma 5. However, it is not at all obvious that propagating from $V(\boldsymbol{s} + \boldsymbol{x}, \sigma^2 + \boldsymbol{x}^\intercal \boldsymbol{x}, n + 1)$ to $V(\boldsymbol{s}, \sigma^2, n)$ by using the second case of (2.6) preserves this structure. This compact representation of the value function is an essential ingredient for a computationally feasible algorithm. In many regret minimization games, the minimax strategy has a computational complexity that apparently scales exponentially with the time horizon. We show that, for squared Euclidean loss, the minimax strategy can be computed in constant amortized time.

The Offline Problem The regret is defined as the difference between the loss of the algorithm and the loss of the best action in hindsight. Here we calculate that action and its loss.

Lemma 5. *For data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathcal{X}$, the loss of the best action in hindsight equals*

$$\inf_{\boldsymbol{a} \in \mathbb{R}^d} \sum_{t=1}^{T} \|\boldsymbol{a} - \boldsymbol{x}_t\|^2 \;=\; \sum_{t=1}^{T} \boldsymbol{x}_t^\intercal \boldsymbol{x}_t - \frac{1}{T} \left(\sum_{t=1}^{T} \boldsymbol{x}_t \right)^\intercal \left(\sum_{t=1}^{T} \boldsymbol{x}_t \right), \tag{2.7}$$

and the minimizer is the mean outcome $\boldsymbol{a}^* = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t$.

Proof. The unconstrained minimizer and value are obtained by equating the derivative to zero and plugging in the solution. □

The best action in hindsight is independent of $\mathcal{X}$, and therefore the follow-the-leader strategy is as well. As we shall see, the minimax strategy does not have this property.

2.3 The Simplex Game

This section analyzes the squared loss game when the adversary is restricted to play on a general simplex, which can be thought of as a non-degenerate (k)-vertex polytope in $\mathbb{R}^d$. We first define a simplex and connect the Euclidean loss game on the simplex with a Mahalanobis loss game on the probability simplex and prove several useful lemmas for optimizing quadratic functions in the simplex. Building off of this, in Section 2.3, we evaluate the value-to-go function of the simplex game and present the minimax and maximin strategies. Finally, we show a regret bound.

Simplex Preliminaries

Recall that the $\boldsymbol{Z}$-simplex is the convex hull of the columns of $\boldsymbol{Z}$, which are in general position. We require that $\boldsymbol{Z}$ be in general position because it allows a short description of the mapping between $\triangle_{\boldsymbol{Z}}$ and $\triangle_k$ that relies on the offset vector, defined below. We will use $\boldsymbol{G} := \boldsymbol{Z}^\intercal \boldsymbol{Z}$ to denote the Gram matrix of $\boldsymbol{Z}$ and $\boldsymbol{g} := \operatorname{diag}(\boldsymbol{G})$ to denote its diagonal.

Definition 6. *For the $\boldsymbol{Z}$-simplex, define the* offset vector *to be any vector $\boldsymbol{v}$ satisfying $\boldsymbol{Z}\boldsymbol{v} = 0$ and $\boldsymbol{1}^\intercal \boldsymbol{v} = 1$.*

Note that such a $\boldsymbol{v}$ always exists: since $\boldsymbol{Z}$ are in general position, $\boldsymbol{Z}$ can have rank at most d and the null space cannot be orthogonal to $\boldsymbol{1}$. In the case when $k = d + 1$, $\begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{1}^\intercal \end{bmatrix}$ has full rank and $\boldsymbol{v}$ is uniquely determined.

We can think of $\boldsymbol{v}$ as the point in $\triangle_k$ that maps to the origin, provided that $\triangle_{\boldsymbol{Z}}$ contains the origin (otherwise, $\boldsymbol{v}$ will have negative entries but will lie in the hyperplane containing $\triangle_k$). The vector $\boldsymbol{v}$ affords us an easily described affine transformation between $\triangle_{\boldsymbol{Z}}$ and $\triangle_k$.

Proposition 7. *Given the $\boldsymbol{Z}$-simplex $\triangle_{\boldsymbol{Z}}$, the linear map $\boldsymbol{Z}$ maps from $\triangle_k$ onto $\triangle_{\boldsymbol{Z}}$. For any $\boldsymbol{z} \in \triangle_{\boldsymbol{Z}}$, the affine map*

$$\triangledown(\boldsymbol{z}) := \boldsymbol{v} + \boldsymbol{S}\boldsymbol{z} \tag{2.8}$$

guarantees that $\boldsymbol{p} = \triangledown(\boldsymbol{z}) \in \triangle_k$ is a solution to $\boldsymbol{Z}\boldsymbol{p} = \boldsymbol{z}$, where $\boldsymbol{v}$ is the offset vector and

$$\boldsymbol{S} \coloneqq \left(\boldsymbol{I} - \boldsymbol{v}\mathbf{1}^{\intercal}\right)\boldsymbol{Z}^{\dagger}. \tag{2.9}$$

Additionally, $\boldsymbol{Z}\boldsymbol{S} = \boldsymbol{I}$.

Proof. Let $\boldsymbol{z} \in \triangle_{\boldsymbol{Z}}$, and we want to find a p such that $\boldsymbol{Z}\boldsymbol{p} = \boldsymbol{z}$. By definition of $\triangle_{\boldsymbol{Z}}$, $\boldsymbol{z}$ must be in the column space of $\boldsymbol{Z}$, and hence we can take $\boldsymbol{p} = \boldsymbol{Z}^{\dagger}\boldsymbol{z} + \alpha\boldsymbol{v}$ for any α (since $\boldsymbol{v}$ is in the null space of $\boldsymbol{Z}$). Solving for the α that requires $\mathbf{1}^{\intercal}\boldsymbol{p} = 1$ yields

$$\mathbf{1}^{\intercal}\left(\boldsymbol{Z}^{\dagger}\boldsymbol{z} + \alpha\boldsymbol{v}\right) = 1 \Leftrightarrow \alpha = 1 - \mathbf{1}^{\intercal}\boldsymbol{Z}^{\dagger}\boldsymbol{z}$$

since $\mathbf{1}^{\intercal}\boldsymbol{v} = 1$. Hence,

$$\begin{aligned} \boldsymbol{p} &= \boldsymbol{Z}^{\dagger}\boldsymbol{z} + \left(1 - \mathbf{1}^{\intercal}\boldsymbol{Z}^{\dagger}\boldsymbol{z}\right)\boldsymbol{v} \\ &= \left(\boldsymbol{I} - \boldsymbol{v}\mathbf{1}^{\intercal}\right)\boldsymbol{Z}^{\dagger}\boldsymbol{z} + \boldsymbol{v}, \end{aligned}$$

and $\boldsymbol{S} = \left(\boldsymbol{I} - \boldsymbol{v}\mathbf{1}^{\intercal}\right)\boldsymbol{Z}^{\dagger}$ as claimed.

The last claim of the theorem is easily checked: $\boldsymbol{Z}\boldsymbol{S} = \boldsymbol{Z}\left(\boldsymbol{I} - \boldsymbol{v}\mathbf{1}^{\intercal}\right)\boldsymbol{Z}^{\dagger} = \boldsymbol{Z}\boldsymbol{Z}^{\dagger} = \boldsymbol{I}$. □

Minimum Enclosing Ball of a Simplex

One of the biggest reasons for working with simplices is that the minimum enclosing ball is easy to compute in closed form, as described in the following lemma, where $\boldsymbol{g} = \operatorname{diag}(\boldsymbol{Z}^{\intercal}\boldsymbol{Z})$ and $\boldsymbol{S}$ is as defined in Proposition 7.

Lemma 8. *Let $\boldsymbol{Z}$ be in general position and $\triangle_{\boldsymbol{Z}}$ be the corresponding simplex. The center and radius of the minimum enclosing ball are*

$$\boldsymbol{c}_{\boldsymbol{Z}} \coloneqq \frac{1}{2}\boldsymbol{S}^{\intercal}\boldsymbol{g}, \quad \text{and} \quad \rho_{\boldsymbol{Z}}^2 \coloneqq \boldsymbol{c}_{\boldsymbol{Z}}^{\intercal}\boldsymbol{c}_{\boldsymbol{Z}} + \boldsymbol{g}^{\intercal}\boldsymbol{v} \tag{2.10}$$

provided that $\boldsymbol{S}\boldsymbol{c}_{\boldsymbol{Z}} + \boldsymbol{v} \succeq 0$, where $\succeq$ denotes an entry-wise inequality between vectors.

The proof of Lemma 8 builds off Lemma 10 and will be presented shortly. We state Lemma 8 to highlight an important condition:

Definition 9. *A simplex $\triangle_{\boldsymbol{Z}}$ is* ball-like *if*

$$\boldsymbol{S}\boldsymbol{c}_{\boldsymbol{Z}} + \boldsymbol{v} \succeq 0. \tag{2.11}$$

The ball-like property holds only if the minimum enclosing ball touches the simplex at every vertex, and allows easy computation of the minimum enclosing ball via (2.10).

We will drop the subscripts on $\boldsymbol{c}$ and ρ^2 when the context is clear.

Optimizing Quadratic Functions on the Simplex

Before we prove Lemma 8 and compute the value-to-go, we will derive a useful result about optimizing quadratic functions over $\triangle_{\boldsymbol{Z}}$.

Lemma 10. *The optimization*

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & -\boldsymbol{p}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{Z} \boldsymbol{p} + 2\boldsymbol{d}^\intercal \boldsymbol{p} \\ \textit{s.t.} \quad & \mathbf{1}^\intercal \boldsymbol{p} = 1. \end{aligned}$$

has solution

$$\boldsymbol{p} = \boldsymbol{v} + \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d}$$

and value

$$\boldsymbol{d}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d} + 2\boldsymbol{d}^\intercal \boldsymbol{v}.$$

Proof. We can equivalently solve the optimization for $\boldsymbol{p}$ or $\boldsymbol{z} = \boldsymbol{Z}\boldsymbol{p}$, since $\boldsymbol{p} \mapsto \boldsymbol{z}$ is a one-to-one mapping between $\{\boldsymbol{p} : \mathbf{1}^\intercal \boldsymbol{p} = 1\}$ and $\mathbb{R}^d$. Using $\boldsymbol{p} = \boldsymbol{v} + \boldsymbol{S}\boldsymbol{z}$ and $\boldsymbol{z} = \boldsymbol{Z}\boldsymbol{p}$, the optimization is equal to

$$\begin{aligned} -\boldsymbol{p}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{Z} \boldsymbol{p} + 2\boldsymbol{d}^\intercal \boldsymbol{p} &= -\boldsymbol{z}^\intercal \boldsymbol{z} + 2\boldsymbol{d}^\intercal \boldsymbol{S}\boldsymbol{z} + 2\boldsymbol{d}^\intercal \boldsymbol{v} \\ &= -\left(\boldsymbol{z} - \boldsymbol{S}^\intercal \boldsymbol{d}\right)^\intercal \left(\boldsymbol{z} - \boldsymbol{S}^\intercal \boldsymbol{d}\right) + \boldsymbol{d}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d} + 2\boldsymbol{d}^\intercal \boldsymbol{v} \end{aligned}$$

We can therefore read off the solution as $\boldsymbol{z} = \boldsymbol{S}^\intercal \boldsymbol{d}$ and the value as

$$\boldsymbol{d}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d} + 2\boldsymbol{d}^\intercal \boldsymbol{v}.$$

Mapping this back, we find that the solution is

$$\boldsymbol{p} = \boldsymbol{v} + \boldsymbol{S}\boldsymbol{z} = \boldsymbol{v} + \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d}.$$

□

With this lemma, we can immediately prove the smallest ball theorem.

Proof of Lemma 8. The smallest ball containing $\triangle_{\boldsymbol{Z}}$ has center $\boldsymbol{c}$ and radius ρ that are solutions to the optimization problem

$$\begin{aligned} \min_{\rho, \boldsymbol{c} \in \triangle_{\boldsymbol{Z}}} \quad & \rho^2 \\ \text{s.t.} \quad & \|\boldsymbol{z}_i - \boldsymbol{c}\|^2 - \rho^2 \leq 0 \quad \forall i = 1, \ldots, k \end{aligned}$$

Using $\boldsymbol{p}$ as the vector of dual variables, the Lagrangian is then

$$\mathcal{L}(\boldsymbol{c}, \rho, \boldsymbol{p}) = \sum_{i=1}^{k} p_i \|\boldsymbol{z}_i - \boldsymbol{c}\|^2 + \rho^2 \left(1 - \sum_{i=1}^{k} p_i\right).$$

If we differentiate with respect to $\boldsymbol{c}$ and set this equal to zero, we must have $\boldsymbol{c} = \sum_i p_i \boldsymbol{z}_i$; using this, we are left with

$$\sum_{i=1}^{k} p_i \|\boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{p}\|^2 + \rho^2 \left(1 - \sum_{i=1}^{k} p_i\right)$$

under the further constraint that $p_i \geq 0$. We can rewrite the first term as

$$\begin{aligned} \sum_i p_i \|\boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{p}\|^2 &= \sum_i p_i \, (\boldsymbol{e}_i - \boldsymbol{p})^\intercal \boldsymbol{Z}^\intercal \boldsymbol{Z} (\boldsymbol{e}_i - \boldsymbol{p}) \\ &= \sum_i p_i \, (\boldsymbol{e}_i^\intercal \boldsymbol{G} \boldsymbol{e}_i - 2\boldsymbol{p}^\intercal \boldsymbol{G} \boldsymbol{e}_i + \boldsymbol{p}^\intercal \boldsymbol{G} \boldsymbol{p}) \\ &= \boldsymbol{g}^\intercal \boldsymbol{p} - \boldsymbol{p}^\intercal \boldsymbol{G} \boldsymbol{p}, \end{aligned}$$

and thus we see that the least-norm minimum ball is equivalent to the optimization problem

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \boldsymbol{g}^\intercal \boldsymbol{p} - \boldsymbol{p}^\intercal \boldsymbol{G} \boldsymbol{p} \\ \text{s.t.} \quad & \boldsymbol{p} \in \triangle_k. \end{aligned} \tag{2.12}$$

Applying Lemma 10 with $\boldsymbol{d} = \boldsymbol{g}/2$ gives

$$\boldsymbol{p}^* = \boldsymbol{v} + \frac{1}{2} \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{g},$$

which implies that

$$\boldsymbol{c}_{\boldsymbol{Z}} = \boldsymbol{Z}\left(\boldsymbol{v} + \boldsymbol{S}\boldsymbol{S}^\intercal \frac{\boldsymbol{g}}{2}\right) = \frac{\boldsymbol{S}^\intercal \boldsymbol{g}}{2}$$

and

$$\rho_Z = \frac{1}{4} \boldsymbol{g}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{g} + \boldsymbol{g}^\intercal \boldsymbol{v}.$$

However, Lemma 10 solves the optimization problem where $\boldsymbol{1}^\intercal \boldsymbol{p} = 1$ is the only constraint. If the solution given by the Lemma also happens to satisfy $p_i \geq 0$, then the solution must also be correct on the subset $\boldsymbol{p} \in \triangle_k$. That is, for the $\boldsymbol{c}_{\boldsymbol{Z}}$ above to be correct, it suffices if

$$\boldsymbol{v} + \boldsymbol{S}\boldsymbol{c}_{\boldsymbol{Z}} \in \triangle_k,$$

which is exactly equivalent to the ball-like condition. □

We need one final ingredient before we can evaluate the value of the game: namely, we need to understand how saddle point problems over $\triangle_{\boldsymbol{Z}}$ behave.

Lemma 11. *For a simplex $\triangle_{\boldsymbol{Z}}$, vector $\boldsymbol{b}$ and constant $\beta \geq 1$, the optimization problem*

$$\min_{\boldsymbol{a} \in \mathbb{R}^d} \max_{\boldsymbol{x} \in \triangle_{\boldsymbol{Z}}} \|\boldsymbol{a} - \boldsymbol{x}\|^2 + \beta \boldsymbol{x}^\intercal \boldsymbol{x} + 2\boldsymbol{b}^\intercal \boldsymbol{x}$$

achieves its value

$$\boldsymbol{d}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d} + 2\boldsymbol{d}^\intercal \boldsymbol{v} = (1+\beta)^2 \boldsymbol{c}^\intercal \boldsymbol{c} + 2(1+\beta)\boldsymbol{c}^\intercal \boldsymbol{b} + \boldsymbol{b}^\intercal \boldsymbol{b} + (1+\beta)\boldsymbol{v}^\intercal \boldsymbol{g}$$

where

$$\boldsymbol{d} = \frac{1+\beta}{2}\boldsymbol{g} + \boldsymbol{Z}^\intercal \boldsymbol{b}.$$

This is accomplished by the saddle point defined by the player strategy

$$\boldsymbol{a}^* = \boldsymbol{S}^\intercal \boldsymbol{d}$$

and the randomized adversary strategy (the maximin strategy randomizes, playing $\boldsymbol{x} = \boldsymbol{z}_i$ with probability p_i^*)

$$\boldsymbol{p}^* = \boldsymbol{v} + \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d}$$

provided $\boldsymbol{p}^* \succeq \boldsymbol{0}$. Note that the expected value of the maximin strategy is precisely $\boldsymbol{a}^*$.

Proof. The objective is convex in $\boldsymbol{x}$ for each $\boldsymbol{a}$ as $\beta \geq 1$, so it is maximized at a corner $\boldsymbol{x} = \boldsymbol{z}_k$. The value is unchanged by allowing the adversary to play linear combinations. That is,

$$\begin{aligned}
\min_{\boldsymbol{a}\in\triangle_{\boldsymbol{Z}}} \max_{\boldsymbol{x}\in\triangle_{\boldsymbol{Z}}} \|\boldsymbol{a}-\boldsymbol{x}\|^2 + \beta \boldsymbol{x}^\intercal \boldsymbol{x} + 2\boldsymbol{b}^\intercal \boldsymbol{x} &= \min_{\boldsymbol{a}\in\triangle_{\boldsymbol{Z}}} \max_{k} \|\boldsymbol{a}-\boldsymbol{z}_k\|^2 + \beta \boldsymbol{z}_k^\intercal \boldsymbol{z}_k + 2\boldsymbol{b}^\intercal \boldsymbol{z}_k \\
&= \max_{\boldsymbol{p}\in\triangle_{d+1}} \min_{\boldsymbol{a}\in\triangle_{\boldsymbol{Z}}} \mathop{\mathbb{E}}_{k\sim\boldsymbol{p}} \left[\|\boldsymbol{a}-\boldsymbol{z}_k\|^2 + \beta \boldsymbol{z}_k^\intercal \boldsymbol{z}_k + 2\boldsymbol{b}^\intercal \boldsymbol{z}_k\right]
\end{aligned}$$

where the second line follows from a straightforward application of a min-max swap (see e.g. [46]).

The properness of the squared loss implies that the minimum of $\mathbb{E}_{k\sim\boldsymbol{p}}\|\boldsymbol{a}-\boldsymbol{z}_k\|^2$ is at $\boldsymbol{a} = \mathbb{E}_{k\sim\boldsymbol{p}}[\boldsymbol{z}_k] = \boldsymbol{Z}\boldsymbol{p}$; plugging this in yields

$$\begin{aligned}
\min_{\boldsymbol{a}\in\triangle_{\boldsymbol{Z}}} \max_{\boldsymbol{x}\in\triangle_{\boldsymbol{Z}}} \|\boldsymbol{a}-\boldsymbol{x}\|^2 + \beta \boldsymbol{x}^\intercal \boldsymbol{x} + 2\boldsymbol{b}^\intercal \boldsymbol{x} &= \max_{\boldsymbol{p}} \mathop{\mathbb{E}}_{k\sim\boldsymbol{p}} \left[\|\boldsymbol{Z}\boldsymbol{p}-\boldsymbol{z}_k\|^2 + \beta \boldsymbol{z}_k^\intercal \boldsymbol{z}_k + 2\boldsymbol{b}^\intercal \boldsymbol{z}_k\right] \\
&= \max_{\boldsymbol{p}} -\boldsymbol{p}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{Z}\boldsymbol{p} + \left((1+\beta)\operatorname{diag}\left(\boldsymbol{Z}^\intercal \boldsymbol{Z}\right) + 2\boldsymbol{Z}^\intercal \boldsymbol{b}\right)^\intercal \boldsymbol{p}
\end{aligned}$$

We immediately find $\boldsymbol{p}^*$ by applying Lemma 10 with the linear term equal to $\boldsymbol{d} = \frac{1+\beta}{2}\boldsymbol{g} + \boldsymbol{Z}^\intercal \boldsymbol{b}$, which can be written as

$$\begin{aligned}
\boldsymbol{p}^* &= \boldsymbol{v} + \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d} \\
&= \boldsymbol{v} + \boldsymbol{S}\boldsymbol{S}^\intercal \left(\frac{1+\beta}{2}\boldsymbol{g} + \boldsymbol{Z}^\intercal \boldsymbol{b}\right).
\end{aligned}$$

Lemma 10 also provides the value

$$\begin{aligned}
\boldsymbol{d}^\intercal \boldsymbol{p} &= \boldsymbol{d}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d} + 2\boldsymbol{d}^\intercal \boldsymbol{v} \\
&= \frac{(1+\beta)^2}{4}\boldsymbol{g}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{g} + (1+\beta)\boldsymbol{g}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{b} + \boldsymbol{b}^\intercal \boldsymbol{Z}\boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{b} + (1+\beta)\boldsymbol{v}^\intercal \boldsymbol{g} \\
&= \frac{(1+\beta)^2}{4}\boldsymbol{g}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{g} + (1+\beta)\boldsymbol{g}^\intercal \boldsymbol{S}\boldsymbol{b} + \boldsymbol{b}^\intercal \boldsymbol{b} + (1+\beta)\boldsymbol{v}^\intercal \boldsymbol{g}.
\end{aligned}$$

Finally, the optimal player action is

$$\boldsymbol{a}^* = \boldsymbol{Z}\boldsymbol{p}^* = \boldsymbol{S}^\intercal \boldsymbol{d}.$$

□

Minimax Analysis of the Simplex Game

With the above lemmas, we can readily compute $V(\boldsymbol{s}, \sigma^2, n)$ as a recursion and specify the minimax and maximin strategies.

```
Input: Game length $T$, ball-like simplex $\triangle_{\boldsymbol{Z}}$
Calculate the smallest enclosing ball center $\boldsymbol{c}$.
Calculate
    $\alpha_T = \frac{1}{T}, \qquad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1},$
Set $\boldsymbol{s} = 0$
for all $n = 1, \dots, T$ do
    Play $\boldsymbol{a}^* = \boldsymbol{c} + \alpha_n \sum_{t=1}^{n-1} (\boldsymbol{x}_t - \boldsymbol{c})$
end for
```

Figure 2.2: Minimax $\triangle_{\boldsymbol{Z}}$ strategy, $\|\cdot\|^2$ loss

Theorem 12. *Consider the T-round game over a ball-like simplex $\triangle_{\boldsymbol{Z}}$ and Euclidean loss $\|\boldsymbol{a} - \boldsymbol{x}\|^2$. After n outcomes $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ with statistics $\boldsymbol{s} = \sum_{t=1}^n \boldsymbol{x}_t$ and $\sigma^2 = \sum_{t=1}^n \boldsymbol{x}_t^\intercal \boldsymbol{x}_t$, the value-to-go is*

$$V(\boldsymbol{s}, \sigma^2, n) \;=\; \alpha_n \boldsymbol{s}^\intercal \boldsymbol{s} - \sigma^2 + (1 - n\alpha_n)\boldsymbol{g}^\intercal \boldsymbol{S}\boldsymbol{s} + \gamma_n,$$

the maximin strategies plays point $\boldsymbol{z}_k$ proportional to

$$\boldsymbol{p}^* = \boldsymbol{v} + \boldsymbol{S}\left((1 - (n-1)\alpha_n)\,\boldsymbol{c} + \alpha_n(n-1)\frac{\boldsymbol{s}}{n-1}\right)$$

and the minimax strategy is

$$\boldsymbol{a}^* = \boldsymbol{Z}\boldsymbol{p}^* = \boldsymbol{c} + \alpha_n \sum_{t=1}^{n-1} (\boldsymbol{x}_t - \boldsymbol{c})$$

where the α_n coefficients are defined in Figure 2.2 and $\gamma_T = 0$ with

$$\gamma_{n-1} = \gamma_n + (1 - (n-1)\alpha_n)^2\, \boldsymbol{c}^\intercal \boldsymbol{c} + \alpha_n \boldsymbol{v}^\intercal \boldsymbol{g}.$$

Proof. We proceed by induction. Recall that $V(\boldsymbol{s}, \sigma^2, T)$, the value at the end of the game is

$$V(\boldsymbol{s}, \sigma^2, T) \;=\; \frac{1}{T}\boldsymbol{s}^\intercal \boldsymbol{s} - \sigma^2,$$

corresponding to $\alpha_T = \frac{1}{T}$ and $\gamma_T = 0$. Now, assume the induction hypothesis for $n \leq T$ rounds. To check the $n-1$ case, we evaluate

$$\begin{aligned}
V(\boldsymbol{s}, \sigma^2, n-1) &= \min_{\boldsymbol{a}\in\mathbb{R}^d} \max_{\boldsymbol{x}\in\triangle_{\boldsymbol{Z}}} \|\boldsymbol{a}-\boldsymbol{x}\|^2 + \alpha_n(\boldsymbol{s}+\boldsymbol{x})^\intercal(\boldsymbol{s}+\boldsymbol{x}) \\
&\quad - (\sigma^2 + \boldsymbol{x}^\intercal\boldsymbol{x}) + (1-n\alpha_n)\boldsymbol{g}^\intercal\boldsymbol{S}(\boldsymbol{s}+\boldsymbol{x}) + \gamma_n. \\
&= \min_{\boldsymbol{a}\in\mathbb{R}^d} \max_{\boldsymbol{x}\in\triangle_{\boldsymbol{Z}}} \|\boldsymbol{a}-\boldsymbol{x}\|^2 + \alpha_n\boldsymbol{x}^\intercal\boldsymbol{x} - \boldsymbol{x}^\intercal\boldsymbol{x} \\
&\quad + 2\left(\frac{1-n\alpha_n}{2}\boldsymbol{S}^\intercal\boldsymbol{g} + \alpha_n\boldsymbol{s}\right)^\intercal \boldsymbol{x} + \alpha_n\boldsymbol{s}^\intercal\boldsymbol{s} \\
&\quad + (1-n\alpha_n)\boldsymbol{g}^\intercal\boldsymbol{S}^\intercal\boldsymbol{s} - \sigma^2 + \gamma_n.
\end{aligned}$$

Applying Lemma 11 with $\beta = (\alpha_n - 1)$ and $\boldsymbol{b} = \frac{1-n\alpha_n}{2}\boldsymbol{S}^\intercal\boldsymbol{g} + \alpha_n\boldsymbol{s}$ achieves its value

$$\boldsymbol{d}^\intercal\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{d} + 2\boldsymbol{d}^\intercal\boldsymbol{v}$$

where

$$\boldsymbol{d} = \left(\frac{\alpha_n}{2}\boldsymbol{I} + \frac{1-n\alpha_n}{2}\boldsymbol{Z}^\intercal\boldsymbol{S}^\intercal\right)\boldsymbol{g} + \alpha_n\boldsymbol{Z}^\intercal\boldsymbol{s}$$

and the maximin strategy is

$$\begin{aligned}
\boldsymbol{p}^* &= \boldsymbol{v} + \boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{d} \\
&= \boldsymbol{v} + \frac{\alpha_n}{2}\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{g} + \frac{1-n\alpha_n}{2}\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{Z}^\intercal\boldsymbol{S}^\intercal\boldsymbol{g} + \alpha_n\boldsymbol{S}\boldsymbol{s} \\
&= \boldsymbol{v} + \frac{1}{2}\left(1-(n-1)\alpha_n\right)\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{g} + \alpha_n\boldsymbol{S}\boldsymbol{s} \\
&= \mathcal{S}\left(\left(1-(n-1)\alpha_n\right)\boldsymbol{c} + \alpha_n(n-1)\frac{\boldsymbol{s}}{n-1}\right)
\end{aligned}$$

Recall that the player plays $\boldsymbol{a}^* = \boldsymbol{Z}\boldsymbol{p}^*$, and so

$$\boldsymbol{a}^* = \left(1-(n-1)\alpha_n\right)\boldsymbol{c} + \alpha_n(n-1)\frac{\boldsymbol{s}}{n-1} = \boldsymbol{c} + \alpha_n\sum_{t=1}^{n-1}\left(\boldsymbol{x}_t - \boldsymbol{c}\right).$$

Next, we need to show that the value-to-go remains in the correct form. Repeatedly using the identity $\boldsymbol{Z}\boldsymbol{S} = \boldsymbol{I}$, we have

$$\begin{aligned}
\boldsymbol{d}^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \boldsymbol{d} + 2\boldsymbol{d}^\intercal \boldsymbol{v} &= \boldsymbol{g}^\intercal \left(\frac{\alpha_n}{2}\boldsymbol{I} + \frac{1-n\alpha_n}{2}\boldsymbol{Z}^\intercal\boldsymbol{S}^\intercal\right)^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal \left(\frac{\alpha_n}{2}\boldsymbol{I} + \frac{1-n\alpha_n}{2}\boldsymbol{Z}^\intercal\boldsymbol{S}^\intercal\right)\boldsymbol{g} \\
&\quad + 2\alpha_n \boldsymbol{g}^\intercal \left(\frac{\alpha_n}{2}\boldsymbol{I} + \frac{1-n\alpha_n}{2}\boldsymbol{Z}^\intercal\boldsymbol{S}^\intercal\right)^\intercal \boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{Z}^\intercal\boldsymbol{s} \\
&\quad + \alpha_n^2 \boldsymbol{s}^\intercal \boldsymbol{Z}\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{Z}^\intercal\boldsymbol{s} + 2\boldsymbol{d}^\intercal\boldsymbol{v} \\
&= \boldsymbol{g}^\intercal \left(\frac{\alpha_n}{2}\boldsymbol{I} + \frac{1-n\alpha_n}{2}\boldsymbol{Z}^\intercal\boldsymbol{S}^\intercal\right)^\intercal \left(\frac{\alpha_n}{2}\boldsymbol{S}\boldsymbol{S}^\intercal + \frac{1-n\alpha_n}{2}\boldsymbol{S}\boldsymbol{S}^\intercal\right)\boldsymbol{g} \\
&\quad + \alpha_n\left(\alpha_n + (1-n\alpha_n)\right)\boldsymbol{g}^\intercal\boldsymbol{S}\boldsymbol{s} + \alpha_n^2\boldsymbol{s}^\intercal\boldsymbol{s} + \alpha_n\boldsymbol{v}^\intercal\boldsymbol{g} \\
&= \left(\frac{\alpha_n}{2} + \frac{1-n\alpha_n}{2}\right)^2 \boldsymbol{g}^\intercal\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{g} \\
&\quad + \alpha_n\left(\alpha_n + (1-n\alpha_n)\right)\boldsymbol{g}^\intercal\boldsymbol{S}\boldsymbol{s} + \alpha_n^2\boldsymbol{s}^\intercal\boldsymbol{s} + \alpha_n\boldsymbol{v}^\intercal\boldsymbol{g}
\end{aligned}$$

and hence the value is

$$\begin{aligned}
V(\boldsymbol{s}, \sigma^2, n-1) &= (\alpha_n^2 + \alpha_n)\boldsymbol{s}^\intercal\boldsymbol{s} + \left((1-n\alpha_n) + \alpha_n\left(\alpha_n + (1-n\alpha_n)\right)\right)\boldsymbol{g}^\intercal\boldsymbol{S}^\intercal\boldsymbol{s} \\
&\quad + \left(\frac{\alpha_n}{2} + \frac{1-n\alpha_n}{2}\right)^2 \boldsymbol{g}^\intercal\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{g} + \alpha_n\boldsymbol{v}^\intercal\boldsymbol{g} - \sigma^2 + \gamma_n \\
&= \alpha_{n-1}\boldsymbol{s}^\intercal\boldsymbol{s} + \left(1-(n-1)\alpha_{n-1}\right)\boldsymbol{g}^\intercal\boldsymbol{S}^\intercal\boldsymbol{s} \\
&\quad + \frac{\left(1-(n-1)\alpha_n\right)^2}{4}\boldsymbol{g}^\intercal\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{g} + \alpha_n\boldsymbol{v}^\intercal\boldsymbol{g} - \sigma^2 + \gamma_n \\
&= \alpha_{n-1}\boldsymbol{s}^\intercal\boldsymbol{s} + \left(1-(n-1)\alpha_{n-1}\right)\boldsymbol{c}^\intercal\boldsymbol{s} \\
&\quad + \left(1-(n-1)\alpha_n\right)^2\boldsymbol{c}^\intercal\boldsymbol{c} + \alpha_n\boldsymbol{v}^\intercal\boldsymbol{g} - \sigma^2 + \gamma_n
\end{aligned}$$

under the update

$$\begin{aligned}
\alpha_{n-1} &= \alpha_n + \alpha_n^2 \\
\gamma_{n-1} &= \gamma_n + \left(1-(n-1)\alpha_n\right)^2\boldsymbol{c}^\intercal\boldsymbol{c} + \alpha_n\boldsymbol{v}^\intercal\boldsymbol{g}.
\end{aligned}$$

Finally, we need to verify that the strategies above respect the constraints of the game, i.e. that $\boldsymbol{a}^* \in \triangle_{\boldsymbol{Z}}$ and $\boldsymbol{p}^* \in \triangle_{d+1}$. It suffices to verify the latter. Otherwise, the above calculations do not correspond to the minimax strategy. We need to show for all $\boldsymbol{s} \geq 0$ with sum $n-1$ that $\boldsymbol{p}^* \geq 0$, i.e.

$$\boldsymbol{v} + \boldsymbol{S}\left(\left(1-(n-1)\alpha_n\right)\boldsymbol{c} + \alpha_n\boldsymbol{s}\right) = \triangledown\left(\left(1-(n-1)\alpha_n\right)\boldsymbol{c} + \alpha_n\boldsymbol{s}\right) \in \triangle_k.$$

By Proposition 7, $\triangledown$ is one-to-one from $\triangle_k$ to $\triangle_{\boldsymbol{Z}}$, and thus the above statement is equivalent to $\left(1-(n-1)\alpha_n\right)\boldsymbol{c} + \alpha_n\boldsymbol{s} \in \triangle_{\boldsymbol{Z}}$, which we can easily verify by noting that the expression is exactly a convex combination of $\boldsymbol{c} \in \triangle_{\boldsymbol{Z}}$ and $\frac{\boldsymbol{s}}{n-1} \in \triangle_{\boldsymbol{Z}}$. □

This full characterization of the game allows us to derive the following minimax regret bound.

Theorem 13. *The minimax regret on the T-round ball-like simplex $\triangle_{\boldsymbol{Z}}$ satisfies*

$$V(\triangle_{\boldsymbol{Z}}) = \rho^2 \sum_{n=1}^{T} \alpha_n \leq \rho^2 \left(1 + \ln(T)\right), \tag{2.13}$$

where ρ^2 is the squared radius of the minimum enclosing ball, $\boldsymbol{c}^\intercal \boldsymbol{c} + \boldsymbol{v}^\intercal \boldsymbol{g}$.

Proof. The regret is equal to the value of the game, $V = V(\boldsymbol{0}, 0, 0) = \gamma_0$; thus, we need to calculate $\sum_{n=1}^{T}(1 - n\alpha_{n+1})^2$. Observe that

$$\begin{aligned}(1 - n\alpha_{n+1})^2 &= 1 - 2n\alpha_{n+1} + n^2\alpha_{n+1}^2 \\ &= 1 - 2n\alpha_{n+1} + n^2(\alpha_n - \alpha_{n+1}) \\ &= \alpha_{n+1} + 1 - (n+1)^2\alpha_{n+1} + n^2\alpha_n.\end{aligned}$$

After summing over n, the last two terms telescope, and we find

$$\sum_{n=0}^{T-1}(1 - n\alpha_{n+1})^2 \;=\; -T^2\alpha_T + \sum_{n=0}^{T-1}(1 + \alpha_{n+1}) \;=\; \sum_{n=1}^{T}\alpha_n,$$

and hence $V = (\boldsymbol{c}^\intercal \boldsymbol{c} + \boldsymbol{g}^\intercal \boldsymbol{v}) \sum_{n=1}^{T} \alpha_n = \rho^2 \sum_{n=1}^{T} \alpha_n$. verifying the first equality.

Each α_n can be bounded by $1/n$, as observed in [50, proof of Lemma 2]. In the base case $n = T$ this holds with equality, and for $n < T$ we have

$$\alpha_n \;=\; \alpha_{n+1}^2 + \alpha_{n+1} \;\leq\; \frac{1}{(n+1)^2} + \frac{1}{n+1} = \frac{1}{n}\frac{n(n+2)}{(n+1)^2} \leq \frac{1}{n}.$$

It follows that $\gamma_0 \propto \sum_{n=1}^{T} \alpha_n \leq \sum_{n=1}^{T} \frac{1}{n} \leq 1 + \ln(T)$ as desired. □

Remark 14. *In fact, [50, Lemma 3] actually proves the slightly more refined bound*

$$\sum_{n=1}^{T} \alpha_n = \ln(T) - \ln(\ln(T)) + O\left(\frac{\ln(\ln(T))}{\ln(T)}\right).$$

2.4 The Ellipsoid Game

This section parallels the previous: we first introduce the Euclidean loss game on an ellipsoid and relate it to the Mahalanobis loss game on a unit sphere. Section 2.4 then proves some useful lemmas for optimizing square functions and saddle points on ellipsoids and uses these tools to compute the value-to-go and optimal strategies.

Here, we consider the game with Euclidean loss and ellipsoid $\mathcal{X} = \bigcirc_{\boldsymbol{W}} := \{\boldsymbol{x} \in \mathbb{R}^K \mid \|\boldsymbol{x}\|_{\boldsymbol{W}} \leq 1\}$. We will first solve the centered case, as the general case easily follows. The centered case is without loss of generality: since the loss is the difference between the player's and adversary's plays, translating $\mathcal{X}$ will simply shift both sets of plays and offer exactly the same loss.

Ellipsoid Preliminaries

As with the simplex, we could equivalently think about the Euclidean loss on a warped space (e.g. an ellipsoid) or the Mahalanobis loss on a Euclidean space.

Lemma 15. *Playing $\boldsymbol{x}$ and $\boldsymbol{a}$ in the $\bigcirc_{\boldsymbol{W}}$ game with Euclidean loss $\|\cdot\|^2$ is equivalent to playing $\boldsymbol{y} = \left(\boldsymbol{W}^\dagger\right)^{\frac{1}{2}} \boldsymbol{x}$ and $\boldsymbol{b} = \left(\boldsymbol{W}^\dagger\right)^{\frac{1}{2}} \boldsymbol{a}$, respectively, in the Euclidean game with Mahalanobis loss $\|\cdot\|^2_{\boldsymbol{W}^\dagger}$.*

This lemma is a simple consequence of noticing that $\boldsymbol{x} \in \bigcirc_{\boldsymbol{W}} \Leftrightarrow \|\left(\boldsymbol{W}^\dagger\right)^{\frac{1}{2}} \boldsymbol{x}\| \leq 1 \Leftrightarrow \boldsymbol{y} \in \bigcirc$, and $\|\boldsymbol{y} - \boldsymbol{b}\|_{\boldsymbol{W}^\dagger} = \|\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{y} - \boldsymbol{b})\| = \|\boldsymbol{x} - \boldsymbol{a}\|$.

For the rest of the section, we will fix $\boldsymbol{W} \succ 0$ and consider the game with loss $\|\cdot\|^2$ and action set $\mathcal{X} = \bigcirc_{\boldsymbol{W}}$.

Optimization on an Ellipsoid

For each step of the backwards induction, we will need to solve a quadratic saddle point problem on the ellipse, which will be provided by the following lemma.

Lemma 16. *Fix a symmetric matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ and assume $\boldsymbol{A} + \boldsymbol{W}^{-1} \succ \mathbf{0}$. Let $\lambda_{\max}$ be the largest eigenvalue of $\boldsymbol{W} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{W}^{\frac{1}{2}}$ and $\boldsymbol{v}_{\max}$ the corresponding eigenvector. If*

$$\boldsymbol{b}^\intercal \boldsymbol{W}^{-\frac{1}{2}}(\lambda \boldsymbol{W}^{-1} - \boldsymbol{A})^{-2}\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{b} \leq 1,$$

then the optimization problem

$$\inf_{\boldsymbol{a} \in \mathbb{R}^d} \sup_{\boldsymbol{x} \in \bigcirc_{\boldsymbol{W}}} \|\boldsymbol{a} - \boldsymbol{x}\|^2 + \boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^\intercal \boldsymbol{x}$$

has value

$$\boldsymbol{b}^\intercal \left(\lambda_{\max}\boldsymbol{W}^{-1} - \boldsymbol{A}\right)^{-1} \boldsymbol{b} + \lambda_{\max}, \tag{2.14}$$

minimax strategy $\boldsymbol{a}^* = (\lambda_{\max}\boldsymbol{W}^{-1} - \boldsymbol{A})^{-1}\boldsymbol{b}$, and a randomized maximin strategy that plays two unit length vectors, with

$$\Pr\left(\boldsymbol{x} = \boldsymbol{a}_\perp \pm \sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{W}^{-1}\boldsymbol{a}_\perp}\boldsymbol{v}_{\max}\right) = \frac{1}{2} \pm \frac{\boldsymbol{a}_\parallel^\intercal \boldsymbol{v}_{\max}}{2\sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{W}^{-1}\boldsymbol{a}_\perp}},$$

where $\boldsymbol{a}_\perp$ and $\boldsymbol{a}_\parallel$ are the components of $\boldsymbol{a}^*$ perpendicular and parallel to $\boldsymbol{v}_{\max}$.

Proof. As the objective is convex, the inner optimum must be on the boundary and hence $\|\boldsymbol{x}\|_W = 1$. Introduce a Lagrange multiplier λ for $\boldsymbol{x}^\intercal \boldsymbol{W}^{-1}\boldsymbol{x} \leq 1$ to get the Lagrangian

$$\inf_{\boldsymbol{a} \in \mathbb{R}^d} \inf_{\lambda \geq 0} \sup_{\boldsymbol{x}} \|\boldsymbol{a} - \boldsymbol{x}\|^2 + \boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{x}^\intercal \boldsymbol{b} + \lambda(1 - \boldsymbol{x}^\intercal \boldsymbol{W}^{-1}\boldsymbol{x}).$$

This is concave in $\boldsymbol{x}$ if $\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1} \preceq \boldsymbol{0}$, that is, $\lambda_{\max} \leq \lambda$. Differentiating yields the optimizer $\boldsymbol{x}^* = \left(\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1}\right)^{-1}(\boldsymbol{a} - \boldsymbol{b})$, which leaves us with an optimization in only $\boldsymbol{a}$ and λ:

$$\inf_{\boldsymbol{a} \in \mathbb{R}^d} \inf_{\lambda \geq \lambda_{\max}} \boldsymbol{a}^\intercal \boldsymbol{a} - (\boldsymbol{a} - \boldsymbol{b})^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} (\boldsymbol{a} - \boldsymbol{b}) + \lambda.$$

Since the infimums are over closed sets, we can exchange their order. We then exploit the fact that, for all $\lambda \geq \lambda_{\max}$, the objective is convex in a. We have

$$\begin{aligned}
&(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1})\boldsymbol{a}^* = -(\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1}\boldsymbol{b} \\
&\Leftrightarrow -(\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1})\boldsymbol{a}^* = \boldsymbol{b} \\
&\Leftrightarrow \boldsymbol{a}^* = (\lambda \boldsymbol{W}^{-1} - \boldsymbol{A})^{-1}\boldsymbol{b}.
\end{aligned}$$

We now rewrite the optimization problem as

$$\inf_{\lambda \geq \lambda_{\max}} \inf_{a \in \mathbb{R}^d} \boldsymbol{a}^\intercal \left(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1}\right) \boldsymbol{a} + 2\boldsymbol{b}^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{a} - \boldsymbol{b}^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b} + \lambda$$

and plug in $\boldsymbol{a} = \boldsymbol{a}^*$. The the quadratic term can be simplified as

$$\begin{aligned}
&\boldsymbol{a}^{*\intercal} \left(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1}\right) \boldsymbol{a}^* \\
&\quad = \boldsymbol{b}^\intercal (\boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \left(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1}\right) (\boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b} \\
&\quad = \boldsymbol{b}^\intercal (\boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1} - \boldsymbol{I})(\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} (\boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b} \\
&\quad = \boldsymbol{b}^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} (\boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b}.
\end{aligned}$$

Plugging this into the objective, we find

$$\begin{aligned}
&\inf_{\lambda \geq \lambda_{\max}} \inf_{a \in \mathbb{R}^d} \boldsymbol{a}^\intercal \left(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1}\right) \boldsymbol{a} + 2\boldsymbol{b}^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{a} - \boldsymbol{b}^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b} + \lambda \\
&\quad = \inf_{\lambda \geq \lambda_{\max}} -\boldsymbol{b}^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} (\boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b} - \boldsymbol{b}^\intercal (\boldsymbol{I} + \boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b} + \lambda \\
&\quad = \inf_{\lambda \geq \lambda_{\max}} -\boldsymbol{b}^\intercal (\boldsymbol{A} - \lambda \boldsymbol{W}^{-1})^{-1} \boldsymbol{b} + \lambda.
\end{aligned}$$

We can use the spectral decomposition $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{W}^{\frac{1}{2}} = \sum_i \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal$ to rewrite this optimization problem as

$$\begin{aligned}
\inf_{\lambda \geq \lambda_{\max}} \boldsymbol{b}^\intercal (\lambda \boldsymbol{W}^{-1} - \boldsymbol{A})^{-1} \boldsymbol{b} + \lambda &= \inf_{\lambda \geq \lambda_{\max}} \boldsymbol{b}^\intercal \boldsymbol{W}^{\frac{1}{2}} (\lambda \boldsymbol{I} - \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{W}^{\frac{1}{2}})^{-1} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b} + \lambda \\
&= \inf_{\lambda \geq \lambda_{\max}} \left(\sum_i \boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b} \boldsymbol{v}_i^\intercal\right) (\lambda \boldsymbol{I} - \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{W}^{\frac{1}{2}})^{-1} \left(\sum_i \boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b} \boldsymbol{v}_i\right) + \lambda \\
&= \inf_{\lambda \geq \lambda_{\max}} \sum_{i,j} \left(\boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b}\right) \boldsymbol{v}_i^\intercal (\lambda \boldsymbol{I} - \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{W}^{\frac{1}{2}})^{-1} \boldsymbol{v}_j \left(\boldsymbol{v}_j^\intercal \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b}\right) + \lambda \\
&= \inf_{\lambda \geq \lambda_{\max}} \sum_{i,j} \left(\boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b}\right) \frac{\boldsymbol{v}_i^\intercal \boldsymbol{v}_j}{\lambda - \lambda_j} \left(\boldsymbol{v}_j^\intercal \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b}\right) + \lambda \\
&= \inf_{\lambda \geq \lambda_{\max}} \sum_i \frac{\left(\boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b}\right)^2}{\lambda - \lambda_i} + \lambda.
\end{aligned}$$

The derivative of the objective with respect to λ is

$$\begin{aligned}
-\sum_i \frac{\left(\boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{b}\right)^2}{(\lambda-\lambda_i)^2}+1 &= -\sum_{i,j}\left(\boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{b}\right)\frac{\boldsymbol{v}_i^\intercal \boldsymbol{v}_j}{(\lambda-\lambda_j)^2}\left(\boldsymbol{v}_j^\intercal \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{b}\right)+1\\
&= -\left(\sum_i \boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{b}\boldsymbol{v}_i^\intercal\right)(\lambda\boldsymbol{I}-\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{W}^{\frac{1}{2}})^{-2}\left(\sum_i \boldsymbol{v}_i^\intercal \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{b}\boldsymbol{v}_i\right)+1\\
&= -\boldsymbol{b}^\intercal\boldsymbol{W}^{\frac{1}{2}}(\lambda\boldsymbol{I}-\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{W}^{\frac{1}{2}})^{-2}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{b}+1\\
&= -\boldsymbol{b}^\intercal\boldsymbol{W}^{-\frac{1}{2}}(\lambda\boldsymbol{W}^{-1}-\boldsymbol{A})^{-2}\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{b}+1,
\end{aligned}$$

and so the objective is an increasing function in $\lambda \geq \lambda_{\max}$ as long as $\boldsymbol{b}^\intercal\boldsymbol{W}^{-\frac{1}{2}}(\lambda\boldsymbol{W}^{-1}-\boldsymbol{A})^{-2}\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{b} \leq 1$. The infimum is attained at $\lambda_{\max}$ and the $\boldsymbol{a}^*$ is minimax for the given $\boldsymbol{x}^*$ when the assumed inequality is satisfied.

To obtain the maximin strategy, we can take the usual convexification where the Adversary plays distributions P over $\bigcirc_{\boldsymbol{W}}$. This allows us to swap the infimum and supremum (see e.g. Sion's minimax theorem [46]) and obtain the equivalent optimization problem

$$V = \sup_P \inf_{\boldsymbol{a}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim P}\left[\boldsymbol{a}^\intercal\boldsymbol{a}-2\boldsymbol{a}^\intercal\boldsymbol{x}+\boldsymbol{x}^\intercal\boldsymbol{x}+\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x}+2\boldsymbol{b}^\intercal\boldsymbol{x}\right].$$

We notice that the objective only depends on the mean $\boldsymbol{\mu} = \mathbb{E}\,\boldsymbol{x}$ and second moment $\boldsymbol{D} = \mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^\intercal$ of P. Since $\boldsymbol{x}$ is supported on the boundary of $\bigcirc_{\boldsymbol{W}}$, we must have $\operatorname{tr}(\boldsymbol{W}^{-1}\boldsymbol{D}) = 1$ and $\boldsymbol{D} \succeq \mu\mu^\intercal$. It is clear that these two properties are necessary for any distribution on the boundary of $\bigcirc_{\boldsymbol{W}}$; the fact that this criteria is also sufficient is proven in [32, Theorem 2.1]. Therefore, the above optimization is equivalent to

$$\begin{aligned}
V &= \sup_{\boldsymbol{\mu},\boldsymbol{D}}\inf_{\boldsymbol{a}} \boldsymbol{a}^\intercal\boldsymbol{a}-2\boldsymbol{a}^\intercal\boldsymbol{\mu}+\operatorname{tr}((\boldsymbol{I}+\boldsymbol{A})\boldsymbol{D})+2\boldsymbol{b}^\intercal\boldsymbol{\mu}\\
&= \sup_{\boldsymbol{\mu},\boldsymbol{D}}\inf_{\boldsymbol{a}} -\boldsymbol{\mu}^\intercal\boldsymbol{\mu}+2\boldsymbol{b}^\intercal\boldsymbol{\mu}+\operatorname{tr}((\boldsymbol{I}+\boldsymbol{A})\boldsymbol{I})\\
&= -\boldsymbol{a}^{*\intercal}\boldsymbol{a}^*+2\boldsymbol{b}^\intercal\boldsymbol{a}^*+\sup_{\substack{\boldsymbol{D}\succeq\boldsymbol{a}^*\boldsymbol{a}^{*\intercal}\\ \operatorname{tr}(\boldsymbol{W}^{-1}\boldsymbol{D})=1}}\operatorname{tr}\left((\boldsymbol{I}+\boldsymbol{A})\boldsymbol{D}\right)
\end{aligned}$$

where the second equality uses $\boldsymbol{a} = \boldsymbol{\mu}$ and the third used the saddle point condition $\boldsymbol{\mu}^* = \boldsymbol{a}^*$.

We can rewrite the matrix trace constraint as $\operatorname{tr}(\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{D}\boldsymbol{W}^{-\frac{1}{2}}) = 1$ and the objective in the supremum as $\operatorname{tr}\left((\boldsymbol{W}+\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{W}^{\frac{1}{2}})\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{D}\boldsymbol{W}^{-\frac{1}{2}}\right)$. We then see that the matrix $\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{D}\boldsymbol{W}^{-\frac{1}{2}}$ seeks to align with the largest eigenvector of $(\boldsymbol{W}+\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{W}^{\frac{1}{2}})$ while respecting the constraint $\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{D}\boldsymbol{W}^{-\frac{1}{2}} \succeq \left(\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{a}^*\right)\left(\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{a}^*\right)^\intercal$. If we reparameterize by $\boldsymbol{C} = \boldsymbol{W}^{-\frac{1}{2}}\left(\boldsymbol{D}-\boldsymbol{a}^*\boldsymbol{a}^{*\intercal}\right)\boldsymbol{W}^{-\frac{1}{2}}$, it becomes clear that we need to find

$$\sup_{\substack{\boldsymbol{C}\succeq\boldsymbol{0}\\ \operatorname{tr}(\boldsymbol{C})=1-\boldsymbol{a}^{*\intercal}\boldsymbol{W}^{-1}\boldsymbol{a}^{*\intercal}}}\operatorname{tr}\left((\boldsymbol{W}+\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{W}^{\frac{1}{2}})\boldsymbol{C}\right).$$

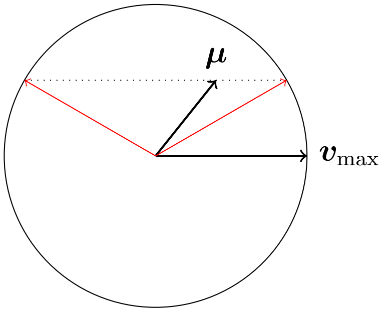

Figure 2.3: Illustration of the maximin distribution from Lemma 16. The mixture of red unit vectors with mean $\boldsymbol{\mu}$ has second moment $\boldsymbol{D} = \boldsymbol{\mu}\boldsymbol{\mu}^\intercal + (1 - \boldsymbol{\mu}^\intercal\boldsymbol{\mu})\boldsymbol{v}_{\max}\boldsymbol{v}_{\max}^\intercal$.

By linearity of the objective, the maximizer can be of rank 1. Hence, this is a (scaled) maximum eigenvalue problem, with solution given by $\boldsymbol{C}^* = (1 - \boldsymbol{a}^{*\intercal}\boldsymbol{W}^{-1}\boldsymbol{a}^*)\boldsymbol{v}_{\max}\boldsymbol{v}_{\max}^\intercal$, so that $\boldsymbol{D}^* = \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{a}^*\boldsymbol{a}^{*\intercal}\boldsymbol{W}^{-\frac{1}{2}} + (1 - \boldsymbol{a}^{*\intercal}\boldsymbol{W}^{-1}\boldsymbol{a}^*)\boldsymbol{v}_{\max}\boldsymbol{v}_{\max}^\intercal$.

It is easy to verify that the mixture in the theorem has the desired mean $\boldsymbol{a}^*$ and second moment $\boldsymbol{D}^*$. See Figure 2.3 for the geometrical intuition.

□

Notice that both the minimax and maximin strategies only depend on $\boldsymbol{W}$ through $\lambda_{\max}$ and $\boldsymbol{v}_{\max}$.

Minimax Analysis of the Ellipsoid Game

With the above lemmas, we can compute the value and strategies for the ellipsoid game in an analogous way to Theorem 12. Again, we find that the value function at the end of the game is quadratic in the state, and, surprisingly, remains quadratic under the backwards induction.

```
**Input:** Game length $T$, $\mathcal{X} = B(\boldsymbol{c}, \boldsymbol{W})$.
Calculate
    $$\boldsymbol{A}_T = \frac{1}{T}\boldsymbol{I}, \qquad \boldsymbol{A}_n = \boldsymbol{A}_{n+1}\left(\lambda_{n+1}^{\max}\boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1}\right)^{-1}\boldsymbol{A}_{n+1} + \boldsymbol{A}_{n+1},$$
where $\lambda_n^{\max} = \lambda_{\max}(\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}_n\boldsymbol{W}^{\frac{1}{2}})$
**for all** $n = 1, \ldots, T$ **do**
    Play $a^* = \boldsymbol{c} + (\lambda_{n+1}^{\max}\boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1})^{-1}\boldsymbol{A}_{n+1}\sum_{t=1}^{n-1}(\boldsymbol{x}_t - \boldsymbol{c})$.
**end for**
```

Figure 2.4: Minimax $\bigcirc_{\boldsymbol{W}}$ strategy

Theorem 17. *Consider the T-round ball game with loss $\|\boldsymbol{a}-\boldsymbol{x}\|^2$ and action space $\bigcirc_{\boldsymbol{W}}$, the matrices defined in Figure 2.4, and the statistics*

$$\boldsymbol{s} = \sum_{t=1}^{n} \boldsymbol{x}_t \quad \textit{and} \quad \sigma^2 = \sum_{t=1}^{n} \boldsymbol{x}_t^\intercal \boldsymbol{x}_t. \tag{2.15}$$

After n rounds, the value-to-go is

$$V(\boldsymbol{s}, \sigma^2, n) \;=\; \frac{1}{2}\boldsymbol{s}^\intercal \boldsymbol{A}_n \boldsymbol{s} - \frac{1}{2}\sigma^2 + \gamma_n,$$

where $\boldsymbol{A}_n$ is as defined in Figure 2.4 and

$$\gamma_T = 0, \qquad\qquad \gamma_n = \gamma_{n+1} + \lambda_{\max}(\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A}_{n+1} \boldsymbol{W}^{\frac{1}{2}}). \tag{2.16}$$

The minimax strategy plays

$$\boldsymbol{a}^*(\boldsymbol{s}, \sigma^2, n) \;=\; \boldsymbol{a}^* = (\lambda_{\max} \boldsymbol{W}^{-1} - \boldsymbol{A}_{n+1} - \boldsymbol{I})^{-1} \boldsymbol{A}_{n+1} \boldsymbol{s}$$

and the maximin strategy plays two unit length vectors with

$$\Pr\left(\boldsymbol{x} = \boldsymbol{a}_\perp \pm \sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{W}^{-1} \boldsymbol{a}_\perp}\, \boldsymbol{v}_{\max}\right) = \frac{1}{2} \pm \frac{\boldsymbol{a}_\parallel^\intercal \boldsymbol{v}_{\max}}{2\sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{W}^{-1} \boldsymbol{a}_\perp}},$$

where $\lambda_{\max}$ and $\boldsymbol{v}_{\max}$ correspond to the largest eigenvalue of $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A}_{n+1} \boldsymbol{W}^{\frac{1}{2}}$ and $\boldsymbol{a}_\perp$ and $\boldsymbol{a}_\parallel$ are the components of $\boldsymbol{a}^*$ perpendicular and parallel to $\boldsymbol{v}_{\max}$.

Proof. Recall that we needed to calculate

$$V(\boldsymbol{s}, \sigma^2, n) \;=\; \inf_{\boldsymbol{a}} \sup_{\boldsymbol{x}} \|\boldsymbol{a} - \boldsymbol{x}\|^2 + (\boldsymbol{s} + \boldsymbol{x})^\intercal \boldsymbol{A}_{n+1} (\boldsymbol{s} + \boldsymbol{x}) - (\sigma^2 + \boldsymbol{x}^\intercal \boldsymbol{x}) + \gamma_{n+1},$$

which may be reorganized to

$$\boldsymbol{s}^\intercal \boldsymbol{A}_{n+1} \boldsymbol{s} - \sigma^2 + \gamma_{n+1} + \inf_{\boldsymbol{a}} \sup_{\boldsymbol{x}} \|\boldsymbol{a} - \boldsymbol{x}\|^2 + \boldsymbol{x}^\intercal \left(\boldsymbol{A}_{n+1} - \boldsymbol{I}\right) \boldsymbol{x} + 2\boldsymbol{s}^\intercal \boldsymbol{A}_{n+1} \boldsymbol{x}.$$

We now apply Lemma 16 with $\boldsymbol{A} = \boldsymbol{A}_{n+1} - \boldsymbol{I}$ and $\boldsymbol{b} = \boldsymbol{A}_{n+1}\boldsymbol{s}$. Let $\lambda_{n+1}^{\max}$ be the largest eigenvalue of $\boldsymbol{W} + \boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{A}_{n+1} - \boldsymbol{I})\boldsymbol{W}^{\frac{1}{2}} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A}_{n+1} \boldsymbol{W}^{\frac{1}{2}}$. If we have

$$1 \;\geq\; \boldsymbol{s}^\intercal \boldsymbol{A}_{n+1} \left(\lambda_{n+1}^{\max} \boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1}\right)^{-2} \boldsymbol{A}_{n+1} \boldsymbol{s}, \tag{2.17}$$

then the value equals

$$\boldsymbol{s}^\intercal \boldsymbol{A}_{n+1} \boldsymbol{s} - \sigma^2 + \gamma_{n+1} + \lambda_{n+1}^{\max} + \boldsymbol{s}^\intercal \boldsymbol{A}_{n+1} \left(\lambda_{n+1}^{\max} \boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1}\right)^{-1} \boldsymbol{A}_{n+1} \boldsymbol{s}$$

with minimax strategy

$$\boldsymbol{a}^* = (\lambda_{n+1}^{\max} \boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1})^{-1} \boldsymbol{A}_{n+1} \boldsymbol{s}.$$

If we assert that $V(\boldsymbol{s}, \sigma^2, n) = \boldsymbol{s}^\intercal \boldsymbol{A}_n \boldsymbol{s} - \sigma^2 + \gamma_n$, we find that we must have the correspondence:

$$\begin{aligned} \boldsymbol{A}_n &= \boldsymbol{A}_{n+1} \left(\lambda_{n+1}^{\max} \boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1}\right)^{-1} \boldsymbol{A}_{n+1} + \boldsymbol{A}_{n+1}, \\ \gamma_n &= \lambda_{n+1}^{\max} + \gamma_{n+1}. \end{aligned}$$

The final piece is to check Equation (2.17). We use the observation that $\boldsymbol{A}_n$ has the same eigenspace as $\boldsymbol{W}^{-1}$ (see the comment after the proof for more discussion). Then, let $\boldsymbol{\nu}$ denote the eigenvalues of $\boldsymbol{W}^{-1}$. We have that the largest eigenvalue of $\boldsymbol{A}_{n+1} \left(\lambda_{n+1}^{\max} \boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1}\right)^{-2} \boldsymbol{A}_{n+1}$ is $\left(\lambda_{n+1}^{\max}\right)^2$ because the mapping

$$\lambda \mapsto \frac{\lambda^2}{\left(\lambda_{n+1}^{\max} \nu + 1 - \lambda\right)^2}$$

is monotonic in $\lambda \leq \lambda^{\max}$.

We can then calculate $\lambda_{n+1}^{\max} = \lambda_{\max}(\boldsymbol{W}^{\frac{1}{2}} A_{n+1} \boldsymbol{W}^{\frac{1}{2}}) = \lambda_{\max}(A_{n+1})/\nu_{\max}$

The proof now follows by combining the recurrence for the largest eigenvalue from Section 2.4 with the bound on α_n from the proof of Theorem 13:

$$\boldsymbol{s}^\intercal \boldsymbol{A}_{n+1} \left(\lambda_{n+1}^{\max} \boldsymbol{W}^{-1} + \boldsymbol{I} - \boldsymbol{A}_{n+1}\right)^{-2} \boldsymbol{A}_{n+1} \boldsymbol{s} \leq \left(\frac{\lambda_{n+1}^{\max}}{\nu_{\max}}\right)^2 \|\boldsymbol{s}\|^2 \leq \alpha_{n+1}^2 n^2 \leq \frac{n^2}{(n+1)^2} < 1.$$

□

Understanding the Eigenvalues of $\boldsymbol{A}_n$

The eigensystem of $\boldsymbol{A}_n$ is always the same as that of $\boldsymbol{W}$. Assume that $\boldsymbol{W}^{-1}$ has eigenvalue, eigenvector pairs $(\nu_i, \boldsymbol{v}_i)$, in decreasing order, and let λ_{n+1}^i denote the eigenvalue of $\boldsymbol{A}_{n+1}$ associated with $\boldsymbol{v}_i$. Assuming that the order of λ_n^i does not change, we can calculate $\lambda_n^{\max} = \frac{\lambda_n^1}{\nu_1}$, set $\lambda_T^i = \frac{1}{T}$ are update λ_n^i as

$$\lambda_n^i = \frac{(\lambda_{n+1}^i)^2}{\frac{\nu_i}{\nu_1}\lambda_{n+1}^1 + 1 - \lambda_{n+1}^i} + \lambda_{n+1}^i,$$

which leaves the order of the λ_n^i unchanged. The largest eigenvalue λ_n^1 satisfies the recurrence $\lambda_T^1 = 1/T$ and $\lambda_n^1 = \left(\lambda_{n+1}^1\right)^2 + \lambda_{n+1}^1$, which, remarkably, is the same recurrence for the α_n parameter in the Brier game, i.e. $\lambda_n^{\max} = \frac{\alpha_n}{\nu_{\max}}$.

This observation is the key to analyzing the minimax regret.

Theorem 18. *The minimax regret of the T-round ball game satisfies*

$$V = \lambda_{\max}(\boldsymbol{W}^{-1}) \sum_{n=1}^{T} \alpha_n \leq \frac{1 + \ln(T)}{2} \lambda_{\max}(\boldsymbol{W}^{-1}).$$

Proof. Notice that $V = V(\boldsymbol{0}, 0, 0) = \gamma_0 = \sum_{n=1}^T \lambda_n^{\max} = \lambda_{\max}(\boldsymbol{W}^{-1}) \sum_{n=1}^T \alpha_n$ and use the bound on $\sum_{n=1}^T \alpha_n$ from the proof of Theorem 13. □

General Ellipsoid

We now generalize the results of the previous section from $B(0, \boldsymbol{W})$ to $B(\boldsymbol{c}, \boldsymbol{W})$.

Theorem 19. *Given a positive semi-definite matrix $\boldsymbol{W}$ and a vector $\boldsymbol{c} \in \mathbb{R}^d$, the T-round ball game with loss $\|\boldsymbol{a} - \boldsymbol{x}\|^2$ and action space $\mathcal{X} = \bigcirc_{\boldsymbol{W}} + \boldsymbol{c}$ has the same regret as the game with $\mathcal{X} = \bigcirc_{\boldsymbol{W}}$, minimax strategy*

$$\boldsymbol{a}^*(\boldsymbol{s}, \sigma^2, n) \;=\; \boldsymbol{a}^* = \boldsymbol{c} + (\lambda_{\max}\boldsymbol{W}^{-1} - \boldsymbol{A}_{n+1} - \boldsymbol{I})^{-1}\boldsymbol{A}_{n+1}\sum_{t=1}^{n}(\boldsymbol{x}_t - \boldsymbol{c})$$

and the maximin strategy plays two unit length vectors with

$$\Pr\left(\boldsymbol{x} = \boldsymbol{c} + \boldsymbol{a}_\perp \pm \sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{W}^{-1}\boldsymbol{a}_\perp}\boldsymbol{v}_{\max}\right) = \frac{1}{2} \pm \frac{\boldsymbol{a}_\parallel^\intercal \boldsymbol{v}_{\max}}{2\sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{W}^{-1}\boldsymbol{a}_\perp}},$$

where $\lambda_{\max}$ and $\boldsymbol{v}_{\max}$ correspond to the largest eigenvalue of $\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{A}_{n+1}\boldsymbol{W}^{\frac{1}{2}}$ and $\boldsymbol{a}_\perp$ and $\boldsymbol{a}_\parallel$ are the components of $\boldsymbol{a}^*$ perpendicular and parallel to $\boldsymbol{v}_{\max}$. That is, the minimax and maximin strategies are just offset versions of the centered-ellipsoid case.

Proof. Let $\boldsymbol{x}_n'$, $\boldsymbol{s}_n'$, $\sigma^{2\prime}$, and $\boldsymbol{a}_n'$ denote the centered versions of $\boldsymbol{x}_n$, $\boldsymbol{s}_n$, σ^2, and $\boldsymbol{a}_n$, respectively. It is straightforward to notice that the value of offline problem with $\boldsymbol{x}_n$ is the same as with $\boldsymbol{x}_n'$ with solution $\boldsymbol{a} = \boldsymbol{a}^* + \boldsymbol{c}$, where $\boldsymbol{a}^*$ is given by Equation (2.7). Tracing through the calculations of the proof of Theorem 17, we find that the calculations with $\boldsymbol{x}_n'$ are exactly the same as for the centered case if we notice that $\|\boldsymbol{a}_n - \boldsymbol{x}_n\| = \|\boldsymbol{a}_n' - \boldsymbol{x}_n'\|$ with $\boldsymbol{a}' = \boldsymbol{a} - \boldsymbol{c}$. Thus, the optimal response is the $\boldsymbol{a}'$ given by Theorem 17, which equates to the $\boldsymbol{a}$ presented in the theorem statement. □

2.5 The Convex Set Game

Building off the two special cases in the previous sections, we are now ready to derive the minimax strategy and value-to-go for bounded convex sets.

The basic argument is as follows. We first show that any convex, compact set $\mathcal{X}$ has some contained ball-like k-simplex contained with the same minimum enclosing ball. The regret of this ball-like simplex must lower bound the regret on $\mathcal{X}$, since it is a smaller set and the adversary's action is more limited. On the other hand, the regret on the minimum enclosing ball must upper bound the regret on $\mathcal{X}$. Since the regret on these two sets is actually equal, the regret on $\mathcal{X}$ must also agree.

Remark 20. *The bounded assumption is necessary. Otherwise, the adversary could play some fixed point for the first $T - 1$ rounds, then play $\boldsymbol{x}_T$ with arbitrary norm on the last round. Regardless of $\boldsymbol{a}_T$, the adversary could pick $\boldsymbol{x}_T$ with $\|\boldsymbol{x}_T\| \gg \|\boldsymbol{a}_T\|$ such that the loss of the comparator is still much smaller (since it is the average of all the $\boldsymbol{x}_t$).*

Mechanically, the backwards induction is not feasible even from the last round with unbounded $\mathcal{X}$ (as we are maximizing a convex function over an unbounded set).

Finding a Supporting Ball-like Simplex

Let $\mathcal{X}$ have smallest enclosing ball with center $\boldsymbol{c}$ and radius ρ. To lower bound the regret of playing on $\mathcal{X}$, we need to find a subset with an easy to calculate regret. To this end, we define:

Definition 21. *Let $\mathcal{X}$ be a convex, compact set. We say that a simplex $\triangle$ of size $k \leq d+1$ supports $\mathcal{X}$ if $\triangle$ is a ball-like simplex and has the same minimum enclosing ball as $\mathcal{X}$.*

This section shows that we can always find a simplex to support $\mathcal{X}$. We shall see that the ball-like simplex is formed from a subset of the intersection of the minimum ball and $\mathcal{X}$.

Lemma 22. *Let $\mathcal{X}$ be a closed and bounded set, possible in some Hilbert space $\mathcal{H}$. The smallest enclosing ball is unique.*

Proof. First, we argue that the minimum enclosing radius is well defined and attained. Define the set $\mathcal{R} = \{\rho \in \mathbb{R} : \exists \boldsymbol{x} \in \mathcal{X} \text{ s.t. } \mathcal{X} \subseteq B(\boldsymbol{x}, \rho)\}$. This set is bounded above since $\mathcal{X}$ is bounded; otherwise, $\mathcal{X}$ must be a singleton and the result is trivial. Hence, it must have a non-zero infimum, which we denote ρ.

We will check uniqueness of the center by contradiction. Assume that both c_1 and c_2 are centers of smallest enclosing balls, i.e. $\mathcal{X} \subseteq B(\boldsymbol{c}_1, \rho) \cap B(\boldsymbol{c}_2, \rho)$. First, note that $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ must both be in the convex hull of $\mathcal{X}$. Otherwise, the projection theorem for Hilbert space implies that the projection of, say, $\boldsymbol{c}_1$ onto the convex hull must be closer to every point.

Defining $\boldsymbol{c}' = \frac{c_1 + c_2}{2}$, we easily derive

$$\|\boldsymbol{x} - \boldsymbol{c}_i\|^2 = \|\boldsymbol{x} - \boldsymbol{c}' + (\boldsymbol{c}' - \boldsymbol{c}_i)\|^2 = \|\boldsymbol{x} - \boldsymbol{c}'\|^2 + \|\boldsymbol{c}' - \boldsymbol{c}_i\|^2 + 2\langle \boldsymbol{x} - \boldsymbol{c}', \boldsymbol{c}' - \boldsymbol{c}_i\rangle.$$

Now, assume that $\boldsymbol{x}$ is closer to $\boldsymbol{c}_1$ than $\boldsymbol{c}_2$. This implies

$$0 \leq \|\boldsymbol{x} - \boldsymbol{c}_2\|^2 - \|\boldsymbol{x} - \boldsymbol{c}_1\|^2 = \|\boldsymbol{c}' - \boldsymbol{c}_2\|^2 - \|\boldsymbol{c}' - \boldsymbol{c}_1\|^2 + 2\langle \boldsymbol{x} - \boldsymbol{c}', \boldsymbol{c}_1 - \boldsymbol{c}_2\rangle,$$

and since $\|\boldsymbol{c}' - \boldsymbol{c}_2\|^2 = \|\boldsymbol{c}' - \boldsymbol{c}_1\|^2$, we have that $2\langle \boldsymbol{x} - \boldsymbol{c}', \boldsymbol{c}_1 - \boldsymbol{c}_2\rangle \geq 0$.

Finally, we expand $\|\boldsymbol{x} - \boldsymbol{c}_2\|^2$ to

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{c}_2\|^2 =& \|\boldsymbol{x} - \boldsymbol{c}_1\|^2 + \|\boldsymbol{c}_1 - \boldsymbol{c}_2\|^2 + 2\langle \boldsymbol{x} - \boldsymbol{c}_1, \boldsymbol{c}_1 - \boldsymbol{c}_2\rangle \\ \geq& \|\boldsymbol{x} - \boldsymbol{c}_1\|^2 + \|\boldsymbol{c}_1 - \boldsymbol{c}_2\|^2,\end{aligned}$$

and hence $\boldsymbol{B}\left(\boldsymbol{c}', \sqrt{\rho^2 - \|\boldsymbol{c}_1 - \boldsymbol{c}_2\|^2}\right)$ is a smaller enclosing ball, yielding a contradiction. This implies that $\boldsymbol{c}$ is unique. □

Next, we turn towards indentifying a subset of $\mathcal{X}$ that has a well understood regret. Any supporting simplex seems to be a good candidate, but we must first prove their existence.

Lemma 23. *Every convex, compact set $\mathcal{X} \in \mathbb{R}^d$ has a supporting simplex of size $2 \leq k \leq d+1$. That is, there exist points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ in general position that share the same minimum enclosing ball as $\mathcal{X}$.*

Proof. Let $\hat{\mathcal{X}} = \partial\mathcal{X} \cap B(\boldsymbol{c}_{\mathcal{X}}, \rho_{\mathcal{X}})$, where $\partial\mathcal{X}$ is the set of boundary points of $\mathcal{X}$. We claim that $\hat{\mathcal{X}}$ and $\mathcal{X}$ share the same minimum ball. Let us consider $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}-\boldsymbol{c}\|$ as a function of $\boldsymbol{c}$. At $\boldsymbol{c}_{\mathcal{X}}$, we have

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}-\boldsymbol{c}\| = \sup_{\boldsymbol{x}\in\hat{\mathcal{X}}}\|\boldsymbol{x}-\boldsymbol{c}\| = \rho$$

and the minimum enclosing ball of $\hat{\mathcal{X}}$ must be the same as the minimum enclosing ball of $\mathcal{X}$ by uniqueness.

Now let $\boldsymbol{c}$ denote this unique minimum. By Caratheodory's theorem, there exist points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k \in \hat{\mathcal{X}}$ with corresponding matrix $\boldsymbol{Z}$ such that, for some $\boldsymbol{p}^* \in \triangle_k$, we have $\boldsymbol{c} = \boldsymbol{Z}\boldsymbol{p}^*$. Without loss of generality, assume that $\boldsymbol{p}^*$ has all non-zero entries (otherwise, the corresponding $\boldsymbol{z}_i$ could be removed without consequence). Let $\triangle_{\boldsymbol{Z}}$ be the convex hull (we will later show that $\triangle_{\boldsymbol{z}}$ is a simplex). Recall that the minimum enclosing ball is the solution to the optimization

$$\min_{\substack{\boldsymbol{c}\in\mathbb{R}^d,\rho^2 \\ \forall i:\|\boldsymbol{c}-\boldsymbol{z}_i\|^2\le\rho^2}} \rho^2.$$

Using $\boldsymbol{p}$ as the dual variables, the Lagrangian is

$$L(\boldsymbol{c},\rho^2,\boldsymbol{p}) \;=\; \max_{\boldsymbol{p}\ge 0}\min_{\boldsymbol{c}\in\mathbb{R}^d,\rho^2} \rho^2 + \sum_i p_i\left(\|\boldsymbol{c}-\boldsymbol{z}_i\|^2 - \rho^2\right).$$

As any finite convex hull is bounded, Slater's condition holds, so we have strong duality and the KKT conditions are necessary and sufficient for optimality. Hence $\boldsymbol{c}$, ρ^2 and $\boldsymbol{p}$ are optimal iff

$$\begin{aligned}
\|\boldsymbol{c}-\boldsymbol{z}_i\|^2 &\le \rho^2 && \text{(primal feasible)}\\
\boldsymbol{p} &\ge \boldsymbol{0} && \text{(dual feasible)}\\
p_i\left(\|\boldsymbol{c}-\boldsymbol{z}_i\|^2-\rho^2\right) &= 0 && \text{(complementary slackness)}\\
0 &= 2\sum_i p_i(\boldsymbol{c}-\boldsymbol{z}_i) && (\partial_{\boldsymbol{c}}L \text{ vanishes})\\
1 &= \sum_i p_i. && (\partial_{\rho^2}L \text{ vanishes})
\end{aligned}$$

These conditions are easily checked for $\boldsymbol{p}^*$. We constructed $\boldsymbol{p}^* \in \triangle_k$ such that $\|\boldsymbol{Z}\boldsymbol{p}^* - \boldsymbol{z}_i\|^2 = \rho^2$ for all i (which covers all but the forth condition) and $\boldsymbol{Z}\boldsymbol{p}^* = \boldsymbol{c}$ (which covers the forth). Hence, by uniqueness of the the smallest ball, $\triangle_{\boldsymbol{Z}'}$ and $\mathcal{X}$ must share the same smallest enclosing ball.

It remains to check that $\triangle_{\boldsymbol{Z}}$ is ball-like. It is sufficient to show that $\boldsymbol{v} + \boldsymbol{S}\boldsymbol{c}_{\boldsymbol{Z}} \in \triangle_k$. This is an easy consequence since $\boldsymbol{c}_{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{p}^*$ and the mapping from $\triangle_{\boldsymbol{Z}}$ to $\triangle_k$ is one-to-one. □

Remark 24. *Playing the minimax strategy for a general closed set requires knowing the center of the minimum enclosing ball. This requirement is very benign under any assumption that the center is at the origin (e.g. symmetry assumptions). However, even in more complicated sets, the problem of finding the smallest enclosing ball is well understood. For example, there are linear time algorithms for finding the smallest ball of a set of points [38] (in our earlier derivations, we formulated this as a quadratic program).*

Main Result

We have finally gathered the tools necessary to prove our main result. Since the adversary's plays are constrained but the player's are not (although the player will naturally play in the same space), games with larger sets will have at least as much regret. Thus, for any $\mathcal{X}$, we can upper bound its regret, $\mathcal{R}_{\mathcal{X}}$, by the regret of the game on the smallest enclosing ball,

$$\mathcal{R}_{\mathcal{X}} \leq \rho_{\mathcal{X}}^2 \sum_{t=1}^{T} \alpha_t.$$

On the other hand, the above section has shown that there is a ball-like simplex that lower bounds the regret on $\mathcal{X}$ by $\rho_{\mathcal{X}} \sum_{t=1}^{T} \alpha_t$. Since these two bounds match, the regret on all sets in the middle must be the same:

Theorem 25. *Let $\mathcal{X}$ be a compact set with center and radius $\mathbf{c}$ and ρ, respectively. The squared loss game has value*

$$V = \rho^2 \sum_{n=1}^{T} \alpha_n, \tag{2.18}$$

which is achieved by the minimax strategy

$$\boldsymbol{a}_n = \boldsymbol{c} + \alpha_n \sum_{t=1}^{n-1} \left(\boldsymbol{x}_t - \boldsymbol{c}\right). \tag{MM}$$

Proof. We will use $V(\mathcal{S})$ to denote the value of the Euclidean game when $\mathcal{X} = \mathcal{S}$. Note that if $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $V(\mathcal{S}_1) \leq V(\mathcal{S}_2)$ since the adversary has strictly more power (he can play from a larger set).

Recall that $\triangle_{\boldsymbol{Z}(\mathcal{X})}$ is the ball-like simplex sharing the minimum enclosing ball with $\mathcal{X}$. Since $\triangle_{\boldsymbol{Z}(\mathcal{X})}$ is ball-like, Theorem 13 implies

$$\rho_{\mathcal{X}}^2 \sum_{n=1}^{T} \alpha_n = V\left(\triangle_{\boldsymbol{Z}(\mathcal{X})}\right) \leq V(\mathcal{X}).$$

On the other hand, the ball $\mathcal{B}(\boldsymbol{c}_{\mathcal{X}}, \rho_{\mathcal{X}})$ is a special case of the ellipsoid with $\boldsymbol{W} = \frac{1}{\rho_{\mathcal{X}}}\boldsymbol{I}$. Since $\mathcal{X} \subseteq B(\boldsymbol{c}_{\mathcal{X}}, \rho_{\mathcal{X}})$, Theorem 18 establishes

$$V(\mathcal{X}) \leq V(\mathcal{B}(\boldsymbol{c}_{\mathcal{X}}, \rho_{\mathcal{X}})) = \lambda_{\max}(\boldsymbol{W}^{-1}) \sum_{n=1}^{T} \alpha_n = \rho_{\mathcal{X}}^2 \sum_{n=1}^{T} \alpha_n.$$

Therefore, we have

$$\sum_{n=1}^{T} \alpha_n \rho_{\mathcal{C}}^2 \geq V(\mathcal{C}) \geq \sum_{n=1}^{T} \alpha_n \rho_{\mathcal{C}}^2,$$

which provides the other side of the inequality. □

Remark 26. *Theorem 25 provides the value and minimax strategy for the convex set game, which is a surprising result. We cannot calculate the value-to-go or the maximin strategy for any of the sets between the supporting ball-like simplex and an enclosing ellipsoid; yet we can calculate the minimax strategy and the value. Recall that the maximin strategy on the ball-like simplex plays (Theorem 12) vertex $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k$ with probability proportional to*

$$\boldsymbol{p}_t = \boldsymbol{v} + \boldsymbol{S}\left((1-(t-1)\alpha_t)\,\boldsymbol{c} + \alpha_t \boldsymbol{S}\boldsymbol{s}_t\right)$$

and the maximin strategy on any ellipsoid $B(\boldsymbol{W}, \boldsymbol{c})$ plays (Theorem)

$$\Pr\left(\boldsymbol{x} = \boldsymbol{c} + \boldsymbol{a}_\perp \pm \sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{a}_\perp} \boldsymbol{v}_{\max}\right) = \frac{1}{2} \pm \frac{\boldsymbol{a}_\parallel^\intercal \boldsymbol{v}_{\max}}{2\sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{a}_\perp}}.$$

where $\lambda_{\max}$ and $\boldsymbol{v}_{\max}$ correspond to the largest eigenvalue of $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A}_{n+1} \boldsymbol{W}^{\frac{1}{2}}$ and $\boldsymbol{a}_\perp$ and $\boldsymbol{a}_\parallel$ are the components of $\boldsymbol{a}^*$ perpendicular and parallel to $\boldsymbol{v}_{\max}$.

2.6 The Hilbert Space Game

In the previous sections, we analyzed the Euclidean game for compact sets and found that the regret only depends on the radius of the smallest containing Euclidean ball. In particular, the regret has no explicit dependence on dimension, which suggests a natural extension to Hilbert space.

Consider the squared loss game where the adversary is constrained to play $\boldsymbol{x}_t \in \mathcal{X}$, where $\mathcal{X} \subset \mathcal{H}$ is some closed and bounded set of a Hilbert space $\mathcal{H}$. In $Reals^d$, the minimax strategy is

$$\boldsymbol{a}_n = \boldsymbol{c} + \alpha_n \sum_{t=1}^{n-1} (\boldsymbol{x}_t - \boldsymbol{c}), \tag{MM}$$

where $\boldsymbol{c}$ is the center of the smallest enclosing ball. By Lemma 22, there is a unique smallest enclosing ball even if $\mathcal{X}$ is in $\mathcal{H}$, and therefore the above strategy is well defined is $\mathcal{H}$. It is easy to compute since it is a linear combination of past data and $\boldsymbol{c}$.

The main result of this section is showing that the finite-dimensional results carry over to Hilbert space: the regret of (MM) is $\rho^2 \sum_{t=1}^T \alpha_t$. We accomplish this by proving a direct upper bound and then constructing a lower bound from a sequence of finite dimensional games (which we analyzed in the previous section).

Lemma 27. *For some fixed point $\boldsymbol{d} \in \mathcal{H}$, the strategy*

$$\boldsymbol{a}_n \;:=\; \boldsymbol{d} + \alpha_n \sum_{t=1}^{n-1} (\boldsymbol{x}_t - \boldsymbol{d}),$$

with the same α_n recursion ($\alpha_T = \frac{1}{T}$ and $\alpha_{n-1} = \alpha_n^2 + \alpha_n$), obtains regret

$$\boldsymbol{R}_T = \sum_{t=1}^T \alpha_t \left\|\boldsymbol{x}_t - \boldsymbol{d}\right\|^2 \le \left(\sup_{\boldsymbol{z}\in\mathcal{X}} \left\|\boldsymbol{z} - \boldsymbol{d}\right\|^2\right) \sum_{t=1}^T \alpha_t.$$

Proof. Recall that the regret is

$$R_T \coloneqq \sum_{t=1}^T \|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2 - \inf_{\boldsymbol{a} \in \mathcal{H}} \sum_{t=1}^T \|\boldsymbol{a} - \boldsymbol{x}_t\|^2,$$

the optimal comparator is $\boldsymbol{a} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t$, and hence

$$\inf_{\boldsymbol{a} \in \mathcal{H}} \sum_{t=1}^T \|\boldsymbol{a} - \boldsymbol{x}_t\|^2 = \sum_{t=1}^T \|\boldsymbol{x}_t\|^2 - \frac{1}{T}\left\|\sum_{t=1}^T \boldsymbol{x}_t\right\|^2.$$

Expanding, we find that

$$\begin{aligned}
R_T &= \sum_{t=1}^T \left(\|\boldsymbol{a}_t\|^2 - 2\langle \boldsymbol{a}_t, \boldsymbol{x}_t\rangle\right) + \frac{1}{T}\left\|\sum_{t=1}^T \boldsymbol{x}_t\right\|^2 \\
&= \sum_{t=1}^T \left(\left\|\boldsymbol{d} + \alpha_t \sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d})\right\|^2 - 2\left\langle \boldsymbol{d} + \alpha_t \sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d}), \boldsymbol{x}_t \right\rangle\right) \\
&\quad + \frac{1}{T}\left\|\sum_{t=1}^T (\boldsymbol{x}_t - \boldsymbol{d})\right\|^2 + 2\left\langle \sum_{t=1}^T (\boldsymbol{x}_t - \boldsymbol{d}), \boldsymbol{d} \right\rangle + T\|\boldsymbol{d}\|^2 \\
&= \sum_{t=1}^T \left(\alpha_t^2 \left\|\sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d})\right\|^2 - 2\alpha_t \left\langle \boldsymbol{x}_t - \boldsymbol{d}, \sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d}) \right\rangle\right) + \frac{1}{T}\left\|\sum_{t=1}^T (\boldsymbol{x}_t - \boldsymbol{d})\right\|^2 \\
&= \sum_{t=1}^T \left((\alpha_{t-1} - \alpha_t) \left\|\sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d})\right\|^2 - 2\alpha_t \left\langle \boldsymbol{x}_t - \boldsymbol{d}, \sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d}) \right\rangle\right) + \alpha_T \left\|\sum_{t=1}^T (\boldsymbol{x}_t - \boldsymbol{d})\right\|^2 \\
&= \sum_{t=1}^T \alpha_t \left(\left\|\sum_{s=1}^{t} (\boldsymbol{x}_s - \boldsymbol{d})\right\|^2 - \left\|\sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d})\right\|^2 - 2\left\langle \boldsymbol{x}_t - \boldsymbol{d}, \sum_{s=1}^{t-1} (\boldsymbol{x}_s - \boldsymbol{d}) \right\rangle\right) \\
&= \sum_{t=1}^T \alpha_t \|\boldsymbol{x}_t - \boldsymbol{d}\|^2
\end{aligned}$$

A straightforward bound on the regret is

$$\sum_{t=1}^T \alpha_t \|\boldsymbol{x}_t - \boldsymbol{d}\|^2 \le \left(\sup_{\boldsymbol{z} \in \mathcal{X}} \|\boldsymbol{z} - \boldsymbol{d}\|^2\right) \sum_{t=1}^T \alpha_t,$$

and optimizing the above guarantee over the choice of $\boldsymbol{d}$ reduces to finding the *squared radius of the smallest ball enclosing* $\mathcal{X}$, i.e.

$$\rho^2 \coloneqq \inf_{\boldsymbol{d} \in \mathcal{H}} \sup_{\boldsymbol{z} \in \mathcal{X}} \|\boldsymbol{z} - \boldsymbol{d}\|^2.$$

It follows that the minimax regret has the bound

$$V(\mathcal{H}) \;\leq\; \rho^2 \sum_{t=1}^{T} \alpha_t. \tag{2.19}$$

Note that this holds even if the $\inf_{\boldsymbol{d}}$ in the definition of ρ^2 is not attained. □

This result matches the upper bound for the finite dimension case. The infimum $\boldsymbol{c}$ is always attained and is unique (Lemma 22), and we have seen that the maximin strategy always plays $\boldsymbol{x}_t$ such that $\|\boldsymbol{x}_t - \boldsymbol{c}\| = \rho$; thus, the upper bound can be met with equality.

Before we prove the matching lower bound, we prove the following proposition, which may be of independent interest.

Proposition 28. *For any closed set $\mathcal{Z} \in \mathcal{H}$, there exists a sequence of finite sets $C_i \subseteq \mathcal{Z}$ such that*

$$\frac{\rho(C_{i+1})}{\rho(\mathcal{Z})} \geq 1 - \sqrt{\frac{2}{i}},$$

where $\rho(C_i)$ is the radius of the smallest enclosing ball of the convex hull of C_i.

Proof. We construct C_i iteratively by adding a point to C_{i-1}. The key argument is showing that we can always pick a point that increases the radius significantly towards ρ. We start with a singleton $C_1 = \{\boldsymbol{z}_1\}$ with $\boldsymbol{z}_1 \in \mathcal{Z}$ arbitrary. This witnesses $\rho(C_1) = 0$ with center $\boldsymbol{c}_1 = \boldsymbol{z}_1$. We construct C_{i+1} from C_i as follows. If $\rho(\operatorname{conv}(C_i)) = \rho(\mathcal{Z})$, we are done. Otherwise, by definition of $\rho(\mathcal{Z})$, there is a point $\boldsymbol{x} \in \mathcal{Z}$ such that $\|\boldsymbol{c} - \boldsymbol{x}\| \geq \rho(\mathcal{Z})$. We then set $C_{i+1} = C_i \cup \{\boldsymbol{x}\}$. The difficulty is in showing that $\rho(\operatorname{conv}(C_i))$ makes significant progress to $\rho(\mathcal{Z})$.

For any finite set C, we can perform a min-max swap,

$$\rho^2(C) \;=\; \min_{\boldsymbol{c}} \max_{\boldsymbol{z} \in \boldsymbol{C}} \|\boldsymbol{z} - \boldsymbol{c}\|^2 \;=\; \max_{q \in \triangle_C} \min_{\boldsymbol{c}} \sum_{\boldsymbol{z} \in C} q(\boldsymbol{z}) \|\boldsymbol{z} - \boldsymbol{c}\|^2,$$

and obtain a distribution q on C_i with mean $\boldsymbol{c}$ such that

$$\rho^2(C_i) \;=\; \sum_{\boldsymbol{z} \in C_i} q(\boldsymbol{z}) \|\boldsymbol{z} - \boldsymbol{c}\|^2.$$

Now, consider the distribution q_λ on C_{i+1} that puts probability $(1-\lambda)q(\boldsymbol{z})$ on $\boldsymbol{z} \in C_i$ and probability λ on $\boldsymbol{x}$. The optimal center w.r.t. this distribution is its mean $\boldsymbol{c}_\lambda \;=\; (1-\lambda)\boldsymbol{c} + \lambda \boldsymbol{x} = \boldsymbol{c} + \lambda(\boldsymbol{x} - \boldsymbol{c})$, and we can lower bound its radius by evaluating the dual problem at q_λ and noting that $\|\boldsymbol{x} - \boldsymbol{c}_\lambda\|^2 =$

$(1-\lambda)^2\|\boldsymbol{x}-\boldsymbol{c}\|^2 \geq (1-\lambda)^2\rho^2(\mathcal{Z})$:

$$\begin{aligned}
\rho^2(C_{i+1}) &\geq (1-\lambda)\sum_{\boldsymbol{z}\in C_i} q(\boldsymbol{z})\|\boldsymbol{z}-\boldsymbol{c}_\lambda\|^2 + \lambda\|\boldsymbol{x}-\boldsymbol{c}_\lambda\|^2 \\
&= (1-\lambda)\sum_{\boldsymbol{z}\in C_i} q(\boldsymbol{z})\left(\|\boldsymbol{z}-\boldsymbol{c}\|^2 - 2\lambda\langle \boldsymbol{z}-\boldsymbol{c}, \boldsymbol{x}-\boldsymbol{c}\rangle + \lambda^2\|\boldsymbol{x}-\boldsymbol{c}\|^2\right) + \lambda\|\boldsymbol{x}-\boldsymbol{c}_\lambda\|^2 \\
&\geq (1-\lambda)\rho^2(C_i) + (1-\lambda)\lambda^2\rho^2(\mathcal{Z}) + \lambda\|\boldsymbol{x}-\boldsymbol{c}_\lambda\|^2 \\
&\geq (1-\lambda)(\rho^2(C_i) + \lambda^2\rho^2(\mathcal{Z})) + \lambda(1-\lambda)^2\rho^2(\mathcal{Z}) \\
&= (1-\lambda)\rho^2(C_i) + \lambda(1-\lambda)\rho^2(\mathcal{Z}).
\end{aligned}$$

This expressions is a lower bound for any $\lambda \in [0,1]$ and is maximized at

$$\lambda = \frac{\rho^2(\mathcal{Z}) - \rho^2(C_i)}{2\rho^2(\mathcal{Z})}$$

where it takes value $\frac{\left(\rho^2(\mathcal{Z})+\rho^2(C_i)\right)^2}{4\rho^2(\mathcal{Z})}$. Therefore,

$$\rho(C_{i+1}) \geq \frac{\rho^2(\mathcal{Z}) + \rho^2(C_i)}{2\rho(\mathcal{Z})},$$

and we may think of $\rho^2(C_i)$ as a sequence with a lower bound generated by the mapping $x \mapsto \frac{\rho^2+x^2}{2\rho}$, which increases to its limit of $\rho^2(\mathcal{Z})$.

To obtain the convergence rate, define $x_i = \rho(C_{i+1})/\rho(\mathcal{Z})$. We then have $x_1 = 0$ and

$$x_{i+1} = \sum_{s=1}^{i}(x_{s+1} - x_s) = \sum_{s=1}^{i}\left(\frac{1+x_s^2}{2} - x_s\right) = \sum_{s=1}^{i}\frac{(1-x_s)^2}{2} \geq \frac{i}{2}(1-x_{i+1})^2,$$

and solving the quadratic inequality yields

$$\frac{\rho(C_{i+1})}{\rho(\mathcal{Z})} \geq 1 + \frac{1}{i} - \sqrt{\frac{2}{i} + \frac{1}{i^2}} \geq 1 - \sqrt{\frac{2}{i}}.$$

□

Proposition 28 makes the lower bound for the Hilbert game immediate, which leads finally to our main result.

Theorem 29. *The value of the squared norm game over any closed, bounded, convex set $\mathcal{X}$ with minimum enclosing ball of radius ρ is*

$$V(\mathcal{X}) = \rho^2\sum_{t=1}^{T}\alpha_T$$

and the minimax strategy (MM).

Proof. Lemma 27 immediately gives an upper bound, $V(\mathcal{X}) \leq \rho \sum_{t=1}^{T} \alpha_T$. The lower bound is an application of Proposition 28. Define $\mathcal{X}_i = \operatorname{conv}(C_i)$, where C_i is the sequence of finite sets guaranteed by the proposition. Since $\mathcal{X}_i \subseteq \mathcal{X}$, we must have $V(\mathcal{X}_i) \leq V(\mathcal{X})$, and so, applying Theorem 25 yields $\rho(\mathcal{X}_i)^2 \sum_{t=1}^{T} \alpha_t \leq V(\mathcal{X})$ for all i, and hence

$$\rho(\mathcal{X})^2 \sum_{t=1}^{T} \alpha_t \leq V(\mathcal{X}) \leq \rho(\mathcal{X})^2 \sum_{t=1}^{T} \alpha_t,$$

completing the proof.

□

2.7 Conclusion

This chapter calculated the value and minimax strategy of the squared loss game where the adversary is constrained to play in some closed, bounded, convex set. We proved that if ρ and $\boldsymbol{c}$ are the radius and center of the minimum enclosing ball, then the simple strategy

$$\boldsymbol{a}_n \;\coloneqq\; \boldsymbol{c} + \alpha_n \sum_{t=1}^{n-1} (\boldsymbol{x}_t - \boldsymbol{c}). \tag{MM}$$

is in fact minimax optimal. The player only needs to know $\boldsymbol{c}$ and T to play. Furthermore, the value is

$$\mathcal{R} \leq \rho^2 \sum_{t=1}^{T} \alpha_t$$

regardless of dimension. We have also calculated the value-to-go for ball-like simplices and ellipsoids and can play with subgame perfection (i.e. optimally respond to a non-optimal adversary). As promised, these methods are computationally efficient as all the α_t coefficients can be precomputed in $O(T)$ time.

Chapter 3

Minimax Online Linear Regression

Given: Covariate budget matrix $\boldsymbol{\Sigma} \succeq 0$
For $t = 1, 2, \ldots$

- Adversary reveals x_t
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

Learner's aim is to minimize regret,

$$\mathcal{R} := \sum_t (\hat{y}_t - y_t)^2 - \min_\theta \sum_t (\boldsymbol{x}_t^\intercal \theta - y_t)^2$$

Figure 3.1: online linear regression protocol

3.1 Introduction

Linear regression is a fundamental prediction problem in machine learning and statistics. In this chapter, we study a sequential version: on round t, the adversary chooses and reveals a covariate vector $\boldsymbol{x}_t \in \mathbb{R}^d$, the learner makes a real-valued prediction $\hat{y}_t$, the adversary chooses and reveals the true outcome y_t, and finally the learner is penalized by the square loss, $(\hat{y}_t - y_t)^2$. The protocol is described in Figure 3.1.

Since it is hopeless to guarantee a small loss (the adversary can always cause constant loss per round), we instead aim to guarantee that we are able to predict almost as well as the best fixed linear predictor in hindsight. Letting $\boldsymbol{x}_s^t$ and y_s^t denote $\boldsymbol{x}_s, \ldots, \boldsymbol{x}_t$ and $y_s, \ldots, y_t$, respectively, the regret of a strategy that predicts $\hat{y}_1^T$ is defined as

$$\mathcal{R}_T\left(\hat{y}_1^T, \boldsymbol{x}_1^T, y_1^T\right) := \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} (\theta^\intercal \boldsymbol{x}_t - y_t)^2.$$

A strategy $s : \bigcup_{t \geq 1} (\mathbb{R}^d \times \mathbb{R})^{t-1} \times \mathbb{R}^d \to \mathbb{R}$, is a map from observations to predictions, and we define $\mathcal{R}_T\left(s, \boldsymbol{x}_1^T, y_1^T\right) := \mathcal{R}_T\left(\hat{y}_1^T, \boldsymbol{x}_1^T, y_1^T\right)$ where $\hat{y}_t = s(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_{t-1}, y_{t-1}, \boldsymbol{x}_t)$. This chapter studies the minimax regret, which is the lowest possible regret of any strategy guaranteeable against all sequences chooses by the adversary, including sequences that adapt to the player strategy. That is, the minimax player strategy s^* is such that

$$\max_{\boldsymbol{x}_1^T, y_1^T} \mathcal{R}_T\left(s^*, \boldsymbol{x}_1^T, y_1^T\right) = \min_s \max_{\boldsymbol{x}_1^T, y_1^T} \mathcal{R}_T\left(s, \boldsymbol{x}_1^T, y_1^T\right),$$

where the min is over all strategies. Phrased another way, every other strategy can be made to suffer at least as much regret as the minimax strategy.

The minimax regret can also be written without explicitly defining a strategy as follows:

$$\mathcal{V} := \max_{\boldsymbol{x}_1} \min_{\hat{y}_1} \max_{y_1} \cdots \max_{\boldsymbol{x}_T} \min_{\hat{y}_T} \max_{y_T} \; \mathcal{R}_T. \tag{3.1}$$

This form may be more intuitive as it suggests how to compute the worst case strategies: solve the optimization problem, starting from time T and working backwards.

We have not specified the domains of any of the optimizations. Throughout, we will allow $\hat{y}_t$ to be unconstrained, but we will need to constrain the adversary to achieve any non-trivial guarantees.

In general, computing minimax strategies is computationally intractable: the optimal prediction $\hat{y}_t$ depends on the complete history $(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_{t-1}, y_{t-1}, \boldsymbol{x}_t)$, and the dependence might be a rather arbitrary function of this enormous space of histories. Hence, we will concentrate on constraint families that permit efficient solutions with a per-round complexity independent of T. Precisely, we search for settings were there are horizon-free minimax strategies. We require a strategy to minimize the worst case regret over covariate and label sequences, but we also insist that, for any time horizon T, the strategy incurs no more regret than any other strategy, *even a strategy that knows* T.

Definition 30. *A strategy s^* is horizon-independent minimax optimal for some class $\mathcal{A}$ of covariate sequences and some class $\mathcal{Y}(\boldsymbol{x}_1^T)$ of label sequences, possibly depending on $\boldsymbol{x}_1^T \in \mathcal{A}$, if*

$$\sup_T \left(\sup_{\boldsymbol{x}_1^T \in \mathcal{A},\, y_1^T \in \mathcal{Y}(\boldsymbol{x}_1^T)} \mathcal{R}_T\left(s^*, \boldsymbol{x}_1^T, y_1^T\right) - \min_s \sup_{\boldsymbol{x}_1^T \in \mathcal{A},\, y_1^T \in \mathcal{Y}(\boldsymbol{x}_1^T)} \mathcal{R}_T\left(s, \boldsymbol{x}_1^T, y_1^T\right) \right) = 0.$$

Outline and Our Contributions Our main result is that given a constraint $\boldsymbol{\Sigma}$ on the scale of the covariates, there is an efficiently computable minimax optimal strategy that can compete with every covariate and label sequence in some class (described in Section 3.7). The restrictions on the adversary essentially ensure that the adversary respects the scale constraint $\boldsymbol{\Sigma}$, so that the player is not led to under-regularize or over-regularize.

We proceed by analyzing a sequence of restrictions on the adversary of decreasing severity. We show that, under these conditions, the minimax strategy is always a simple, linear predictor. After t rounds, define the summary statistics

$$\boldsymbol{s}_t := \sum_{q=1}^{t} y_q \boldsymbol{x}_q\,, \qquad \sigma_t^2 := \sum_{q=1}^{t} y_q^2, \quad \text{and} \quad \boldsymbol{\Pi}_t := \sum_{q=1}^{t} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal. \tag{3.2}$$

The minimax strategy (we call it **mms**) predicts

$$\hat{y}_{t+1} \;=\; \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t, \tag{\textbf{mms}}$$

where the $\boldsymbol{P}_t$ matrices have an intricate definition we will describe shortly. We first note that taking $\boldsymbol{P}_t = \boldsymbol{\Pi}_t^\dagger$ or $\boldsymbol{P}_t = (\boldsymbol{\Pi}_t + \lambda \boldsymbol{I})^\dagger$ corresponds to follow the leader or ridge regression, respectively. The minimax algorithm is more sophisticated. For $\boldsymbol{\Sigma}_0 \succeq 0$, a limit on the size of the covariates, define

$$\boldsymbol{P}_0 := \boldsymbol{\Sigma}_0^\dagger, \; \boldsymbol{P}_t := \boldsymbol{P}_{t-1} - \frac{a_t}{b_t^2} \boldsymbol{P}_{t-1} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal \boldsymbol{P}_{t-1} \tag{3.3}$$

where

$$b_t^2 := \boldsymbol{x}_t^\intercal \boldsymbol{P}_{t-1} \boldsymbol{x}_t, \;\; a_t := \frac{\sqrt{4b_t^2+1}-1}{\sqrt{4b_t^2+1}+1}. \tag{3.4}$$

Hence, (**mms**) has per-round complexity $O(d^2)$. The rest of the chapter explains how we arrived at (3.3) and presents several families of constraints on the data, $\{\boldsymbol{x}_t, y_t\}$ where (**mms**) is the (efficiently computable) minimax strategy.

Intuitively, the minimax strategy begins by regularizing against all possible future covariate sizes defined by $\boldsymbol{\Sigma}_0$. Upon seeing $\boldsymbol{x}_t$, it lessens the regularization in the direction of $\boldsymbol{x}_t\boldsymbol{x}_t^\intercal$ since there is less budget left in that direction.

We will always impose the following *Adversarial Covariate Conditions*. For $\boldsymbol{P}_t$ defined by (3.3), we denote

$$\mathcal{A}(\boldsymbol{\Sigma}) := \left\{ \boldsymbol{x}_1^T : \text{for } \boldsymbol{P}_0 = \boldsymbol{\Sigma}^\dagger,\, \boldsymbol{P}_t^\dagger \succeq \boldsymbol{\Pi}_t \right\} \tag{3.5}$$

$$\overline{\mathcal{A}}(\boldsymbol{\Sigma}) := \left\{ \boldsymbol{x}_1^T : \text{for } \boldsymbol{P}_0 = \boldsymbol{\Sigma}^\dagger,\, \boldsymbol{P}_t^\dagger \succeq \boldsymbol{\Pi}_t \text{ and } \boldsymbol{P}_T^\dagger = \Pi_T \right\}. \tag{3.6}$$

The conditions essentially require that the scale of all the covariates $\boldsymbol{x}_1^T$ at the conclusion of the game are compatible with $\boldsymbol{\Sigma}$, the scale promised to the player at the start. We provide more interpretations of this condition and show that, without this condition, the adversary can cause arbitrarily high regret.

We begin our analysis in Section 3.3 with the easier case of fixed-design regression where all the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ are chosen by the adversary and revealed at the start of the game. This allows us to explicitly solve for the value-to-go starting from the back of the game, provided that the labels are bounded and the covariates satisfy an additional alignment condition. The backwards induction results in (**mms**) with $\boldsymbol{P}_t$ matrices defined by

$$\boldsymbol{P}_T = \boldsymbol{\Pi}_T^\dagger, \qquad\qquad \boldsymbol{P}_t = \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \tag{3.7}$$

which corresponds to the forwards $\boldsymbol{P}_t$ upon setting $\boldsymbol{\Sigma}_0$ to be the unique matrix that leads to $\boldsymbol{P}_T = \boldsymbol{\Pi}_T$. The $\boldsymbol{P}_t$ matrices are shown to have the alternative form

$$\boldsymbol{P}_t^{-1} = \sum_{q=1}^{t} \boldsymbol{x}_q\boldsymbol{x}_q^\intercal + \sum_{q=t+1}^{T} \frac{\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q}{1+\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q\boldsymbol{x}_q^\intercal, \tag{3.8}$$

and we immediately see that $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{\Sigma})$. The $\boldsymbol{P}_t$ matrices also have the form which intuitively is ridge regression with some future-instance-weighted regularization. We can view each term $\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$ as round t's contribution to the regret, and hence the $\boldsymbol{P}_t$ can be interpreted as an inverse second moment matrix when the outer products $\boldsymbol{x}_q\boldsymbol{x}_q^\intercal$ for unseen data are weighted according to their contribution to the minimax regret. We emphasize that $\boldsymbol{P}_t$ is a product of the minimax analysis; its elegant form showcases the beauty of the minimax approach.

Section 3.4 demonstrates that we can also be minimax in the fixed design case if we fix some $R \geq 0$ and require

$$\sum_{t=1}^{T} y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t \;\leq\; R.$$

There are no bounded label assumptions, and since the strategy is (**mms**), it clearly does not need knowledge of R and it is automatically adaptive in this sense. This suggests that the above quantity is a natural measure of the complexity of the constellation of labels and covariates for regression.

In order to study more general adversarial-design games, Section 3.5 shows that the forwards and backwards recursions are equivalent so long as $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{\Sigma})$. This immediately lifts the two previous fixed-design results to the adversarial-design setting. That is, we show that (**mms**) is minimax optimal for all sequences $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{\Sigma})$ and label constraints studied in Section 3.3 and 3.4.

However, our goal is to be minimax against sequences in $\mathcal{A}(\boldsymbol{\Sigma})$. To this end, we examine a hybrid game in Section 3.6 where the covariate sequence is fixed but the adversary is allowed to adaptively choose when to stop the game. Then, Section 3.7 relaxes the condition that $\boldsymbol{P}_t = \boldsymbol{\Pi}_T^\dagger$ with equality and show that, as long as the covariates satisfy a continuation condition, (**mms**) remains optimal. We proceed by analyzing a variant of the game where the adversary is allowed to stop at any point, and we derive conditions that ensure that the adversary causes maximum regret by not doing so.

Section 3.8 continues the intuition of (**mms**) as a modified ridge regression by casting the minimax algorithm as Follow-the-Regularized-Leader with a specific, time-dependent regularizer. Finally, Section 3.9 derives an $O(\log(T))$ regret bound for the worst case covariates.

Related Work

Linear regression is one of the classical problems in statistics and has been studied for over a century. The online version of linear regression is much more recent. [22] considered online linear regression with binary labels and ℓ_1-constrained parameters $\boldsymbol{w}$, and gave an $O(d\log(dT))$ regret bound for an ℓ_2-regularized follow-the-leader strategy. [11] considered ℓ_2-constrained parameters, and gave $O(\sqrt{T})$ regret bounds for a gradient descent algorithm with ℓ_2 regularization. [28] showed that an Exponentiated Gradient algorithm, based on relative entropy regularization, gives $O(\sqrt{T})$ regret. All of these results depend on the scale of the instances and labels. [53] applied the Aggregating Algorithm [52] to continuously many experts to arrive at an algorithm for online linear regression. This algorithm uses the inverse second moment matrix of past and current covariates, whereas the minimax strategy that we present uses the entire covariate sequence (see (3.11)). Vovk's algorithm was interpreted and re-analyzed in various ways [21, 4]: it is minimax optimal for the last trial, and it satisfies a $O(\log T)$ scale-dependent regret bound. The scale dependence is perhaps not surprising when future instances are not available. The regret bound we obtain for the minimax strategy is $O(\log T)$ with no dependence on the scale of the covariates. Refined work on "last-step minimax" was done by [39].

We take the approach of [50] and [30], who studied minimax optimal strategies for prediction games with squared loss: rather than proposing an algorithm that explicitly involves regularization and proving a regret bound, we identify the optimal minimax strategy for square loss; the ideal regularization emerges.

3.2 Fixed Design Linear Regression

In the fixed design setting, the game length T and covariates $\boldsymbol{x}_1^T$ are chosen and revealed before the game begins. The goal of this section will be to calculate the value-to-go and the minimax and maximin strategies of this game. Recall that the value is

$$\mathcal{V} := \min_{\hat{y}_1} \max_{y_1} \cdots \min_{\hat{y}_T} \max_{y_T} \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} (\theta^\intercal \boldsymbol{x}_t - y_t)^2$$

where we constrain $y_1^T \in \mathcal{Y}$ for some $\mathcal{Y}$ (we will provide several amenable examples later) but do not explicitly constrain $\hat{y}_1^t$. This formula encodes the requirement that $\hat{y}_t$ is a function of y_1^{t-1} and $\hat{y}_1^{t-1}$ and y_t is a function of y_1^t and $\hat{y}_1^{t-1}$. We emphasize that in the fixed design setting, every y_t and $\hat{y}_t$ can be a function of $\boldsymbol{x}_1^T$.

The standard technique to solve for the value is known as backwards induction. Define the value to go function recursively with base case

$$V\left(\boldsymbol{s}_T, \sigma_T^2, T\right) := -\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} \left(\theta^\intercal \boldsymbol{x}_t - y_t\right)^2$$

and induction

$$V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) := \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left((\hat{y}_{t+1} - y_{t+1})^2 + V\left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right) \right)$$

The value-to-go function at time n, $V(y_1^n, \boldsymbol{x}_1^T)$, is the regret if y_1^n and $\hat{y}_1^n$ have been played and both players play optimally from round $n+1$ onward with the loss of the fixed actions of the past $\sum_{t=1}^{n} (\hat{y}_n - y_n)^2$ subtracted off (they do not affect the strategies). These equations are naturally solved working backwards from $n = T$, hence the name.

Generally, calculating the value-to-go is computationally intractable: we must evaluate $V(y_1^n, \boldsymbol{x}_1^T)$ for every possible sequence of plays, which suggests, and often results in, complexity exponential in T. The contribution of this work can alternatively be understood as specifying when this exponential blow-up does not occur. In fact, the value-to-go can be written as a succinct function of $\boldsymbol{s}_n$ and σ_n only.

The Offline Problem

Lemma 31. *Fix data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$. The loss of the best linear predictor in hindsight is*

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{t=1}^{T} (\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t)^2 &= \sum_{t=1}^{T} y_t^2 - \left(\sum_{t=1}^{T} y_t \boldsymbol{x}_t\right)^\intercal \left(\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^\dagger \left(\sum_{t=1}^{T} y_t \boldsymbol{x}_t\right) \\ &= \sigma_T^2 - \boldsymbol{s}_T^\intercal \boldsymbol{\Pi}_T^\dagger \boldsymbol{s}_T \end{aligned}$$

where $\dagger$ denotes pseudo-inverse (any generalized inverse will do). It is minimized by

$$\boldsymbol{w} = \boldsymbol{\Pi}_T^\dagger \boldsymbol{s}_T.$$

3.3 Minimax Analysis for Bounded Labels

In this section we perform a minimax analysis of fixed-design linear regression with bounded labels y_t and give an exact expression for the minimax regret. As discussed in the introduction, the following problem-weighted inverse covariate matrices are central to the analysis and algorithm.

Recall the backwards recursion for the $\boldsymbol{P}_t$ matrices:

$$\boldsymbol{P}_T = \left(\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal \right)^\dagger, \qquad \boldsymbol{P}_t = \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1} \boldsymbol{x}_{t+1} \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}.$$

In the proof, it becomes clear that the $\boldsymbol{P}_t$ arise exactly from solving the minimax problem.

Theorem 32. *Fix a constant $B > 0$ and a sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{R}^d$. Consider the following T-round game. On round $t \in \{1, \ldots, T\}$, the player first chooses $\hat{y}_t \in \mathbb{R}$, then the adversary chooses $y_t \in [-B, B]$ and the player incurs loss $(\hat{y}_t - y_t)^2$. The value of this game is*

$$\min_{\hat{y}_1} \max_{y_1} \cdots \min_{\hat{y}_T} \max_{y_T} \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{t=1}^{T} (\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t)^2.$$

Assume that the following covariate condition holds:

$$\sum_{q=1}^{t-1} |\boldsymbol{x}_q^\intercal \boldsymbol{P}_t \boldsymbol{x}_t| \;\le\; 1 \quad \textit{for all } 1 \le t \le T. \tag{3.9}$$

Then the value of the game is $B^2 \sum_{t=1}^{T} \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$, the optimal strategy is (**mms**)*: $\hat{y}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t$, where $\boldsymbol{s}_t = \sum_{q=1}^{t} y_q \boldsymbol{x}_q$, and the maximin probability distribution assigns*

$$\Pr(y_{t+1} = \pm B) = 1/2 \pm \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t / (2B).$$

The proof shows that the minimax strategy optimizes the value-to-go, and therefore optimally exploits suboptimal play by the adversary.

Proof. The proof proceeds by explicitly performing the backwards induction and showing

$$V(\boldsymbol{s}_t, \sigma_t^2, t) = \boldsymbol{s}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_t - \sigma_t^2 + \gamma_t,$$

where the γ_t coefficients are recursively defined as

$$\gamma_T = 0, \qquad \gamma_t = \gamma_{t+1} + B^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{x}_{t+1}.$$

This implies that the value of the game is $V(\boldsymbol{0}, 0, 0) = \gamma_0 = B^2 \sum_{t=1}^{T} \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$, as desired. Lemma 31 establishes the base case $V\left(\boldsymbol{s}_T, \sigma_T^2, T\right) = \boldsymbol{s}_T^\intercal \boldsymbol{P}_T \boldsymbol{s}_T - \sigma_T^2$. Now, assuming the induction hypothesis

$$V(\boldsymbol{s}_{t+1}, \sigma_{t+1}^2, t+1) = \boldsymbol{s}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_{t+1} - \sigma_{t+1}^2 + \gamma_{t+1},$$

we have

$$\begin{aligned}
V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &= \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left(\hat{y}_{t+1} - y_{t+1}\right)^2 + V\left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right) \\
&= \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left(\hat{y}_{t+1} - y_{t+1}\right)^2 + \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right)^\intercal \boldsymbol{P}_{t+1} \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right) \\
&\qquad - \left(\sigma_t^2 + y_{t+1}^2\right) + \gamma_{t+1} \\
&= \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left(\hat{y}_{t+1}^2 - 2\hat{y}_{t+1}y_{t+1} + 2y_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t + y_{t+1}^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right. \\
&\qquad + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1} \\
&= \min_{\hat{y}_{t+1}} \hat{y}_{t+1}^2 + \left(\max_{y_{t+1}} 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \hat{y}_{t+1}\right) y_{t+1} + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1} y_{t+1}^2\right) \\
&\qquad + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1}.
\end{aligned}$$

The inner maximization is a quadratic in $y_{t+1} \in [-B, B]$ with a non-negative second derivative, so it is maximized by an extreme $y_{t+1} \in \{-B, B\}$, giving

$$\begin{aligned}
V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) = \min_{\hat{y}_{t+1}} &\left(\hat{y}_{t+1}^2 + 2B\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \hat{y}_{t+1}\right|\right) \\
&\qquad + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1} B^2 + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1}.
\end{aligned}$$

The minimization over $\hat{y}_{t+1}$ is of a convex function, which is minimized when 0 is in the subgradient, so that

$$\hat{y}_{t+1} = \begin{cases} -B & \text{if } \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t < -B, \\ B & \text{if } \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t > B, \\ \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t & \text{otherwise.} \end{cases} \tag{3.10}$$

Under the assumption (3.9) of the theorem, only the last case occurs:

$$\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t\right| = \left|\sum_{q=1}^{t} \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_q y_q\right| \le \sum_{q=1}^{t} \left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_q\right| |y_q| \le B,$$

so we have $\hat{y}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t$. Plugging this solution in, we find

$$\begin{aligned}
V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &= \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t + B^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1} + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1} \\
&= \boldsymbol{s}_t^\intercal \left(\boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\right)\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1} + B^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1},
\end{aligned}$$

verifying the recursion for $\boldsymbol{P}_t$ and γ_t. From the perspective of the adversary, we need to solve

$$\begin{aligned}
&\max_{p\in[0,1]} \min_{\hat{y}_{t+1}} \hat{y}_{t+1}^2 + 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \hat{y}_{t+1}\right)(2p-1)B + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1} B^2 \\
= \quad &\max_{p\in[0,1]} B(2p-1)^2 + 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - B(2p-1)\right)(2p-1)B + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1} B^2.
\end{aligned}$$

because the minimizer of $\hat{y}_{t+1}$ is the mean $B(2p-1)$. Setting the p-derivative to zero results in worst-case probability $1/2 \pm \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t/(2B)$ on $\pm B$. □

As discussed in the introduction, the $\boldsymbol{P}_t$ have an alternative form:

Lemma 33. *For the $\boldsymbol{P}_t$ matrices defined in* (3.7)*, we have*

$$\boldsymbol{P}_t^{-1} = \sum_{q=1}^{t} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal + \sum_{q=t+1}^{T} \frac{\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q}{1 + \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal. \tag{3.11}$$

Proof. The proof is by induction. We start with

$$\boldsymbol{P}_T^{-1} = \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal.$$

Suppose the equation in the lemma is true for $1 < t \leq T$. Then by the Sherman-Morrison formula,

$$\begin{aligned}
\boldsymbol{P}_{t-1}^{-1} &= \left(\boldsymbol{P}_t + \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal \boldsymbol{P}_t\right)^{-1} \\
&= \boldsymbol{P}_t^{-1} - \frac{\boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} \\
&= \sum_{q=1}^{t} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal + \sum_{q=t+1}^{T} \frac{\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q}{1 + \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal - \frac{\boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} \\
&= \sum_{q=1}^{t-1} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal + \sum_{q=t}^{T} \frac{\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q}{1 + \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal.
\end{aligned}$$

□

Condition (3.9) can be easily tested; it does not involve the labels y_t. It can be viewed as forbidding outlier covariates: an x_t that is large relative to the others will cause the condition to fail, leading to clipping in (3.10). The condition appears to be restrictive: it is satisfied if the covariates are approximately orthonormal, which essentially corresponds to playing d interleaved independent one-dimensional regression problems, but we do not know of other problem instances that satisfy the condition.

The condition arises because of the uniform constraint on the labels. There are, however, many other constraint sets for which the same strategy is still minimax optimal, but the corresponding conditions are milder. In particular, it is clear that the proof extends immediately to the case in which the adversary is constrained to choose label sequences from

$$\mathcal{Y}_B := \left\{(y_1, \dots, y_T) : |y_t| \leq B_t\right\}, \tag{3.12}$$

provided that the $B = (B_1, \dots, B_T)$ are compatible with the data by satisfying

$$B_t \geq \sum_{q=1}^{t-1} \left|\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_q\right| B_q. \tag{3.13}$$

In this case, the minimax regret is $\sum_{t=1}^T B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$ and the maximin probability distribution for y_{t+1} puts weight $1/2 \pm \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t/(2B_{t+1})$ on $\pm B_{t+1}$. Condition (3.9) is a special case of these compatibility constraints (3.13) corresponding to $B_1 = \cdots = B_T$.

We end this section by proving one useful (and maybe surprising) fact.

Lemma 34. *Let $\boldsymbol{x}_1^T$ be any covariate sequence and $\boldsymbol{P}_1, \dots, \boldsymbol{P}_T$ the associated precision matrices given by the backwards recursion* (3.7). *For any invertible matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, let $\boldsymbol{x}'_t = \boldsymbol{W}\boldsymbol{x}_t$. Then the precision matrices of $\boldsymbol{x}'_1, \dots, \boldsymbol{x}'_T$ are exactly $\boldsymbol{P}'_t = \boldsymbol{W}^{\dagger\intercal}\boldsymbol{P}_t\boldsymbol{W}^\dagger$ and $\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t = \boldsymbol{x}_t'^\intercal \boldsymbol{P}'_t \boldsymbol{x}'_t$.*

Proof. First, we can easily check that $\boldsymbol{P}'_T = \left(\sum_{t=1}^T \boldsymbol{x}'_t \boldsymbol{x}_t'^\intercal\right)^\dagger = (\boldsymbol{W}^\intercal)^\dagger \left(\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^\dagger \boldsymbol{W}^\dagger$. Now, assume that the hypothesis holds for t. Then

$$\begin{aligned}\boldsymbol{P}'_{t-1} =& \boldsymbol{P}'_t + \tilde{\boldsymbol{P}}_t \boldsymbol{x}'_t \boldsymbol{x}_t'^\intercal \boldsymbol{P}'_t \\ =& \boldsymbol{W}^{\dagger\intercal}\boldsymbol{P}_t\boldsymbol{W}^\dagger + \boldsymbol{W}^{\dagger\intercal}\left(\boldsymbol{P}_t\boldsymbol{W}^\dagger\boldsymbol{W}\boldsymbol{x}_t\boldsymbol{x}_t\boldsymbol{W}^\intercal\boldsymbol{W}^{\dagger\intercal}\boldsymbol{P}_t\right)W^\dagger \\ =& \boldsymbol{W}^{\dagger\intercal}\boldsymbol{P}_{t-1}\boldsymbol{W}^\dagger.\end{aligned}$$

□

3.4 Ellipsoidal Constraints

In addition to the box constraints, we are also able to provide the minimax strategy for another family of constraints. In this section we investigate another way of budgeting that is suggested by the problem. Namely, for some $R \geq 0$, we consider the set

$$\mathcal{Y}_R := \left\{(y_1, \dots, y_T) \in \mathbb{R} : \sum_{t=1}^T y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t = R\right\} \tag{3.14}$$

of label sequences with a certain weighted 2-norm, where the weights are related to the hardness of the covariates. We analyze the minimax fixed design linear regression problem (3.1) on $\mathcal{Y}_R$, and show that the minimax strategy is again the simple linear strategy (**mms**). Recall that this strategy predicts

$$\hat{y}_{t+1} \;=\; \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t.$$

This is surprising for two reasons. First, this predictor does not incorporate knowledge of R. Second, there is no easy relation between R and the maximum label magnitude $B_{\max} := \max_t |y_t|$. As the minimax regret bound of Section 3.9 deteriorates with B^2, one might conjecture that the performance also degenerates. However, to the contrary, we show that the regret of the predictor (**mms**) now *equals*

$$\mathcal{R}_T \;=\; \sum_{t=1}^T y_t^2 \boldsymbol{x}_t \boldsymbol{P}_t \boldsymbol{x}_t.$$

This means that this algorithm has two very special properties. First, it is a strong equalizer in the sense that it suffers the same regret on all 2^T sign-flips of the labels. And second, it is adaptive to the complexity R of the labels.

The regret this algorithm incurs is better than the minimax regret with $B = B_{\max}$ under Condition (3.9). Still, it inherits the $B_{\max}^2 d \log T$ bound. In addition, minimax optimality for the family of constraints $\mathcal{Y}_R$ is stronger than the corresponding result for the family of box constraints $\mathcal{Y}_B$ defined in (3.12), in the sense that, given some budget R and a sequence of B_ts that satisfy the compatibility inequalities (3.13), we can rescale the B_ts so that $\mathcal{Y}_B$ is contained in $\mathcal{Y}_R$, but the minimax regret is the same in both cases.

We proceed in two steps. We characterize the worst-case regret of the simple linear predictor (**mms**) on the set $\mathcal{Y}_R$. Then we argue that the worst-case regret of any predictor is at least as large.

Lemma 35. *Let $\boldsymbol{P}_t$ be as defined in* (3.7). *For all $y_1, \ldots, y_T$, strategy* (**mms**) *has regret*

$$\mathcal{R}_T \;=\; \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{t=1}^{T} (\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t)^2 \;=\; \sum_{t=1}^{T} y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t. \tag{3.15}$$

Proof. The worst-case (over labels) slack in (3.15) can be recursively calculated by

$$\begin{aligned} F\left(\boldsymbol{s}_T, \sigma_T^2, T\right) &\coloneqq - \min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{t=1}^{T} \left(\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t\right)^2 - \sum_{t=1}^{T} y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t, \\ F\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &\coloneqq \max_{y_{t+1}} \left(\left(\hat{y}_{t+1} - y_{t+1}\right)^2 + F\left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right)\right). \end{aligned}$$

Note that the max over $y_1, \ldots, y_T$ of the difference between left and right hand side of (3.15) is equal to $F(\boldsymbol{0}, 0, 0)$. We now show by induction that

$$F(\boldsymbol{s}_t, \sigma_t^2, t) = \boldsymbol{s}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t} y_q^2 \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q.$$

Lemma 31 verifies the base case. To check the inductive step, we calculate

$$\begin{aligned} F\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &= \max_{y_{t+1}} \left(\left(\hat{y}_{t+1} - y_{t+1}\right)^2 + F\left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right)\right) \\ &= \hat{y}_{t+1}^2 + \max_{y_{t+1}} 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \hat{y}_{t+1}\right) y_{t+1} + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{x}_{t+1} y_{t+1}^2 \\ &\qquad\qquad + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t+1} y_q^2 \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q \\ &= \hat{y}_{t+1}^2 + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t} y_q^2 \boldsymbol{x}_q \boldsymbol{P}_q \boldsymbol{x}_q + \max_{y_{t+1}} 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \hat{y}_{t+1}\right) y_{t+1}, \end{aligned}$$

where the $\max$ term on the right is zero by the choice of $\hat{y}_{t+1}$. Finally, we obtain

$$F\left(\boldsymbol{s}_t, \sigma_t^2, t\right) = \boldsymbol{s}_t^\intercal \left(\boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\right)\boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t} y_q^2 \boldsymbol{x}_q \boldsymbol{P}_q \boldsymbol{x}_q$$

as desired. The theorem statement is immediate upon noting that $F(\mathbf{0}, 0, 0) = 0$. □

Using this result, we can show that our predictor is minimax optimal.

Theorem 36. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ be fixed and let $\boldsymbol{P}_t$ be the corresponding prediction matrices. Then for every R, strategy* (**mms**) *is minimax optimal on the set of labelings $\mathcal{Y}_R$ as defined in* (3.14).

Proof. First, note that strategy (**mms**) suffers regret R on every y_t sequence in $\mathcal{Y}_R$. Now fix any predictor $\mathfrak{X}$ and consider the label sequence $0, \dots, 0, \pm\sqrt{\frac{R}{\boldsymbol{x}_T^\intercal \boldsymbol{P}_T \boldsymbol{x}_T}}$, where the sign of the label in the last round is chosen to oppose the sign of the predictor's prediction. Our predictor (**mms**) predicts the first $T-1$ perfectly and is at least as good on the Tth round. So on *every* round, predictor $\mathfrak{X}$ incurs at least the loss of (**mms**), and hence its worst-case regret is at least R. Thus, (**mms**), which incurs regret exactly R, is minimax optimal. □

For some fixed sequence of label budgets $B_1, B_2, \dots > 0$, define

1. *Label constraints on y_t*: $\mathcal{L}(B_t) := \{y_1^T : |y_t| \le B_t\}$
2. *Box constraints on $\boldsymbol{x}_t$*: $\mathcal{B}(B_t) := \left\{x_t : B_t \ge \sum_{s=1}^{t-1} |\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_s|\, B_s \text{ for } 2 \le t\right\}$.
3. *Implicit box constraints*: $\mathcal{B}(\boldsymbol{x}_1^T, b) := \mathcal{B}\left(B_1^T(\boldsymbol{x}_1^T, b)\right)$, where $B_1^T(\boldsymbol{x}_1^T, b)$ is the component-wise minimum of the set of compatible box constraints, $\mathcal{C}(\boldsymbol{x}_1^t, b))$, defined by

$$\mathcal{C}(\boldsymbol{x}_1^T, b) := \left\{ B_1^T : B_1 = b,\ B_t \ge \sum_{s=1}^{t-1} |\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_s|\, B_s \text{ for } 2 \le t \le T \right\}. \tag{3.16}$$

4. *Ellipsoidal constraints*: $\mathcal{E}(\boldsymbol{x}_1^T, R) := \left\{y_1^T : \sum_{t=1}^T y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t \le R\right\}$.

To summarize the fixed-design results of the previous two sections, we have the following corollary to Theorem 32.

Corollary 37. *For each x_1^T, the corresponding strategy* (**mms**) *is minimax optimal with respect to $\mathcal{B}(B_1^T)$, $\mathcal{B}(\boldsymbol{x}_1^T, b)$, and $\mathcal{E}(\boldsymbol{x}_1^T, R)$, for any $B_t > 0$, $b > 0$, and $R > 0$, in the following sense.*

(1) *If B_1^T satisfies $B_1^T \in \mathcal{C}(\boldsymbol{x}_1^T, B_1)$, then*

$$\sup_{y_1^T \in \mathcal{B}(B_1^T)} R_T(s^*, \boldsymbol{x}_1^T, y_1^T) = \min_s \sup_{y_1^T \in \mathcal{B}(B_1^T)} R_T(s, \boldsymbol{x}_1^T, y_1^T) = \sum_{t=1}^T B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t,$$

(3)
$$\sup_{y_1^T \in \mathcal{E}(\boldsymbol{x}_1^T, R)} R_T(s^*, \boldsymbol{x}_1^T, y_1^T) = \min_s \sup_{y_1^T \in \mathcal{E}(\boldsymbol{x}_1^T, R)} R_T(s, \boldsymbol{x}_1^T, y_1^T) = R.$$

3.5 The Forward Algorithm

The previous section described the fixed-design minimax strategy and established sufficient conditions for its optimality. Unfortunately, $\boldsymbol{P}_t$ is recursively defined as a function of the entire $\boldsymbol{x}_1^T$ sequence, so, at first, it seems difficult to remove the fixed-design requirement.

The key insight is that if $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{\Sigma})$, then the forward recursion is exactly equal to the backwards recursion found in the value-to-go of the last section.

If we provide the player with an initial precision matrix $\boldsymbol{P}_0$, then the natural algorithm that emerges is to play (**mms**), i.e., $\hat{y}_t := \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_{t-1}$, except with the $\boldsymbol{P}_t$ matrices defined by the forwards recursion. The prediction $\hat{y}_t$ is a function of $\boldsymbol{x}_1^{t-1}$ and y_1^{t-1} only and can be used in online covariate settings. The algorithm needs $O(d^2)$ memory and at each round the computational complexity is $O(d^2)$.

Our first result is that this algorithm is actually minimax optimal if we constrain the adversary to play in $\overline{\mathcal{A}}(\boldsymbol{\Sigma})$. Another interpretation is that $\boldsymbol{\Sigma}$ encodes all the necessary scale information the learner needs to optimally respond. In particular, the learner does not need to know T.

Theorem 38. *For all positive semidefinite $\boldsymbol{\Sigma}$, $B_1, B_2, \ldots > 0$, $b > 0$, and $R > 0$, the minimax strategy* (**mms**) *using the forward recursion* (3.3) *is horizon-independent minimax optimal, i.e.*

$$\sup_T \sup_{\boldsymbol{x}_1^T \in \mathcal{A}} \left(\sup_{y_1^T \in \mathcal{Y}(\boldsymbol{x}_1^T)} R_T(s^*, \boldsymbol{x}_1^T, y_1^T) - \min_s \sup_{y_1^T \in \mathcal{Y}(\boldsymbol{x}_1^T)} R_T(s, \boldsymbol{x}_1^T, y_1^T) \right) = 0.$$

with respect to the following $(\mathcal{A}, \mathcal{Y}(\boldsymbol{x}_1^t))$:

$$(\overline{\mathcal{A}}(\boldsymbol{\Sigma}), \mathcal{E}(\boldsymbol{x}_1^T, R)), \qquad (\mathcal{A}(B_1^T) \cap \overline{\mathcal{A}}(\boldsymbol{\Sigma}), \mathcal{B}(B_1^T)), \qquad (\overline{\mathcal{A}}(\boldsymbol{\Sigma}), \mathcal{B}(\boldsymbol{x}_1^T, b)).$$

That is, s^* performs as well as the best strategy that sees the covariate sequence in advance.

The proof is intuitively simple: if the adversary plays in $\overline{\mathcal{A}}(\boldsymbol{\Sigma})$, then the forward recursion begins at the matrix that make the $\boldsymbol{P}_t$ generated by the forwards and backwards recurrences identical; hence, the strategy exactly corresponds to the minimax strategy.

Lemma 39. *For any fixed covariate sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ satisfying*

$$\boldsymbol{\Sigma} = \sum_{s=1}^{T} \frac{\boldsymbol{x}_s^\intercal \boldsymbol{P}_s \boldsymbol{x}_s}{1 + \boldsymbol{x}_s^\intercal \boldsymbol{P}_s \boldsymbol{x}_s} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal,$$

the forward matrices $\boldsymbol{P}_t$ defined by (3.3) *are identical to the $\boldsymbol{P}_t$ matrices defined by the backwards recursion* (3.7).

Proof. Let $\boldsymbol{P}_t$ be defined in the forwards recursion and $\boldsymbol{P}_t'$ denote the backwards recursion in Equation (3.7). The lemma assumes the base case $\boldsymbol{P}_T = \boldsymbol{P}_T'$. Now, given the induction hypothesis

$\boldsymbol{P}'_t = \boldsymbol{P}_t$, we show

$$\begin{aligned}
\boldsymbol{P}'_{t-1} &= \boldsymbol{P}'_t + \boldsymbol{P}'_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal \boldsymbol{P}'_t \\
&= \boldsymbol{P}_{t-1} - \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} + (\boldsymbol{P}_{t-1} - \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1})\boldsymbol{x}_t\boldsymbol{x}_t^\intercal(\boldsymbol{P}_{t-1} - \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1}) \\
&= \boldsymbol{P}_{t-1} - \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} + \boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} - \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} \\
&- \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} + \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} \\
&= \boldsymbol{P}_{t-1} - \frac{a_t}{b_t^2}\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} + \boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} - 2a_t\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} \\
&\quad + a_t^2\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} \\
&= \boldsymbol{P}_{t-1} + \left(-\frac{a_t}{b_t^2} + 1 - 2a_t + a_t^2\right)\boldsymbol{P}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_{t-1} \\
&= \boldsymbol{P}_{t-1},
\end{aligned}$$

which is true since $\left(-\frac{a_t}{b_t^2} + 1 - 2a_t + a_t^2\right) = (-(1-a_t)^2 + (1-a_t)^2) = 0.$ □

We can now complete the proof.

Proof. **of Theorem 38** First, assume that $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{\Sigma})$. Theorem 32 along with Lemma 39 gives that

$$\sup_{y_1^T \in \mathcal{B}(B_1^T)} R_T(s^*, \boldsymbol{x}_1^T, y_1^T) - \min_s \sup_{y_1^T \in \mathcal{B}(B_1^T)} R_T(s, \boldsymbol{x}_1^T, y_1^T) = 0.$$

Since this holds for all $\boldsymbol{x}_1^T$, we actually get the stronger result

$$\sup_T \sup_{\boldsymbol{x}_1^T \in \mathcal{A}(B_1^T) \cap \overline{\mathcal{A}}(\boldsymbol{\Sigma})} \left(\sup_{y_1^T \in \mathcal{B}(B_1^T)} R_T(s^*, \boldsymbol{x}_1^T, y_1^T) - \min_s \sup_{y_1^T \in \mathcal{B}(B_1^T)} R_T(s, \boldsymbol{x}_1^T, y_1^T) \right) = 0.$$

Identical reasoning extends Theorem 36 to the adversarial covariate context. □

Before ending this section, we prove the alternative interpretation of $\mathcal{A}(\boldsymbol{\Sigma})$ described in the introduction. The *Adversarial Covariate Conditions* can alternatively be written using the backwards recursion,

$$\begin{aligned}
\mathcal{A}(\boldsymbol{\Sigma}) &= \left\{ \boldsymbol{x}_1^T : \text{for } \boldsymbol{P}_0, \ldots, \boldsymbol{P}_T \text{ defined by (3.7), } \boldsymbol{P}_0^\dagger \preceq \boldsymbol{\Sigma} \right\}, \\
\overline{\mathcal{A}}(\boldsymbol{\Sigma}) &= \left\{ \boldsymbol{x}_1^T : \text{for } \boldsymbol{P}_0, \ldots, \boldsymbol{P}_T \text{ defined by (3.7), } \boldsymbol{P}_0^\dagger = \boldsymbol{\Sigma} \right\},
\end{aligned} \tag{3.17}$$

The equivalence of the two definitions of $\overline{\mathcal{A}(\boldsymbol{\Sigma})}$ is a simple consequence of Lemma 39. The equivalence of $\mathcal{A}(\boldsymbol{\Sigma})$ requires the following lemma.

Lemma 40. *For any $t \geq 0$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$, and symmetric matrix $\boldsymbol{P} \succeq 0$, the following two conditions are equivalent:*

1. $\boldsymbol{P}^\dagger \succeq \boldsymbol{\Pi}_t$
2. *For any $T \geq t + k$, where $k = \operatorname{rank}\left(\boldsymbol{P}^\dagger - \boldsymbol{\Pi}_t\right)$, there is a continuation of the covariate sequence, $\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_T$, such that setting $\boldsymbol{P}_t = \boldsymbol{P}$ and defining $\boldsymbol{P}_{t+1}, \ldots, \boldsymbol{P}_T$ by the forward recursion* (3.3) *gives $\boldsymbol{P}_T^\dagger = \boldsymbol{\Pi}_T$.*

Proof. To see that Condition 1 implies Condition 2, we will consider the forward algorithm recursion, starting from $\boldsymbol{P}_t = \boldsymbol{P}$, and show that we can find suitable covariate vectors $\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_{t+k}$, so that

$$\operatorname{rank}\left(\boldsymbol{P}_{t+i}^\dagger - \sum_{s=1}^{t+i} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal\right) = k - i,$$

which implies the result for $T = t + k$. It suffices to show that, at each step, we can reduce this rank by one. Consider the spectral decomposition

$$\boldsymbol{P}^\dagger - \boldsymbol{\Pi}_t = \sum_{i=1}^{m} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal,$$

for orthonormal $v_1, \ldots, v_k$ and non-negative $\lambda_1 \geq \cdots \geq \lambda_k > 0$. Choosing $\boldsymbol{x}_{t+1} = \beta \boldsymbol{v}_k$, there is a $\beta \geq 0$ such that

$$\boldsymbol{P}_{t+1}^\dagger - \boldsymbol{\Pi}_{t-1} = \sum_{i=1}^{k-1} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal,$$

which implies the result. Indeed, we have

$$\begin{aligned}
\boldsymbol{P}_{t+1}^\dagger - \boldsymbol{\Pi}_{t+1} &= \boldsymbol{P}_t^\dagger + \frac{a_{t+1}\beta^2}{(1 - a_{t+1})b_{t+1}^2} \boldsymbol{v}_k \boldsymbol{v}_k^\intercal - \boldsymbol{\Pi}_t - \beta^2 \boldsymbol{v}_{t+1} \boldsymbol{v}_{t+1}^\intercal \\
&= \sum_{i=1}^{k-1} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal + \left(\lambda_k - \beta^2 + \frac{a_{t+1}\beta^2}{(1 - a_{t+1})b_{t+1}^2}\right) \boldsymbol{v}_k \boldsymbol{v}_k^\intercal.
\end{aligned}$$

Recall

$$\begin{aligned}
b_{t+1}^2 &= \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_t \boldsymbol{x}_{t+1} \\
&= \beta^2 \boldsymbol{v}_k^\intercal \left(\boldsymbol{\Pi}_t + \sum_{i=1}^{k} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal\right)^\dagger \boldsymbol{v}_k \\
&= \beta^2 c^2,
\end{aligned}$$

where we have defined $c^2 > 0$. We need to choose $\beta \geq 0$ so that

$$\begin{aligned}\lambda_k &= \beta^2\left(1 - \frac{a_{t+1}}{(1-a_{t+1})b_{t+1}^2}\right)\\ &= \beta^2\left(1 - \frac{\sqrt{4b_t^2+1}-1}{2b_t^2}\right)\\ &= \beta^2\left(1 - \frac{\sqrt{4\beta^2c^2+1}-1}{2\beta^2c^2}\right)\\ \Leftrightarrow \qquad c^2\lambda_k &= \beta^2c^2 - \frac{\sqrt{4\beta^2c^2+1}-1}{2}.\end{aligned}$$

Since $c^2\lambda_k \geq 0$ and the function on the right hand side maps to $[0,\infty)$ for $\beta \geq 0$, there is a suitable choice of β. To see that this implies the result for any $T \geq t+k$, notice that by choosing a smaller value of β, the rank is not diminished.

To see the other direction, notice that Condition 2 and Lemma 39 together imply that there is a T and a completion of the sequence, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$, so that plugging the sequence into the backwards recurrence (3.7) gives $\boldsymbol{P}_t = \boldsymbol{P}$. But then Equation (3.11) shows that

$$\boldsymbol{P}_t^\dagger = \boldsymbol{\Pi}_t + \sum_{s=t+1}^{T} \frac{\boldsymbol{x}_s^\intercal \boldsymbol{P}_s \boldsymbol{x}_s}{1 + \boldsymbol{x}_s^\intercal \boldsymbol{P}_s \boldsymbol{x}_s} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal \succeq \boldsymbol{\Pi}_t,$$

which is Condition 1. □

3.6 Exhausting the Budget

The last section presented the minimax strategy when $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{\Sigma})$. This section examines a different method for relaxing the fixed-design assumption: fixing a covariate sequence $\boldsymbol{x}_1^T$ but allowing the adversary to control the length of the game. We perform a minimax analysis and establish conditions that ensure that the adversary maximizes regret by using up the entire covariance budget $\boldsymbol{\Sigma}$.

We encapsulate the adversary's decision by the variable e_t, which is equal to one if the adversary decides to play round t and zero otherwise. For some fixed T and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$, the game protocol is: at round t, the player predicts $\hat{y}_t$, the adversary chooses e_t and if $e_t = 1$, plays the response y_t and causes the learner $(\hat{y}_t - y_t)^2$ loss. We call this the fixed design game with early stopping.

We first define the additional loss suffered by the comparator from playing t rounds instead of $t-1$ rounds as

$$\Delta_t^* := \min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} (\theta^\intercal \boldsymbol{x}_s - y_s)^2 - \min_{\theta' \in \mathbb{R}^d} \sum_{s=1}^{t-1} (\theta'^\intercal \boldsymbol{x}_s - y_s)^2 \tag{3.18}$$

so that $\Delta_t^* \geq 0$ and $L_T^* = \sum_{t=1}^{t} \Delta_t^*$. The regret of the game with early stopping can be written as

$$\mathcal{R}_T = \sum_{t=1}^{T} \left(\prod_{s=1}^{t} e_s\right) \left((y_t - \hat{y}_t)^2 - \Delta_t^*\right).$$

If the adversary chooses $y_t = {\theta^*_{t-1}}^\intercal \boldsymbol{x}_t$ where θ^*_{t-1} is the ordinary least squares solution on data through time $t-1$, then $\delta^*_t = 0$. Since the loss of the player is always non-negative, the regret increases. However, this y_t might not be feasible, and so this section identifies necessary and sufficient conditions where the adversary maximizing regret by playing all rounds. This intuition will be used in the adversarial covariates section to argue when the adversary will use the entire covariance budget.

We proceed by computing the minimax strategy for this augmented game and characterize when (**mms**) is, once again, the minimax strategy. To this end, we define the instantaneous value-to-go $W_t(\boldsymbol{s}_t, \sigma^2_t, \boldsymbol{\Pi}_t)$ by $W_T(\boldsymbol{s}_T) = 0$ and

$$W_{t-1}(\boldsymbol{s}_{t-1}, \sigma^2_{t-1}, \boldsymbol{\Pi}_{t-1}) = \max_{e_t \in \{0,1\}} e_t \left(\min_{\hat{y}_t} \max_{y_t} (\hat{y}_t - y_t)^2 - \Delta^*_t + W_t(\boldsymbol{s}_{t-1} + y_t \boldsymbol{x}_t) \right).$$

It is easy to check that $\boldsymbol{W}_0$ is the minimax regret for this game and that it equals the regret of the fixed design game when the adversary plays every round.

Calculating Δ^*_t

To proceed, we need a closed for solution for Δ^*_t. To this end, and using $\mathcal{R}(\boldsymbol{M})$ to denote the row space of matrix $\boldsymbol{M}$, we can prove:

Lemma 41. *The marginal loss for the comparator of playing another round with covariate $\boldsymbol{x} = \boldsymbol{x}_\parallel + \boldsymbol{x}_\perp$, where $\boldsymbol{x}_\parallel \in \mathcal{R}(\boldsymbol{\Pi}_{t-1})$ and $\boldsymbol{x}_\perp$ is its orthogonal complement, is*

$$\Delta^*_t = y_t^2 \left(1 - \boldsymbol{x}_t^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t\right) - 2 y_t \boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t + \left(\boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t\right)^2 \frac{\boldsymbol{x}_t^\intercal \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t}{\boldsymbol{x}_t^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t}.$$

The rest of this section proves this result but can be skipped without impacting the rest of the presentation.

While the update of $\boldsymbol{P}_t$ is given by the forward recursion, the rank one update of $\boldsymbol{\Pi}_t$ is more complicated; Sherman-Morrison cannot be used directly. Instead, we show the following lemma.

Lemma 42. *Using $\boldsymbol{x}_\perp \coloneqq \boldsymbol{x} - \boldsymbol{\Pi}_{t-1}\boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}$ to denote the projection of $\boldsymbol{x}_t$ onto the orthogonal complement of $\boldsymbol{\Pi}_{t-1}$, we have*

$$\boldsymbol{\Pi}_t^\dagger = \begin{cases} \boldsymbol{\Pi}_{t-1}^\dagger - \dfrac{\boldsymbol{x}_\perp \boldsymbol{x}^\intercal \boldsymbol{\Pi}_{t-1}^\dagger + \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}\boldsymbol{x}_\perp^\intercal}{\boldsymbol{x}_\perp^\intercal \boldsymbol{x}_\perp} + \dfrac{\boldsymbol{x}_\perp \left(1 + \boldsymbol{x}^\intercal \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}\right) \boldsymbol{x}_\perp^\intercal}{(\boldsymbol{x}_\perp^\intercal \boldsymbol{x}_\perp)^2} & \textit{if } \boldsymbol{x} \notin \mathcal{R}(\boldsymbol{\Pi}_{t-1}), \textit{ and} \\[2ex] \boldsymbol{\Pi}_{t-1}^\dagger + \dfrac{\boldsymbol{\Pi}_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal \boldsymbol{\Pi}_{t-1}}{1 - \boldsymbol{x}_t^\intercal \boldsymbol{\Pi}_{t-1} \boldsymbol{x}_t^\intercal} & \textit{otherwise .} \end{cases}$$

Proof. We will write $\boldsymbol{X}$ as the matrix with columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}$. Thus, we have

$$\boldsymbol{\Pi}_t = \boldsymbol{\Pi}_{t-1} + \boldsymbol{x}\boldsymbol{x}^\intercal = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{x} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^\intercal \\ \boldsymbol{x}^\intercal \end{bmatrix},$$

and since $\boldsymbol{X}$ has linearly independent columns, (without loss of generality; we shall see why later), $\begin{bmatrix}\boldsymbol{X} & \boldsymbol{x}\end{bmatrix}$ has linearly independent columns since x is not in the column space of $\boldsymbol{X}$. Therefore, we have

$$\begin{bmatrix}\boldsymbol{X} & \boldsymbol{x}\end{bmatrix}^{\dagger} = \left(\begin{bmatrix}\boldsymbol{X}^{\top} \\ \boldsymbol{x}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{X} & \boldsymbol{x}\end{bmatrix}\right)^{-1}\begin{bmatrix}\boldsymbol{X}^{\top} \\ \boldsymbol{x}^{\top}\end{bmatrix}$$

and

$$\boldsymbol{\Pi}_t^{\dagger} = (\boldsymbol{\Pi}_{t-1} + \boldsymbol{x}\boldsymbol{x}^{\top})^{\dagger} = \begin{bmatrix}\boldsymbol{X} & \boldsymbol{x}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{X}^{\top} \\ \boldsymbol{x}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{X} & \boldsymbol{x}\end{bmatrix}\right)^{-2}\begin{bmatrix}\boldsymbol{X}^{\top} \\ \boldsymbol{x}^{\top}\end{bmatrix}.$$

Now, recall that the matrix that projects onto the column space of $\boldsymbol{X}$ is $\mathcal{P} := \boldsymbol{X}\boldsymbol{X}^{\dagger}$ and define $\boldsymbol{x}_{\parallel} := \mathcal{P}\boldsymbol{x}$ and $\boldsymbol{x}_{\perp} = \boldsymbol{x} - \boldsymbol{x}_{\parallel}$. We can calculate the middle matrix by using the block matrix inversion formula:

$$\begin{aligned}
\left(\begin{bmatrix}\boldsymbol{X}^{\top} \\ \boldsymbol{x}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{X} & \boldsymbol{x}\end{bmatrix}\right)^{-1} &= \begin{bmatrix}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1} + \frac{\boldsymbol{X}^{\dagger}\boldsymbol{x}\boldsymbol{x}^{\top}\boldsymbol{X}^{\top\dagger}}{\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}^{\top}\mathcal{P}\boldsymbol{x}} & \frac{-\boldsymbol{X}^{\dagger}\boldsymbol{x}}{\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}^{\top}\mathcal{P}\boldsymbol{x}} \\ \frac{-\boldsymbol{x}^{\top}\boldsymbol{X}^{\top\dagger}}{\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}^{\top}\mathcal{P}\boldsymbol{x}} & \frac{1}{\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}^{\top}\mathcal{P}\boldsymbol{x}}\end{bmatrix} \\
&= \frac{1}{\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}_{\parallel}^{\top}\boldsymbol{x}_{\parallel}}\begin{bmatrix}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}_{\parallel}^{\top}\boldsymbol{x}_{\parallel}\right) + \boldsymbol{X}^{\dagger}\boldsymbol{x}\boldsymbol{x}^{\top}\boldsymbol{X}^{\top\dagger} & -\boldsymbol{X}^{\dagger}\boldsymbol{x} \\ -\boldsymbol{x}^{\top}\boldsymbol{X}^{\top\dagger} & 1\end{bmatrix},
\end{aligned}$$

and so

$$\left(\begin{bmatrix}\boldsymbol{X}^{\top} \\ \boldsymbol{x}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{X} & \boldsymbol{x}\end{bmatrix}\right)^{-1}\begin{bmatrix}\boldsymbol{X}^{\top} \\ \boldsymbol{x}^{\top}\end{bmatrix} = \frac{1}{\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}_{\parallel}^{\top}\boldsymbol{x}_{\parallel}}\begin{bmatrix}\boldsymbol{X}^{\dagger}\left(\boldsymbol{x}^{\top}\boldsymbol{x} - \boldsymbol{x}_{\parallel}^{\top}\boldsymbol{x}_{\parallel}\right) - \boldsymbol{X}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\top} \\ \boldsymbol{x}_{\perp}^{\top}\end{bmatrix}$$

Using the Pythagorean theorem (i.e. that $\boldsymbol{x}^{\top}\boldsymbol{x} = \boldsymbol{x}_{\parallel}^{\top}\boldsymbol{x}_{\parallel} + \boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}$) and that $\boldsymbol{\Pi}_{t-1}^{\dagger} = \boldsymbol{X}^{\top\dagger}\boldsymbol{X}^{\dagger}$, we have

$$\begin{aligned}
\boldsymbol{\Pi}_t^{\dagger} &= \frac{1}{\left(\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}\right)^2}\begin{bmatrix}\boldsymbol{X}^{\dagger\top}\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp} - \boldsymbol{x}_{\perp}\boldsymbol{x}^{\top}\boldsymbol{X}^{\top\dagger} & \boldsymbol{x}_{\perp}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}^{\dagger}\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp} - \boldsymbol{X}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\top} \\ \boldsymbol{x}_{\perp}^{\top}\end{bmatrix} \\
&= \frac{1}{\left(\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}\right)^2}\left(\boldsymbol{\Pi}_{t-1}^{\dagger}\left(\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}\right)^2 - \boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}\left(\boldsymbol{x}_{\perp}\boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger} + \boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\top}\right)\right) \\
&\quad + \frac{\boldsymbol{x}_{\perp}\boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\top}}{\left(\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}\right)^2} + \frac{\boldsymbol{x}_{\perp}\boldsymbol{x}_{\perp}^{\top}}{\left(\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}\right)^2} \\
&= \boldsymbol{\Pi}_{t-1}^{\dagger} - \frac{\boldsymbol{x}_{\perp}\boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger} + \boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\top}}{\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}} + \frac{\boldsymbol{x}_{\perp}\left(1 + \boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\right)\boldsymbol{x}_{\perp}^{\top}}{(\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp})^2}.
\end{aligned}$$

Thus, we can evaluate

$$\begin{aligned}
\boldsymbol{x}^{\top}\boldsymbol{\Pi}_t^{\dagger}\boldsymbol{x} &= \boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x} - \frac{\boldsymbol{x}^{\top}\boldsymbol{x}_{\perp}\boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x} + \boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}}{\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp}} + \frac{\boldsymbol{x}^{\top}\boldsymbol{x}_{\perp}\left(1 + \boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\right)\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}}{(\boldsymbol{x}_{\perp}^{\top}\boldsymbol{x}_{\perp})^2} \\
&= \boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x} - 2\boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x} + 1 + \boldsymbol{x}^{\top}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x} \\
&= 1,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{x}^{\intercal}\boldsymbol{\Pi}_t^{\dagger}\boldsymbol{s} &= \boldsymbol{x}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{s} - \boldsymbol{x}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{s} - \frac{\boldsymbol{x}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{s}}{\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{x}_{\perp}} + \frac{\left(1+\boldsymbol{x}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\right)\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{s}}{\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{x}_{\perp}} \\
&= -\frac{\boldsymbol{x}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{s}}{\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{x}_{\perp}} + \frac{\left(1+\boldsymbol{x}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}\right)\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{s}}{\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{x}_{\perp}} \\
&= 0.
\end{aligned}$$

Finally, notice that

$$\boldsymbol{s}^{\intercal}\boldsymbol{\Pi}_t^{\dagger}\boldsymbol{s} = \boldsymbol{s}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{s}$$

since $\boldsymbol{x}_{\perp}^{\intercal}\boldsymbol{s} = 0$.

The second case is a consequence of the Sherman-Morrison formula. Since $\boldsymbol{\Pi}_t$, $\boldsymbol{\Pi}_{t-1}$, and $\boldsymbol{x}_t$ are all in the same eigenspace, we can without loss of generality assume full rank and apply Sherman-Morrison. A precise formulation can also be found in e.g. [24]. □

With this form of $\boldsymbol{\Pi}_t^{\dagger}$, we can easily check the following equalities.

Corollary 43. *For a PSD symmetric matrix $\boldsymbol{\Pi}$, $s \in \mathcal{R}(\boldsymbol{\Pi})$ and $x \notin \mathcal{R}(\boldsymbol{\Pi})$, we have*

$$\begin{aligned}
x^{\intercal}\left(\boldsymbol{\Pi} + xx^{\intercal}\right)^{\dagger} x &= 1, \\
s^{\intercal}\left(\boldsymbol{\Pi} + xx^{\intercal}\right)^{\dagger} x &= 0, \\
s^{\intercal}\left(\boldsymbol{\Pi} + xx^{\intercal}\right)^{\dagger} s &= s^{\intercal}\boldsymbol{\Pi}^{\dagger}s.
\end{aligned}$$

Proof of Lemma 41. We have

$$\Delta_t^* = \sigma_t^2 - \sigma_{t-1}^2 - \left(\boldsymbol{s}_{t-1} + y_t\boldsymbol{x}_t\right)^{\intercal}\boldsymbol{\Pi}_t^{\dagger}\left(\boldsymbol{s}_{t-1} + y_t\boldsymbol{x}_t\right) + \boldsymbol{s}_{t-1}^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{s}_{t-1}.$$

First, assume that $\boldsymbol{x}_{\perp} = 0$. Then $\boldsymbol{x}_t$ is in the column space of $\boldsymbol{\Pi}_t$ and $\boldsymbol{\Pi}_{t-1}$, and an application of the generalized Sherman-Morrison formula (see e.g. [24]) yields

$$\boldsymbol{\Pi}_{t-1}^{\dagger} = \left(\boldsymbol{\Pi}_t - \boldsymbol{x}_t\boldsymbol{x}_t^{\intercal}\right)^{\dagger} = \boldsymbol{\Pi}_t^{\dagger} + \frac{\boldsymbol{\Pi}_t^{\dagger}\boldsymbol{x}_t\boldsymbol{x}_t^{\intercal}\boldsymbol{\Pi}_t^{\dagger}}{1 - \boldsymbol{x}_t^{\intercal}\boldsymbol{\Pi}_t^{\dagger}\boldsymbol{x}_t}, \tag{3.19}$$

and so

$$\begin{aligned}
\Delta_t^* &= \sigma_t^2 - \sigma_{t-1}^2 - \left(\boldsymbol{s}_{t-1} + y_t\boldsymbol{x}_t\right)^{\intercal}\Pi_t^{\dagger}\left(\boldsymbol{s}_{t-1} + y_t\boldsymbol{x}_t\right) + \boldsymbol{s}_{t-1}^{\intercal}\Pi_{t-1}^{\dagger}\boldsymbol{s}_{t-1} \\
&= y_t^2 - 2\boldsymbol{y}_t\boldsymbol{s}_{t-1}\Pi_t^{\dagger}\boldsymbol{x}_t - y_t^2\boldsymbol{x}_t^{\intercal}\Pi_t^{\dagger}\boldsymbol{x}_t + \boldsymbol{s}_{t-1}^{\intercal}\left(\Pi_{t-1}^{\dagger} - \Pi_t^{\dagger}\right)\boldsymbol{s}_{t-1}.
\end{aligned}$$

Finally, notice that (3.19) implies

$$\boldsymbol{x}_t^{\intercal}\boldsymbol{\Pi}_{t-1}^{\dagger}\boldsymbol{x}_t = \frac{\boldsymbol{x}_t^{\intercal}\boldsymbol{\Pi}_t^{\dagger}\boldsymbol{x}_t}{1 - \boldsymbol{x}_t^{\intercal}\boldsymbol{\Pi}_t^{\dagger}\boldsymbol{x}_t}$$

when $\boldsymbol{x}_\perp = 0$, yielding the claim in that case.

Now, assume that $\boldsymbol{x}_\perp \neq 0$. Then

$$\begin{aligned}&(\boldsymbol{s}_{t-1} + y_t\boldsymbol{x}_t)^\intercal \boldsymbol{\Pi}_t^\dagger(\boldsymbol{s}_{t-1} + y_t\boldsymbol{x}_t)\\ &= \boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{s}_{t-1} + 2y_t \boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t + y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t\\ &= \boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{s}_{t-1} + y_t^2,\end{aligned}$$

where we applied the three claims of Lemma 43 to obtain the second equality. Therefore, $\Delta_t^* = 0$, verifying the formula. □

Calculating the Value

We can now explicitly calculate the value-to-go for this game.

Theorem 44. *Consider the fixed-design game with early stopping, with covariates $\boldsymbol{x}_1^T$. Define the $\boldsymbol{P}_t$ by the backwards recursion* (3.7) *and define $\gamma_t = \sum_{s=t+1}^T B_s^2 \boldsymbol{x}_s^\intercal \boldsymbol{P}_s \boldsymbol{x}_s$. Suppose that, for all t,*

$$\gamma_t \geq \boldsymbol{s}_t^\intercal \left(\boldsymbol{\Pi}_t^\dagger - \boldsymbol{P}_t\right) \boldsymbol{s}_t. \tag{3.20}$$

Then the instantaneous value-to-go is equal to

$$W_t(\boldsymbol{s}_t) = \boldsymbol{s}_t^\intercal \left(\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger\right) \boldsymbol{s}_t + \gamma_t, \tag{3.21}$$

the adversary causes more regret by continuing the game, and the optimal player strategy is (**mms**).

Proof. The proof is by induction: assume that $\boldsymbol{W}_t(\boldsymbol{s}_t) = \boldsymbol{s}_t^\intercal \left(\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger\right) \boldsymbol{s}_t + \gamma_t$. The base case is easily established with $\gamma_T = 0$ and $\boldsymbol{P}_T = \boldsymbol{\Pi}_T^\dagger$ yielding the base case of $\boldsymbol{W}_T = 0$. Now, we assume

$\boldsymbol{W}_t$ is correct and want to verify the formula for $\boldsymbol{W}_{t-1}$. Hence, we need to calculate

$$
\begin{aligned}
&W_{t-1}(\boldsymbol{s}_{t-1}) \\
&= \max_{e_t \in \{0,1\}} e_t \left(\min_{\hat{y}_t} \max_{y_t} (\hat{y}_t - y_t)^2 - \Delta_t^* + W_t(\boldsymbol{s}_{t-1} + y_t \boldsymbol{x}_t) \right) \\
&= \Bigg(\min_{\hat{y}} \max_{y} \, (\hat{y} - y)^2 - y^2 \left(1 - \boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \right) + 2y \boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t - \left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \right)^2 \frac{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t}{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t} \\
&\qquad + (\boldsymbol{s}_{t-1} + y \boldsymbol{x}_t)^\mathsf{T} \left(\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger \right) (\boldsymbol{s}_{t-1} + y \boldsymbol{x}_t) + \gamma_t \Bigg)_+ \\
&= \Bigg(\min_{\hat{y}} \max_{y} \; \hat{y}^2 + 2y \left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t + \boldsymbol{s}_{t-1}^\mathsf{T} \left(\boldsymbol{P}_t + \boldsymbol{\Pi}_t^\dagger \right) \boldsymbol{x}_t - \hat{y} \right) + y^2 \boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \\
&\qquad - \left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \right)^2 \frac{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t}{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t} + \boldsymbol{s}_{t-1}^\mathsf{T} \left(\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger \right) \boldsymbol{s}_{t-1} + y^2 \boldsymbol{x}_t^\mathsf{T} \left(\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger \right) \boldsymbol{x}_t + \gamma_t \Bigg)_+ \\
&= \Bigg(\min_{\hat{y}} \max_{y} \hat{y}^2 + 2y \left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{P}_t \boldsymbol{x}_t - \hat{y} \right) + y^2 \boldsymbol{x}_t^\mathsf{T} \boldsymbol{P}_t \boldsymbol{x}_t \\
&\qquad - \left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \right)^2 \frac{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t}{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t} + \boldsymbol{s}_{t-1}^\mathsf{T} (\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger) \boldsymbol{s}_{t-1} + \gamma_t \Bigg)_+ .
\end{aligned}
$$

The objective is convex in y and therefore the optimum will be on the boundary at $\pm B_t$. Thus,

$$
\begin{aligned}
W_{t-1}(\boldsymbol{s}_{t-1}) = \Bigg(&\min_{\hat{y}} \hat{y}^2 + 2B_t \left| \boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{P}_t \boldsymbol{x}_t - \hat{y} \right| - B_t^2 \boldsymbol{x}_t^\mathsf{T} \boldsymbol{A}_t \boldsymbol{x}_t \\
&- \left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \right)^2 \frac{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t}{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t} + \boldsymbol{s}_{t-1}^\mathsf{T} (\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger) \boldsymbol{s}_{t-1} + \gamma_t \Bigg)_+ .
\end{aligned}
$$

This objective is convex in $\hat{y}$ as well, and hence we can minimize it by setting the subgradient to zero. Under the condition that $\left| \boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{B}_t \boldsymbol{x}_t \right| \leq B_t$, the subgradient at $\hat{y} = \boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{P}_t \boldsymbol{x}_t$ contains zero. Therefore,

$$
\begin{aligned}
&W_{t-1}(\boldsymbol{s}_{t-1}) \\
&= \left(\left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{P}_t \boldsymbol{x}_t \right)^2 + B_t^2 \boldsymbol{x}_t^\mathsf{T} \boldsymbol{P}_t \boldsymbol{x}_t - \left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \right)^2 \frac{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t}{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t} + \boldsymbol{s}_{t-1}^\mathsf{T} \left(\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger \right) \boldsymbol{s}_{t-1} + \gamma_t \right)_+ .
\end{aligned}
$$

If $\boldsymbol{x}_t \in \mathcal{R}(\boldsymbol{\Pi}_{t-1})$, then we can use a generalized Sherman-Morrison lemma (see Lemma 42 for details) to calculate $\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t = \frac{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t}{1 - \boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t}$, and therefore

$$
\begin{aligned}
\left(\boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \right)^2 \frac{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{x}_t}{\boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t} + \boldsymbol{s}_{t-1}^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{s}_{t-1} &= \boldsymbol{s}_{t-1}^\mathsf{T} \left(\boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \frac{1}{1 - \boldsymbol{x}_t^\mathsf{T} \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t} + \boldsymbol{\Pi}_t^\dagger \right) \boldsymbol{s}_{t-1} \\
&= \boldsymbol{s}_{t-1} \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{s}_{t-1} .
\end{aligned}
$$

If instead $\boldsymbol{x}_t \notin \mathcal{R}(\boldsymbol{\Pi}_{t-1})$, then a standard fact for the ordinary least squares solution is $\boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{x}_t = 0$ and $\boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_t^\dagger \boldsymbol{s}_{t-1} = \boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{s}_{t-1}$ (a proof of this fact is provided in Lemma 43). In either case, we have

$$\begin{aligned} W_{t-1}(\boldsymbol{s}_{t-1}) &= \left(\boldsymbol{s}_{t-1}^\intercal \left(\boldsymbol{P}_t + \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal \boldsymbol{P}_t\right) \boldsymbol{s}_{t-1} + B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t - \boldsymbol{s}_{t-1}^\intercal \boldsymbol{\Pi}_{t-1}^\dagger \boldsymbol{s}_{t-1} + \gamma_t\right)_+ \\ &= \left(\boldsymbol{s}_{t-1}^\intercal \left(\boldsymbol{P}_{t-1} - \boldsymbol{\Pi}_{t-1}^\dagger\right) \boldsymbol{s}_{t-1} + \gamma_{t-1}\right)_+ , \end{aligned}$$

verifying the $\boldsymbol{P}_t$ and γ_t recurrence. If $\gamma_{t-1} \geq \boldsymbol{s}_{t-1}^\intercal \left(\boldsymbol{\Pi}_{t-1}^\dagger - \boldsymbol{P}_{t-1}\right) \boldsymbol{s}_{t-1}$ holds for all t, then the instantaneous value-to-go is always positive, an optimal adversary will always continue, and the data sequence seen by the player is $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{P}_0)$. In this case, the minimax strategy is confirmed to be (**mms**) by Theorem 38. □

3.7 Minimax Game

From the previous section, we have the intuition that if the adversary plays actions that keep the instantaneous value-to-go positive, then the game will not terminate early. This section builds on this intuition to define a set of constraints where (**mms**) is the minimax response. For convenience, we restate constraints $\mathcal{A}$ and $\mathcal{B}$ and define a third constraint related to the continuation conditions (3.20). For some fixed B_t sequence, define the bounded label constraint on the y_ts, $\mathcal{L}(B_t) := \{y_t : |y_t| \leq B_t\}$, and the following constraints on the $\boldsymbol{x}_t$:

$$\begin{gathered} \mathcal{A}\left(\boldsymbol{\Sigma}_0\right) := \left\{\boldsymbol{x}_t : \boldsymbol{P}_t^\dagger \succeq \boldsymbol{\Pi}_t\right\}, \quad \mathcal{B}\left(\boldsymbol{\Sigma}_0\right) := \left\{\boldsymbol{x}_t : B_t \geq \sum_{s=1}^{t-1} |\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_s|\right\}, \text{ and} \\ \mathcal{C}\left(\boldsymbol{\Sigma}_0, \gamma_0\right) := \left\{\boldsymbol{x}_t : \gamma_{t-1} \geq \left\| \left(\boldsymbol{B}_{t-1} \boldsymbol{X}_{t-1}^\intercal \left(\boldsymbol{\Pi}_{t-1}^\dagger - \boldsymbol{P}_{t-1}\right) \boldsymbol{X}_{t-1} \boldsymbol{B}_{t-1}\right)^{\frac{1}{2}} \right\|_{\infty,2} \forall t = 1, \dots, T\right\}, \end{gathered}$$

where $\gamma_t = \gamma_{t-1} - B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$ and $\boldsymbol{P}_t$ are derived from the forward recursion (3.3), $\boldsymbol{X}_t$ is the matrix with columns $\boldsymbol{x}_1^t$, and $\boldsymbol{B}_t := \operatorname{diag}(B_1, \dots, B_t)$. The form of the $\mathcal{C}$ constraint is explained by writing

$$\begin{aligned} \boldsymbol{s}_{t-1}^\intercal \left(\boldsymbol{\Pi}_{t-1}^\dagger - \boldsymbol{P}_{t-1}\right) \boldsymbol{s}_{t-1} &\leq \max_{\xi : \|\xi\|_\infty \leq 1} \xi \boldsymbol{B}_{t-1} \boldsymbol{X}_{t-1}^\intercal \left(\boldsymbol{\Pi}_{t-1}^\dagger - \boldsymbol{P}_{t-1}\right) \boldsymbol{X}_{t-1} \boldsymbol{B}_{t-1} \xi \\ &= \left\| \left(\boldsymbol{B}_{t-1} \boldsymbol{X}_{t-1}^\intercal \left(\boldsymbol{\Pi}_{t-1}^\dagger - \boldsymbol{P}_{t-1}\right) \boldsymbol{X}_{t-1} \boldsymbol{B}_{t-1}\right)^{\frac{1}{2}} \right\|_{\infty,2} . \end{aligned}$$

The $\mathcal{C}$ constraint ensures that $\boldsymbol{s}_{t-1}^\intercal \left(\boldsymbol{\Pi}_{t-1}^\dagger - \boldsymbol{P}_{t-1}\right) \boldsymbol{s}_{t-1} \leq \gamma_{t-1}$ for all possible $y_1^t \in \mathcal{L}(B_t)$.

We use the shorthand $\mathcal{ABC}(\boldsymbol{\Sigma}_0, \gamma_0) := \mathcal{A}(\boldsymbol{\Sigma}_0) \cup \mathcal{B}(\boldsymbol{\Sigma}_0) \cup \mathcal{C}(\boldsymbol{\Sigma}_0, \gamma_0)$, and also define $\overline{\mathcal{ABC}}(\boldsymbol{\Sigma}_0, \gamma_0)$ to be any sequence $\boldsymbol{x}_1^T \in \overline{\mathcal{A}}(\boldsymbol{\Sigma}_0)$ and $\gamma_T = 0$, i.e., the $\boldsymbol{\Sigma}_0$ and γ_0 budgets are exactly met. Note that $\overline{\mathcal{ABC}}(\boldsymbol{\Sigma}_0, \gamma_0) \subset \mathcal{ABC}(\boldsymbol{\Sigma}_0, \gamma_0)$. As we shall show, the covariate sequence $\boldsymbol{x}_1^T$ cannot be optimal for the adversary unless $\boldsymbol{x}_1^T \in \overline{\mathcal{ABC}}(\boldsymbol{\Sigma}_0, \gamma_0)$.

Given: $\boldsymbol{\Sigma} \succeq 0$, $\gamma_0 > 0$.
For $t = 1, 2, \ldots,$

- Adversary may choose to end the game
- Adversary reveals $\boldsymbol{x}_t \in \mathcal{ABC}(\Sigma_0, \gamma_0)$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals y_t such that $|y_t| \leq B_t$
- Learner incurs loss $(\hat{y}_t - y_t)^2$
- The game ends if no $\boldsymbol{x}_{t+1}$ exists such that $\boldsymbol{x}_1^{t+1} \in \mathcal{ABC}(\boldsymbol{\Sigma}_0, \gamma_0)$

Figure 3.2: Adversarial Covariates Protocol

The adversarial covariates protocol is presented in Figure 3.7 and allows the adversary to choose covariates (subject to $\mathcal{ABC}$) and to choose when to end the game. The first step in the analysis is to check that the constraint set is non-trivial.

Lemma 45. *Consider the game defined by $\boldsymbol{\Sigma}_0 \succeq 0$, $\gamma_0 \geq \|B_t\|_\infty$ and a B_t sequence. If there exists some T such that*

$$\sum_{t=1}^{T} \frac{B_t^2}{t + \log(T+1)} \geq \gamma_0, \tag{3.22}$$

then there exists a covariate sequence $\boldsymbol{x}_1^T \in \overline{\mathcal{ABC}}(\boldsymbol{\Sigma}_0, \gamma_0)$. In particular, any B_t that are bounded below satisfy this condition.

Proof. It actually suffices to take the simplest of sequences, $\boldsymbol{x}_t = \boldsymbol{e}_1$. For any fixed T, $\boldsymbol{P}_T = \frac{1}{T}\boldsymbol{e}_1\boldsymbol{e}_1^\intercal$, where all the $\boldsymbol{P}_t$ for the remainder of the proof are with respect to the covariate sequence of T copies of $\boldsymbol{e}_1$. In this case, the $\boldsymbol{P}_t$ matrices are all zero except for the first element which evolves like $\boldsymbol{P}_{t-1} = \boldsymbol{P}_t + \boldsymbol{P}_t^2$. This is the same recursion studied by [50], who proved a lower bound of $(t + \log(T+1) - \log(t+1))^{-1}$. Thus, we can bound

$$\sum_{t=1}^{T} B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t \geq \sum_{t=1}^{T} \frac{B_t^2}{t + \log(T+1) - \log(t+1)} \geq \sum_{t=1}^{T} \frac{B_t^2}{t + \log(T+1)},$$

and thus condition (3.22) implies that there is an $\boldsymbol{x}_1^T$ sequence that produces an upper bound on γ_0.

Next, notice that if we choose any index t' with $B_{t'} \leq \|B_t\|_\infty$, then the covariate sequence $\boldsymbol{x}_t = e_1\{t = t'\}$, where $\{\cdot\}$ is the indicator function, produces $\sum_{t=1}^{T} B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t = B_{t'}^2 \leq \gamma_0$. Now, $\sum_{t=1}^{T} B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t \leq \gamma_0$ is a continuous function of $\boldsymbol{x}_1^T$, and hence, by the intermediate value theorem, there is a $\boldsymbol{x}_1^T$ with $\sum_{t=1}^{T} B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t = \gamma_0$.

Next, we check the $\mathcal{B}$ constraint. First, observe that it suffices to check that we can construct some $\boldsymbol{x}_1^T$ using the construction of the previous paragraph. On $[0, 1/2]$, $x/(1+x) \geq x/2$ and the $\boldsymbol{P}_t$

sequence is decreasing, so $\sum_{s=t+1}^{T} \boldsymbol{x}_s^2 \frac{\boldsymbol{x}_s^2 \boldsymbol{P}_s}{1+\boldsymbol{x}_s^2 \boldsymbol{P}_s} \geq \frac{1}{2}x \sum_{s=t+1}^{T} \boldsymbol{x}_s^4 \boldsymbol{P}_s$, and combined with (3.11), we have

$$\sum_{s=1}^{t} |\boldsymbol{x}_t^\intercal \boldsymbol{P}_s \boldsymbol{x}_s| \leq |\boldsymbol{x}_t| \frac{\sum_{s=1}^{t} |\boldsymbol{x}_s|}{\boldsymbol{\Pi}_t + \sum_{s=t+1}^{T} \boldsymbol{x}_s^2 \frac{\boldsymbol{x}_s^2 \boldsymbol{P}_s}{1+\boldsymbol{x}_s^2 \boldsymbol{P}_s}} \leq |\boldsymbol{x}_t| \frac{\sum_{s=1}^{t} |\boldsymbol{x}_s|}{\boldsymbol{\Pi}_t + \sum_{s=t+1}^{T} \boldsymbol{x}_s^2 \frac{\boldsymbol{x}_s^2 \boldsymbol{P}_s}{2}}.$$

The arguments from the previous section show that $\sum_{s=t+1}^{T} \boldsymbol{x}_s^2 \frac{\boldsymbol{x}_s^2 \boldsymbol{P}_s}{2}$ can be made to grow without bound (in particular, by taking $\boldsymbol{x}_s = \boldsymbol{e}_1$), and so we can always find a long enough covariate sequence such that the $\mathcal{B}$ constraint is met.

Now, fix any $\boldsymbol{x}_1^T$ sequence that achieves the $\mathcal{B}$ and $\mathcal{C}$ constraints. By Lemma 34, we can, for any invertible matrix $\boldsymbol{A}$, rescale the covariate sequence to form $\boldsymbol{x}_t' = \boldsymbol{A}\boldsymbol{x}_t$ to obtain the corresponding $\boldsymbol{P}_t' = \boldsymbol{W}^{-1}\boldsymbol{P}_t\boldsymbol{W}^{-1}$. Since we have $\boldsymbol{x}_s^\intercal \boldsymbol{P}_t \boldsymbol{x}_t = \boldsymbol{x}_s'^\intercal \boldsymbol{P}_t' \boldsymbol{x}_t'$ for any s and t, the $\mathcal{B}$ and $\mathcal{C}$ constraints hold automatically. Therefore, we are free to choose $\boldsymbol{A}$ such that $\boldsymbol{P}_0' = \boldsymbol{\Sigma}_0$, and therefore $\boldsymbol{x}_1^{T\prime} \in \mathcal{ABC}(\boldsymbol{\Sigma}_0, \gamma_0)$.

□

We now present the main result of this work:

Theorem 46. *Consider the two player game defined in Figure 3.7. The player strategy* (**mms**) *is minimax optimal against an adversary in the sense that*

$$\sup_T \left(\sup_{(\boldsymbol{x}_1^T, y_1^T) \in \mathcal{ABC}(\boldsymbol{\Sigma}_0, \gamma_0)} R_T(s^*, \boldsymbol{x}_1^T, y_1^T) - \min_s \sup_{(\boldsymbol{x}_1^T, y_1^T) \in \mathcal{ABC}(\boldsymbol{\Sigma}_0, \gamma_0)} R_T(s, \boldsymbol{x}_1^T, y_1^T) \right) = 0.$$

and it attains the minimax regret, γ_0.

Proof. We will actually show something stronger: the optimal adversary strategy for the game in Figure 3.7 plays an $\boldsymbol{x}_1^T$ sequence in $\overline{ABC}$ and causes exactly γ_0 regret against the optimal player strategy (**mms**).

Now, assume that the game stops before round $T+1$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ have been played. There are four possible scenarios depending on whether the $\boldsymbol{\Sigma}_0$ or γ_0 budgets exhausted.

If neither budget is exhausted, we apply Lemma 40 to conclude that there exists a covariate sequence $\boldsymbol{x}_{T+1}^{T+k}$ that uses up the $\boldsymbol{\Sigma}_0$ budget. The $\mathcal{C}(\boldsymbol{\Sigma}_0, \gamma_0)$ contraint guarantees that the adversary can cause more regret by playing these rounds. Hence, an adversary that exhausts neither budget is suboptimal.

Since $\boldsymbol{P}_t - \boldsymbol{\Pi}_t^\dagger \succeq 0$, we cannot exhaust the γ_0 before the $\boldsymbol{\Sigma}_0$ budget and still satisfy the $\mathcal{C}$ constraint.

If the $\boldsymbol{\Sigma}_0$ budget is exhausted, then $\boldsymbol{x}_1^T \in \mathcal{A}$ and hence the minimax regret is $\sum_{t=1}^T B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$ by Theorem 38. Since $\gamma_T = \gamma_0 - \sum_{t=1}^T B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$, the adversary strategy is suboptimal if $\gamma_T > 0$ since it is possible to cause γ_0 regret.

These arguments cover all four cases, we can conclude that the adversary can cause at most γ_0 regret and that any strategy that causes γ_0 regret must exhaust the $\boldsymbol{\Sigma}_0$ and γ_0 budgets. □

The Necessity of a γ_0 Bound

Requiring a γ_0 bound may seem artificial at first, especially since it translates directly into a bound on the regret. However, it is a reasonable constraint to impose, for several reasons. First, the regret can be infinite if only condition $\mathcal{A}$ is imposed; Section 3.9 argues that the regret may grow like $\log(T)$ (while the results are upper bounds, lower bounds of the kind used in Lemma 45 supply lower bounds of the same order for the covariate sequence of constant magnitude). Additionally, the restriction of the adversary can be mild, as the γ_0 budget can be increased and the game history will remain consistent with it. Finally, we emphasize that the player does not need to know γ_0 to play (**mms**).

3.8 Follow the Regularized Leader

The minimax strategy (**mms**) can be interpreted as playing follow-the-regularized-leader with a certain data-dependent regularizer. In particular, if we define

$$\boldsymbol{R}_0 \coloneqq \boldsymbol{P}_0^{-1}, \text{ and } \boldsymbol{R}_t \coloneqq \boldsymbol{R}_{t-1} + \frac{1}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal - \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}^\intercal, \tag{3.23}$$

then we can prove the following correspondence.

Lemma 47. *The minimax strategy* (**mms**) *is exactly follow-the-regularized-leader, predicting* $\hat{y}_t = \theta^\intercal \boldsymbol{x}_t$ *at round* t*, where* θ *is the solution to*

$$\min_\theta \sum_{s=1}^{t-1} (\theta^\intercal \boldsymbol{x}_s - y_s)^2 + \theta^\intercal \boldsymbol{R}_t \theta,$$

and the regularization matrices $\boldsymbol{R}_t$ *are given by* (3.23).

Proof. Since θ minimizes a convex unconstrained objective, we set the derivative to zero and obtain the solution $\theta^* = \left(\sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal + \boldsymbol{R}_t\right)^{-1} \boldsymbol{s}_{t-1}$. Thus, we need to verify that $\sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal + \boldsymbol{R}_t = \boldsymbol{P}_t^{-1}$ for all t. This also guarantees that $\boldsymbol{R}_t \succeq 0$.

This is true for $t = 0$ by definition of R_0. Now, proceeding by induction, assume that the statement holds for $t - 1$. Then,

$$\begin{aligned}
\sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal + \boldsymbol{R}_t &= \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal + \boldsymbol{R}_{t-1} + \frac{\boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} - \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}^\intercal \\
&= \sum_{s=1}^{t-2} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal + \boldsymbol{R}_{t-1} + \frac{\boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} \\
&= \boldsymbol{P}_{t-1}^{-1} + \frac{\boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} = \boldsymbol{P}_t^{-1},
\end{aligned}$$

where the last equality is Sherman-Morrison. □

We can also write the $\boldsymbol{R}_t$ recursion without referring to $\boldsymbol{P}_t$.

Lemma 48. *The definition of* $\boldsymbol{R}_t$ *in Equation* (3.23) *is equivalent to defining* $\boldsymbol{R}_0 = \boldsymbol{P}_0^{-1}$ *and*

$$\boldsymbol{R}_t = \boldsymbol{R}_{t-1} + \frac{2\boldsymbol{x}_t\boldsymbol{x}_t^\intercal}{\sqrt{1 + 4\boldsymbol{x}_t^\intercal\left(\boldsymbol{R}_{t-1} + \sum_{s=1}^{t-2}\boldsymbol{x}_s\boldsymbol{x}_s^\intercal\right)^{-1}\boldsymbol{x}_t} + 1} - \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^\intercal. \tag{3.24}$$

Proof. First, we can calculate

$$4b_t^2 = \left(\frac{2}{1-a_t} - 1\right)^2 - 1 = \left(\frac{1+a_t}{1-a_t}\right)^2 - 1 = \frac{4a_t}{(1-a_t)^2}, \tag{3.25}$$

which implies that $b_t^2 = \frac{a_t}{(1-a_t)^2}$. Using the forward recursion (3.7) of $\boldsymbol{P}_t$, we have

$$\boldsymbol{x}_t^\intercal\boldsymbol{P}_t\boldsymbol{x}_t = b_t^2 - a_t b_t^2 = \frac{a_t}{1-a_t},$$

and

$$\frac{1}{1+\boldsymbol{x}_t^\intercal\boldsymbol{P}_t\boldsymbol{x}_t} = 1 - a_t = \frac{2}{\sqrt{1+4b_t^2}+1},$$

which, when combined with $b_t^2 = \boldsymbol{x}_t^\intercal\left(\boldsymbol{R}_{t-1} + \sum_{s=1}^{t-2}\boldsymbol{x}_s\boldsymbol{x}_s^\intercal\right)^{-1}\boldsymbol{x}_t$, yields the desired statement. □

Last step minimax The last step minimax algorithm [4] plays $\hat{y}_t = \boldsymbol{\Pi}_t^\dagger\boldsymbol{s}_{t-1}$, so we can also view the minimax algorithm as last step minimax with a regularization of

$$R_t = \sum_{s=t+1}^{T}\frac{\boldsymbol{x}_s^\intercal\boldsymbol{P}_s\boldsymbol{x}_s}{1+\boldsymbol{x}_s^\intercal\boldsymbol{P}_s\boldsymbol{x}_s}\boldsymbol{x}_s\boldsymbol{x}_s^\intercal.$$

3.9 Regret Bound

The main result of this chapter is to show that if the adversary plays in $\mathcal{ABC}(\Sigma_0,\gamma_0)$, then (**mms**) is the minimax optimal strategy and the adversary can cause at most γ_0 regret by fixing any $\boldsymbol{x}_1^T$ with $\overline{\mathcal{ABC}}(\boldsymbol{\Sigma}_0,\gamma_0)$ and playing the maximin strategy for the fixed-design game.

The regret of the games analyzed in Sections 3.5 and 3.6 need a different bound. Specifically,

Theorem 49. *For any fixed* T *and* B_1^T*, we can bound the minimax regret of the box-constrained game by*

$$\sup_{\boldsymbol{x}_1^T\in\overline{\mathcal{A}}(\boldsymbol{\Sigma}_0)}\ \sup_{y_1^T\in\mathcal{L}(B_1^T)} R_T(s^*,\boldsymbol{x}_1^T,y_1^T) \le \frac{d\|B_1^T\|_\infty}{\|\boldsymbol{\Sigma}\|_2}\left(1 + 2\ln\left(1 + \frac{||\boldsymbol{\Sigma}||_2^2}{2\|B_1^T\|_\infty^2}||B_1^T||_2^2\right)\right).$$

The minimax analysis shows that the minimax regret is equal to $\sup_{\boldsymbol{x}_1^T \in \mathcal{A}(\boldsymbol{\Sigma})} \sum_t B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$, which we bound by defining the worst case regret function,

$$\varphi_t(\boldsymbol{\Sigma}, B_1^t) = \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t} \left\{ \sum_{s=1}^t B_s^2 \boldsymbol{x}_s^\intercal \boldsymbol{P}_s(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_s)\boldsymbol{x}_s : \boldsymbol{\Sigma} \succeq \boldsymbol{P}_t(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t) \sum_{s=1}^t \boldsymbol{x}_s \boldsymbol{x}_s^\intercal \right\}.$$

We drop the explicit dependence of $\boldsymbol{P}_t$ on $\boldsymbol{x}_1^T$ and reparameterize by $\boldsymbol{r}_t^2 = \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal$:

$$\varphi_t(\boldsymbol{\Sigma}, B_1^t) = \max_{\boldsymbol{r}_1,\ldots,\boldsymbol{r}_t} \left\{ \sum_{s=1}^t B_s^2 \operatorname{tr}(\boldsymbol{r}_t^2) : \boldsymbol{\Sigma} \succeq \boldsymbol{P}_t \sum_{s=1}^t \boldsymbol{P}_s^{-1} \boldsymbol{r}_s^2 \right\}.$$

Noting that $\boldsymbol{P}_{t-1}\boldsymbol{P}_t^{-1} = \boldsymbol{I} + \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal = \boldsymbol{I} + \boldsymbol{r}_t^2$, we can derive an induction for φ_t:

$$\begin{aligned}
\varphi_t(\boldsymbol{\Sigma}, B_1^t) &= \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t} B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t + \left\{ \sum_{s=1}^{t-1} B_s^2 \boldsymbol{x}_s^\intercal \boldsymbol{P}_s \boldsymbol{x}_s : \boldsymbol{\Sigma} - \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal \succeq \boldsymbol{P}_t \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal \right\} \\
&= \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t} B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t + \left\{ \sum_{s=1}^{t-1} B_s^2 \boldsymbol{x}_s^\intercal \boldsymbol{P}_s \boldsymbol{x}_s : (\boldsymbol{\Sigma} - \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal)\boldsymbol{P}_{t-1}\boldsymbol{P}_t^{-1} \succeq \boldsymbol{P}_{t-1} \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal \right\} \\
&= \max_{\boldsymbol{r}_t,\ldots,\boldsymbol{r}_t} B_t^2 \operatorname{tr}(\boldsymbol{r}_t^2) + \left\{ \sum_{s=1}^{t-1} B_s^2 \operatorname{tr}(\boldsymbol{r}_s^2) : (\boldsymbol{\Sigma} - \boldsymbol{r}_t^2)(\boldsymbol{I} + \boldsymbol{r}_s^2) \succeq \boldsymbol{P}_{t-1} \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal \right\} \\
&= \max_{\boldsymbol{r}_t} B_t^2 \operatorname{tr}(\boldsymbol{r}_t^2) + \varphi_{t-1}\left((\boldsymbol{\Sigma} - \boldsymbol{r}_t^2)(\boldsymbol{I} + \boldsymbol{r}_s^2), B_1^{t-1} \right).
\end{aligned}$$

As a first step, we will bound φ_t in one dimension where $\varphi_t(\Sigma, B_1^t) = \max_{r_t} B_t^2 r_t^2 + \varphi_{t-1}((\Sigma - r_t^2)(1 + r_t^2), B_1^{t-1})$. We have omitted the bolding to emphasize that we are in the scalar case.

Lemma 50. *For every T and every B_1^T with $||B_1^T||_\infty \le \Sigma$,*

$$\varphi_T(\Sigma, B_1^T) \le \min\left\{ -\ln(1-\Sigma), 1 + 2\log\left(1 + \frac{||B_1^T||_2^2}{2}\right) \right\}.$$

Proof. In fact, we will prove the slightly stronger statement: for any positive function $f(T)$ with $f(0) \ge 0$ and $B_{T+1}^2 e^{-f(T)/2} + f(T) \le f(T+1)$, we have

$$\varphi_T(\Sigma, B_1^T) \le \min\{-\ln(1-\Sigma), f(T)\}.$$

We prove this by induction on T. The base case is trivial. Assume that the induction hypothesis holds for T. Then,

$$\begin{aligned}
\varphi_{T+1}(\Sigma, B_1^T) &= \max_{r_{T+1}^2} B_{T+1}^2 r_{T+1}^2 + \varphi_T\left((\Sigma - r_t^2)(1 + r_t^2), B_1^T \right) \\
&= \max_{0 \le x \le \Sigma} B_{T+1}^2 \frac{\sqrt{(1+\Sigma)^2 - 4x} - (1-\Sigma)}{2} + \varphi_T(x, B_1^{T-1}) \\
&\le \max_{0 \le x \le \Sigma} B_{T+1}^2 \frac{\sqrt{(1+\Sigma)^2 - 4x} - (1-\Sigma)}{2} + \min\{-\ln(1-x), f(T)\}.
\end{aligned}$$

Define $\hat{x} = 1 - \exp(-f(T))$, which is where the minimum switches from the first to the second argument. To find the maximizing x, we will calculate when the derivative is positive:

$$\begin{aligned}
&\frac{-B_{T+1}^2}{\sqrt{(1+\Sigma)^2 - 4x}} + \frac{1}{1-x} \geq 0 \\
\Leftrightarrow &(1+\Sigma)^2 - 4x - B_{T+1}^4(1-x)^2 \geq 0 \\
\Leftrightarrow &(1+\Sigma)^2 - B_{T+1}^4(1+x)^2 + 4(B_{T+1}^4 - 1)x \geq 0, \qquad (3.26)
\end{aligned}$$

which is true for all $x \leq \Sigma$ and $B^4 \leq \Sigma$. In fact, B_{T+1}^4 may be bigger than Σ without violating the constraint, but in particular $B_t \leq \Sigma$ is enough.

The sign of the derivative changes at $\hat{x}$. If $\Sigma \leq \hat{x}$, then the maximum is at Σ and we have

$$\begin{aligned}
\varphi_{T+1}(\Sigma, B_1^T) \leq & B_{T+1}^2 \frac{\sqrt{(1+\Sigma)^2 - 4\Sigma} - (1-\Sigma)}{2} + \varphi_T(\Sigma) \\
= & \varphi_T(\Sigma).
\end{aligned}$$

Otherwise, if $\hat{x} \leq \Sigma$, the maximum is at $\hat{x}$ and we have

$$\begin{aligned}
\varphi_{T+1}(\Sigma, B_1^T) \leq & B_{T+1}^2 \frac{\sqrt{(1+\Sigma)^2 - 4\hat{x}} - (1-\Sigma)}{2} + f(T) \\
\leq & B_{T+1}^2 \sqrt{1-\hat{x}} + f(T) \\
= & B_{T+1}^2 \exp(-f(T)/2) + f(T)
\end{aligned}$$

where the second line was from using $\Sigma \leq 1$. This allows any $f(T)$ that satisfies

$$B_{T+1}^2 e^{-f(T)/2} + f(T) \leq f(T+1).$$

To check that $f(T) = 1 + 2\log(1 + 1/2 \sum_t B_t^2)$ indeed works, we calculate:

$$\begin{aligned}
f(T+1) - f(T) &= -2\log\left(\frac{2 + \sum_{t=1}^T B_t^2}{2 + \sum_{t=1}^{T+1} B_t^2}\right) \\
&= -2\log\left(1 - \frac{B_{T+1}^2}{2 + \sum_{t=1}^{T+1} B_t^2}\right) \\
&\geq \frac{B_{T+1}^2}{1 + \frac{1}{2}\sum_{t=1}^{T+1} B_t^2} \\
&\geq e^{-1/2} \frac{B_{T+1}^2}{1 + \frac{1}{2}\sum_{t=1}^{T+1} B_t^2} \\
&= B_{T+1}^2 e^{-f(T)/2}.
\end{aligned}$$

□

The general multidimensional case can be bounded by first relaying the assumption that $\boldsymbol{r}_t^2 = \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal$ to allow general matrices $\boldsymbol{R}_t$ (relaxing the rank one assumption), which only increases the value of the maximization. We can then apply the one-dimensional bound in every direction:

Lemma 51. *For any $\Sigma \geq 0$, $\psi_t(\Sigma \boldsymbol{I}, B_1^t) = \sum_{i=1}^d \varphi_t(\Sigma, B_1^t)$, where $\varphi_t(\Sigma)$ is the one-dimensional regret bound.*

Proof. The base case is trivial since both sides are zero. For the inductive hypothesis, assume that $\psi_{t-1}(\Sigma \boldsymbol{I}, B_1^{t-1}) = \sum_{i=1}^d \varphi_{t-1}(\Sigma, B_1^{t-1})$. Denoting the eigenvalues of $\boldsymbol{R}$ by $\lambda_1, \ldots, \lambda_d$, we have

$$\begin{aligned}
\psi_t(\Sigma \boldsymbol{I}, B_1^t) &= \max_{\boldsymbol{R}} B_t^2 \operatorname{tr}(\boldsymbol{R}) + \psi_{t-1}\left((\Sigma \boldsymbol{I} - \boldsymbol{R})(\boldsymbol{I} + \boldsymbol{R}), B_1^{t-1}\right) \\
&= \max_{\boldsymbol{R}} \left\{ \sum_{i=1}^d \lambda_i + \sum_{i=1}^d \varphi_{t-1}\left((1+\lambda_i)(\Sigma - \lambda_i), B_1^{t-1}\right) \right\} \;=\; \sum_{i=1}^d \varphi_t(\Sigma, B_1^t).
\end{aligned}$$

□

Proof. **(of Theorem 49)** Recall from Theorem 32 that for given T and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, the regret of the box constrained game is precisely $\sum_{t=1}^T B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$. Lemma 51 bounds $\sum_{t=1}^T B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$ by a quantity that does not depend on $\boldsymbol{x}_t$. To invoke Lemma 50, we need that $B_t \leq \max_i \lambda_i$ for all t, which is exactly $||B_1^T||_\infty \leq ||\boldsymbol{\Sigma}||_2$. Rescaling the B_t sequence (and hence the regret bound) gives the result. □

3.10 Conclusion

We have presented the minimax optimal strategy for linear regression in a number of scenarios where it is efficient, including fixed-design and adversarial design with certain covariate scale and budget constraints. We found that, as long as the adversary respects these constraints, the minimax algorithm remains optimal for any length of game without having knowledge of it. Furthermore, this strategy competes even against strategies that are allowed to know the length of the game. We have also provided an intuitive view of the algorithm as follow-the-regularized-leader with a specific data-dependent regularizer, which automatically adjusts to the scale of the data and to how much data budget remains.

Chapter 4

Regret Bounds for MDP Learning

4.1 Introduction

The Markov Decision Process planning problem is to find a good policy given complete knowledge of the transition dynamics and loss function. Much work has been done by the reinforcement learning community; the earliest approaches with convergence guarantees date back to value iteration [7], policy iteration [27], and other dynamic programming ideas. Another thread has been the linear programming formulation [36]. In general, the planning problem is well understood for state-spaces small enough to permit computation of the value function [8]. However, in large state space problems, both the dynamic programming and linear program approaches are computationally infeasible as complexity scales quadratically with the number of states.

A popular approach to large-scale problems is to search for the optimal value function within the linear span of a small number of features with the hope that the optimal value function will be well approximated and will lead to a near optimal policy. Two popular methods are Approximate Dynamic Programming (ADP) and Approximate Linear Programming (ALP). For a survey on theoretical results for ADP, see [9, 47], [8, Vol. 2, Chapter 6], and more recent papers [49, 48, 33, 34].

Our goal is to find an almost-optimal policy in some low dimensional space such that the complexity scales with the low dimensional space but is sublinear in the size of the state space. In contrast, all prior work on ALP either scales badly or requires access to samples from a distribution that depends on the optimal policy. To accomplish this, we will use randomized algorithms to optimize policies that are parameterized by linear functions in the dual LP. We provide performance bounds in the average loss and discounted loss cases. In particular, we introduce new proof techniques and tools for average cost and discounted cost MDP problems and use these techniques to derive a reduction to stochastic convex optimization with accompanying error bounds.

Markov Decision Process

Markov decision processes have become a popular approach to modeling an agent interacting with an environment, and, most notably, are the model assumed by reinforcement learning. The MDP is parameterized by:

1. a discrete state space $\mathcal{X} = \{1, 2, \ldots, X\}$,
2. a discrete action space $\mathcal{A} = \{1, 2, \ldots, A\}$,
3. transition dynamics $P : \mathcal{X} \times \mathcal{A} \to \triangle_{\mathcal{X}}$ that describes the distribution of the next states given a current state and action, and
4. loss function $\ell : \mathcal{X} \times \mathcal{A} \to [0, 1]$ that provides the cost of taking an action in a given state.

The (fully observed) state encapsulates all the persistent information of the environment, and the influence of the agent is captured through the transition distribution, which is a function of the current state and the current action. We will use x_t to denote the state at time t and a_t the action chosen at time t. The state evolves in a Markov fashion: x_t is conditionally independent of the past given x_{t-1} and a_{t-1}. A policy π provides a distribution of actions for every state, and the agent's goal is to find a policy with small loss. The inclusion of the state dynamics forces the agent to consider policies that seek long term over myopic rewards.

The goal of the planning problem is to find a policy with small long-term loss; this requires the learner to plan for the entire trajectory, as myopic actions may guide the state to areas of high loss. We study the two most common costs: average cost and discounted. The first measures the the average loss under the stationary distribution under the policy, and the second measures the loss starting from a starting state with an exponential decay. In this way, average cost captures the long-term dynamics of a policy, and discounted cost captures the transient loss of a particular starting condition.

For example, the average cost of π is the expected loss of the stationary distribution of P^π and hence represents the average loss of a Markov chain after all the transient dynamics have vanished. On the other hand, discounted cost weights the future loss by an exponentially decreasing amount and is designed to capture the transient behavior. We will provide technical definitions in the next section, as we obtain results for both costs.

Notation

We will use the standard matrix notation for Markov chains. Throughout, we will use x_i for the ith component of vector x and M_{ij} for the element of $\boldsymbol{M}$ in row i, column j, and $M_{i,:}$ and $M_{:,j}$ for ith row and jth column of matrix M, respectively. Thus, we will write a distribution over states as a row vector $x \in \mathbb{R}^X$, where $p(X = i) = x_i$, and we will write the transition dynamics from state X to state X' as a matrix $\boldsymbol{P} \in \mathbb{R}^{X \times X}$ with $p(X' = j | X = i) = P_{ij}$ so that the product xP will be the marginal distribution of X'.

Given a fixed policy π, the induced state transition matrix is P^π where

$$(P^\pi)_{ij} = \sum_k P(X' = j | X = i, A = k)\pi(A = k | X = i).$$

We will also study distributions over state-action pairs, usually denoted $\mu \in \mathbb{R}^{X \times A}$, that can be thought of as a marginal on $\mathcal{X}$ multiplied by a policy, i.e. $\mu_{x,a} = P(X = x)\pi(A = x | X = x)$. This implies that μP provides the distribution of X' with the X marginal and policy π above. We also define the marginalization matrix $B \in \{0, 1\}^{XA \times X}$ to be the binary matrix such that the ith column has A ones in rows $1 + (i - 1)A$ to iA; thus, μB is the marginal distribution of the states of μ. A distribution over state-action pairs is a stationary distribution of P, in the sense that the state marginal μB is the stationary distribution under the corresponding P^π, if $\mu P = B\mu$.

We will use the norms $\|v\|_{1,c} = \sum_i c_i |v_i|$ and $\|v\|_{\infty,c} = \max_i c_i |v_i|$ (for a positive vector c). The constant one and zero vector are $\mathbf{1}$ and $\mathbf{0}$, and $\wedge$ and $\vee$ refer to the element-wise minimum and maximum. We can then compactly define $[v]_- = v \wedge 0$ and $[v]_+ = v \vee 0$ as the negative and positive parts of a vector v, respectively. Finally, $v \leq w$ for two vectors means element-wise inequality, i.e. $v_i \leq w_i$ for all i.

Evaluating a Policy

The goal in MDP planning is to find an optimal policy, but there are many different ways to assign a numeric value to the performance of a policy. The two most common are average cost and discounted cost. Average cost is roughly the expected loss of the policy once the Markov chain has reached stationarity and therefore disregards all the transient dynamics. Discounted cost minimizes the cost where future losses t rounds into the future are discounted by γ^t, where $\gamma \in (0,1)$ is some discounting factor. Therefore, discounted cost emphasized the short-term reward and roughly only considers $1/(1-\gamma)$ rounds into the future. Precisely,

$$\lambda_\pi(x) := \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{t=0}^{n} \ell(X_t, \pi(X_t)) \,\middle|\, X_0 = x\right] \qquad \text{(average cost), and} \qquad (4.1)$$

$$J_\pi(x) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(X_t, \pi(X_t)) \,\middle|\, X_0 = x\right] \qquad \text{(discounted cost)} \qquad (4.2)$$

where X_t is distributed according to $x_0(P^\pi)^t$. The initial state is very relevant for J but irrelevant for λ. We study the average cost in Section 4.2 and the discounted cost in Section 4.4.

Linear Programming for Average Cost

For the average cost, let $h \in \mathbb{R}^X$ be a vector and $\lambda\mathbb{R}$ a scalar. The *Bellman operator for average cost* is

$$\boldsymbol{L}h(x) := \min_{a\in\mathcal{A}} \left[\ell(x,a) + \sum_{x'\in\mathcal{X}} P_{(x,a),x'} h(x')\right],$$

and h and λ correspond to an optimal policy if they satisfy the Bellman optimality equation,

$$\lambda + h(x) = \boldsymbol{L}h(x) \quad \forall x.$$

We will call such an h and λ the differential value function and the average cost, respectively. When the Bellman optimtality equation is satisfied, the greedy policy (taking the action that achieves the minimum in the operator with probability 1) achieves the optimal loss, and finding a solution to the Bellman optimality condition implies that the greedy policy of h^* is optimal [41]. Additionally, every policy induces a stationary distribution over state-action pairs; the average cost is precisely the expected loss under this stationary distribution.

We can formulate the Bellman optimality equation as a linear program [36] by first noticing that if we have $\boldsymbol{L}h \geq h + \lambda\mathbf{1}$ for some λ and h, we must have $\lambda \leq \lambda^*$. Therefore, the optimal λ and h are the solution to

$$\begin{aligned} &\max_{\lambda,h} \lambda\,, \\ &\text{s.t.} \quad h + \lambda\mathbf{1} \leq \boldsymbol{L}h. \end{aligned}$$

Now, notice that $h+\lambda\mathbf{1} \le \min_a \left[\ell(x,a) + \sum_y P(y|x,a)h(y)\right]$ is equivalent to requiring $h(x)+\lambda \le \ell(x,a) + \sum_y P(y|x,a)h(y)$ for all x and a. In our matrix notation, this is precisely $B(\lambda\mathbf{1}+h) \le \ell + Ph$. Hence, the Bellman optimality equation is equivalent to the linear program

$$\begin{aligned} &\max_{\lambda,h} \lambda\,, \\ &\text{s.t.} \quad B(\lambda\mathbf{1}+h) \le \ell + Ph\,. \end{aligned} \tag{4.3}$$

A standard computation shows that the dual of LP (4.3) has the form of

$$\begin{aligned} &\min_{\mu\in\mathbb{R}^{XA}} \mu^\intercal \ell\,, \\ &\text{s.t.} \quad \mu^\intercal\mathbf{1} = 1,\ \mu \ge \mathbf{0},\ \mu^\intercal(P-B) = \mathbf{0}\ , \end{aligned} \tag{4.4}$$

The dual variable, μ, has an important interpretation: it is a stationary distribution over state-action pairs. The first two constraints ensure that μ is a probability distribution over state-action space and the third constraint forces μ to be a stationary distribution. To be precise, define the policy π_μ

$$\pi_\mu(a|x) = \frac{\mu(x,a)}{\sum_{a'\in\mathcal{A}} \mu(x,a')}.$$

Then, μ has a distribution where the state marginal is the stationary distribution of P^{π_μ} and the action is drawn according to $\pi_\mu(a|x)$. The objective, $\mu^\intercal\ell$, is the average loss under μ (i.e. the average loss of π_μ).

Linear Programming for Discounted Cost

There are analogous notions for the discounted cost setting. We define a value function $J : \mathcal{X} \to$ as a mapping from states to discounted costs. The hope is to find J^*, where $J^*(x)$ is the discounted cost starting in state x if the optimal policy is used.

We define the *Bellman operator for discounted cost*

$$\boldsymbol{L}^\gamma J(x) := \min_{a\in\mathcal{A}} \left[\ell(x,a) + \gamma \sum_{x'\in\mathcal{X}} P_{(x,a),x'} J(x')\right]$$

and the optimal value function will be the fixed point of the Bellman operator,

$$L^\gamma J^* = J^*.$$

It is easy to check that $J \le \boldsymbol{L}^\gamma J$ implies $J \le J^*$, and therefore, for any strictly positive vector α, the optimal value function is the solution to the linear program

$$\begin{aligned} &\max_{J} \alpha^\intercal J \\ &\text{s.t.} \quad \boldsymbol{L}^\gamma J \ge J. \end{aligned} \tag{4.5}$$

We also have an interpretable dual LP. Let $\alpha \in \mathbb{R}^X$ be an arbitrary positive vector such that $\alpha^\intercal \mathbf{1} = 1$. The linear program for discounted MDPs in the dual space has the form of

$$\begin{aligned} &\min_{\nu \in \mathbb{R}^{XA}} \nu^\intercal \ell \,, \\ &\text{s.t.} \quad (B - \gamma P)^\intercal \nu = \alpha, \quad \nu \geq 0, \quad \nu^\intercal \mathbf{1} = 1. \end{aligned} \tag{4.6}$$

Unlike the average cost case, the dual variable ν cannot be interpreted as a stationary distribution. However, it can be thought of as the discounted number of visits, as made explicit in the following theorem from [41]:

Theorem 52. *1. For each randomized Markovian policy π and state x and action a, define $\nu_\pi(x, a)$ by*

$$\nu_\pi(x, a) = \sum_{x'} \alpha(x') \sum_{t=1}^{\infty} \gamma^{t-1} P^\pi(x_t = x, a_t = a \mid x_1 = x') \,.$$

Then ν_π is a feasible solution to the dual problem.

2. Suppose ν is a feasible solution to the dual problem, then, for each $x \in \mathcal{X}$, $\sum_a \nu(x, a) > 0$. Define the randomized stationary policy π_ν by

$$\pi_\nu(a|x) = \frac{\nu(x, a)}{\sum_{a'} \nu(x, a')} \,.$$

Then, ν_{π_ν} is a feasible solution to the dual LP and $\nu_{\pi_\nu} = \nu$.

Thus, we can approximately solve the planning problem if we find a vector z such that the discounted cost of the policy defined by z, namely ν_{π_z}, is small. To handle possibly negative entries of $\boldsymbol{z}$, we more generally define

$$\pi_z(a|x) = \frac{[z(x, a)]_+}{\sum_{a'} [z(x, a')]_+} \,.$$

In this case, the precise relationship between ν_{π_z} and the value function can be found in [41]: for any vector z,

$$\sum_{x,a} \nu_{\pi_z}(x, a) = \frac{1}{1 - \gamma} \quad \text{and} \quad \nu_{\pi_z}^T \ell = \alpha^T J_{\pi_z}, \tag{4.7}$$

where J is the value function corresponding to policy π_z.

Approximate Linear Programming

If we ignore computational constraints, we can solve the planning problem by solving the linear programs (4.4) and (4.6). Unfortunately, state spaces are frequently very large and often grow exponentially with the complexity of the system (e.g. number of queues in the queuing network), and therefore any method polynomial in X becomes intractable. As any general optimality guarantee is impossible with computation sublinear in X without special knowledge of the problem, we instead aim for optimality with respect to some smaller policy class.

In contrast to previous work, we reduce the dimensionality by limiting the dual variables to lie in some d-dimensional affine subspace. Let $\Phi \in \mathbb{R}^{X \times A}$ by a feature matrix and μ_0 some know stationary distribution (that can be taken to be zero but allows a user to start with a good policy). For the average cost case, we will limit our search to $\mu = \mu_0 + \Phi\theta$ for $\theta \in \Theta \subset \mathbb{R}^d$; that is, we will study the *approximate average cost dual LP*,

$$\begin{aligned} &\min_{\theta} (\mu_0 + \Phi\theta)^\intercal \ell \,, \\ &\text{s.t.} \quad (\mu_0 + \Phi\theta)^\intercal \mathbf{1} = 1,\, \mu_0 + \Phi\theta \geq \mathbf{0},\, (\mu_0 + \Phi\theta)^\intercal (P - B) = \mathbf{0} \,. \end{aligned} \tag{4.8}$$

For every θ, we associate a policy

$$\pi_\theta(a|x) = \frac{[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_+}{\sum_{a'} [\mu_0(x,a') + \Phi_{(x,a'),:}\theta]_+} \tag{4.9}$$

and a stationary distribution μ_θ the actual stationary distribution of running policy π_θ. Thus, the average cost corresponding to the policy π_θ is $\ell^\intercal \mu_\theta$.

For the discounted cost case with feature matrix Φ, we restrict the dual variable to $\nu = \Phi\theta$ and define the *approximate discounted cost dual LP*

$$\begin{aligned} &\min_{\theta \in \mathbb{R}^d} \ \ell^\intercal \Phi\theta \,, \\ &\text{s.t.} \quad (B - \gamma P)^\intercal \Phi\theta = \alpha, \quad \Phi\theta \geq 0. \end{aligned} \tag{4.10}$$

For every θ, we define a policy

$$\pi_\theta(a|x) = \frac{[\Phi_{(x,a),:}\theta]_+}{\sum_{a'} [\Phi_{(x,a'),:}\theta]_+}, \tag{4.11}$$

and let ν_θ be corresponding dual variables (i.e. the discounted number of visits); hence, $\ell^\intercal \nu_\theta$ is the discounted cost as in (4.7).

Problem Definition

This chapter solves the following problem.

Definition 53 (Efficient Large-Scale Dual ALP). *For an MDP specified by ℓ and P with the dual variables ξ_θ corresponding to $\theta \in \Theta$, the efficient large-scale dual ALP problem is to find a $\widehat{\theta}$ such that*

$$\ell^\intercal \xi_{\widehat{\theta}} \leq \min \left\{ \ell^\intercal \xi_\theta : \xi_\theta \textit{ feasible for } (4.4) \textit{ or } (4.6) \right\} + O(\epsilon) \tag{4.12}$$

in time polynomial in d and $1/\epsilon$. The model of computation allows access to arbitrary entries of Φ, ℓ, P, μ_0, $P^\intercal \Phi$, and $\ell^\intercal \Phi$ in unit time.

The computational complexity cannot scale with X and we do not assume any knowledge of the optimal policy. In fact, as we shall see, we solve a harder problem, which we define as follows.

Definition 54 (Expanded Efficient Large-Scale Dual ALP). *Let $V : \mathbb{R}^d \to \mathbb{R}_+$ be some "violation function" that represents how far ξ_θ is from satisfying the constraints of* (4.4) *or* (4.6) *and has $V(\theta) = 0$ if θ is feasible.*

The expanded efficient large-scale dual ALP problem is to produce parameters $\widehat{\theta}$ such that

$$\ell^\intercal \xi_{\widehat{\theta}} \le \min_{\theta \in \Theta} \ell^\intercal \xi_\theta + V(\theta) + O(\epsilon), \tag{4.13}$$

in time polynomial in d and $1/\epsilon$, under the same model of computation as in Definition 53.

Note that the expanded problem is strictly more general as guarantee (4.13) implies guarantee (4.12). Also, many feature vectors Φ may not admit any feasible points. In this case, the dual ALP problem is trivial, but the expanded problem is still meaningful.

Having access to arbitrary entries of the quantities in Definition 53 arises naturally in many situations. In many cases, entries of $P^\intercal \Phi$ are easy to compute. For example, suppose that for any state x' there are a small number of state-action pairs (x, a) such that $P(x'|x, a) > 0$. Consider Tetris; although the number of board configurations is large, each state has a small number of possible neighbors. Dynamics specified by graphical models with small connectivity also satisfy this constraint. Computing entries of $P^\intercal \Phi$ is also feasible given reasonable features. If a feature φ_i is a stationary distribution, then $P^\intercal \varphi_i = B^\intercal \varphi_i$. Otherwise, it is our prerogative to design sparse feature vectors, hence making the multiplication easy. We shall see an example of this setting later.

Related Work

One of the first dimensionality reduction methods, proposed in Schweitzer and Seidmann [43], was to project the primal LP into a subspace. The first theoretical analysis of ALP methods, Farias and Van Roy [19], analyzed the discounted primal LP (4.6) performance when the value function was constrained; i.e. $J = \Phi\theta$. Given some vector $u \in \{u \in \mathbb{R}^X : u \ge \mathbf{1}, u \in \text{span}(\Psi), \beta_u < 1\}$ and a "goodness-of-fit" parameter $\beta_u = \gamma \max_{x,a} \sum_{x'} P_{(x,a),x'} u(x')/u(x)$, the authors showed that θ^*, the solution to the ALP, satisfies

$$\|J_* - \Psi w_*\|_{1,c} \le \frac{2c^\intercal u}{1 - \beta_u} \min_w \|J_* - \Psi w\|_{\infty,1/u} . \tag{4.14}$$

Unfortunately, this result has a number of limitations. First, solving ALP can be computationally expensive as the number of constraints is large. Second, it assumes that the feasible set of ALP is non-empty. Finally, Inequality (4.14) implies that $c = \mu_{\pi_{\Psi w_*},\nu}$ is an appropriate choice to obtain performance bounds. However, w_* itself is function of c and is not known before solving ALP.

One approach is to solve ALP iteratively, using $c = \mu_{\pi_{\Psi w_*},\nu}$ from last iteration. They showed that for an arbitrary probability distribution $\nu \in \Delta_{\mathcal{X}}$ and accompanying $\mu_{\pi,\nu} = (1-\gamma)\nu^\intercal (I - \gamma P^\pi)^{-1}$, we must have

$$\|J_{\pi_J} - J_*\|_{1,\nu} \le \frac{1}{1-\gamma} \|J - J_*\|_{1,\mu_{\pi_J,\nu}} .$$

Farias and Van Roy [18] propose a computationally efficient algorithm that is based on a constraint sampling technique. The idea is to sample a relatively small number of constraints and solve the resulting LP. Let $\mathcal{N} \subset \mathbb{R}^d$ be a known set that contains w_* (solution of ALP). Let $\mu^V_{\pi,c}(x) = \mu_{\pi,c}(x)V(x)/(\mu^\top_{\pi,c}V)$ and define the distribution $\rho^V_{\pi,c}(x,a) = \mu^V_{\pi,c}(x)/A$. Let $\delta \in (0,1)$ and $\epsilon \in (0,1)$. Let $\overline{\beta}_u = \gamma \max_x \sum_{x'} P_{(x,\pi_*(x)),x'} u(x')/u(x)$ and

$$D = \frac{(1+\overline{\beta}_V)\mu^\top_{\pi_*,c}V}{2c^\top J_*} \sup_{w\in\mathcal{N}} \|J_* - \Psi w\|_{\infty,1/V} , \qquad m \geq \frac{16AD}{(1-\gamma)\epsilon}\left(d \log \frac{48AD}{(1-\gamma)\epsilon} + \log\frac{2}{\delta}\right) .$$

Let $\mathcal{S}$ be a set of m random state-action pairs sampled under $\rho^V_{\pi_*,c}$. Let $\widehat{w}$ be a solution of the following sampled LP:

$$\begin{aligned} &\max_{w\in\mathbb{R}^d} c^\top \Psi w , \\ &\text{s.t.} \quad w \in \mathcal{N},\, \forall (x,a) \in \mathcal{S},\, \ell(x,a) + \gamma P_{(x,a),:}\Psi w \geq (\Psi w)(x) . \end{aligned}$$

Farias and Van Roy [18] prove that with probability at least $1-\delta$, we have

$$\|J_* - \Psi\widehat{w}\|_{1,c} \leq \|J_* - \Psi w_*\|_{1,c} + \epsilon \|J_*\|_{1,c} .$$

This result has a number of limitations. First, vector $\mu_{\pi_*,c}$ (that is used in the definition of D) depends on the optimal policy, but an optimal policy is what we want to compute in the first place. Second, we can no longer use Inequality (**??**) to obtain a performance bound (a bound on $\|J_{\pi_{\Psi\widehat{w}}} - J_*\|_{1,c}$), as $\Psi\widehat{w}$ does not necessarily satisfy all constraints of ALP.

Known as approximate linear programming (ALP), these methods were later improved by Farias and Van Roy [19, 17], Hauskrecht and Kveton [25], Guestrin, Hauskrecht, and Kveton [23], Petrik and Zilberstein [40], and Desai, Farias, and Moallemi [15]. As noted by Desai, Farias, and Moallemi [15], the prior work on ALP either requires access to samples from a distribution that depends on optimal policy or assumes the ability to solve an LP with as many constraints as states. Our objective is to design algorithms for very large MDPs that do not require knowledge of the optimal policy.

Let $c \in \mathbb{R}^X$ be a vector with positive components and $\gamma \in (0,1)$ be a discount factor. Let $L : \mathbb{R}^X \to \mathbb{R}^X$ be the Bellman operator defined by $(LJ)(x) = \min_{a\in\mathcal{A}}(\ell(x,a) + \gamma\sum_{x'\in\mathcal{X}} P_{(x,a),x'}J(x'))$ for $x \in \mathcal{X}$. Let $\Psi \in \mathbb{R}^{X\times d}$ be a feature matrix. The exact and approximate LP problems are as follows:

$$\begin{array}{ll} \displaystyle\max_{J\in\mathbb{R}^X} c^\top J , & \qquad \displaystyle\max_{w\in\mathbb{R}^d} c^\top \Psi w , \\ \text{s.t.} \quad LJ \geq J , & \qquad \text{s.t.} \quad L\Psi w \geq \Psi w . \end{array}$$

which can also be written as

$$\begin{array}{ll} \displaystyle\max_{J\in\mathbb{R}^X} c^\top J , & \displaystyle\max_{w\in\mathbb{R}^d} c^\top \Psi w , \\ \text{s.t.} \quad \forall (x,a),\, \ell(x,a) + \gamma P_{(x,a),:}J \geq J(x) , & \text{s.t.} \quad \forall (x,a),\, \ell(x,a) + \gamma P_{(x,a),:}\Psi w \geq (\Psi w)(x) . \end{array} \tag{4.15}$$

The optimization problem on the RHS is an approximate LP with the choice of $J = \Psi w$. Let $J_\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(x_t, \pi(x_t)) | x_0 = x\right]$ be the value of policy π, J_* be the solution of LHS, and $\pi_J(x) = \arg\min_{a\in\mathcal{A}}(\ell(x,a) + \gamma P_{(x,a),:} J)$ be the greedy policy with respect to J.

Desai, Farias, and Moallemi [15] study a smoothed version of ALP, in which slack variables are introduced that allow for some violation of the constraints. Let D' be a violation budget. The smoothed ALP (SALP) has the form of

$$\begin{aligned} &\max_{w,s} c^\top \Psi w\,, && \max_{w,s} c^\top \Psi w - \frac{2\mu_{\pi_*,c}^\top s}{1-\gamma}\,, \\ &\text{s.t.} \quad \Psi w \le L\Psi w + s,\ \mu_{\pi_*,c}^\top s \le D',\ s \ge \mathbf{0}, && \text{s.t.} \quad \Psi w \le L\Psi w + s,\ s \ge \mathbf{0}\,. \end{aligned}$$

The ALP on RHS is equivalent to LHS with a specific choice of D'. Let $\overline{U} = \{u \in \mathbb{R}^X \,:\, u \ge \mathbf{1}\}$ be a set of weight vectors. Desai, Farias, and Moallemi [15] prove that if w_* is a solution to above problem, then

$$\|J_* - \Psi w_*\|_{1,c} \le \inf_{w,u\in\overline{U}} \|J_* - \Psi w\|_{\infty,1/u} \left(c^\top u + \frac{2(\mu_{\pi_*,c}^\top u)(1+\beta_u)}{1-\gamma} \right)\,.$$

The above bound improves (4.14) as $\overline{U}$ is larger than U and RHS in the above bound is smaller than RHS of (4.14). Further, they prove that if η is a distribution and we choose $c = (1-\gamma)\eta^\top(I - \gamma P^{\pi_{\Psi w_*}})$, then

$$\|J_{\mu_{\Psi w_*}} - J_*\|_{1,\eta} \le \frac{1}{1-\gamma} \left(\inf_{w,u\in\overline{U}} \|J_* - \Psi w\|_{\infty,1/u} \left(c^\top u + \frac{2(\mu_{\pi_*,\nu}^\top u)(1+\beta_u)}{1-\gamma} \right) \right)\,.$$

Similar methods are also proposed by Petrik and Zilberstein [40]. One problem with this result is that c is defined in terms of w_*, which itself depends on c. Also, the smoothed ALP formulation uses π_* which is not known. Desai, Farias, and Moallemi [15] also propose a computationally efficient algorithm. Let $\mathcal{S}$ be a set of S random states drawn under distribution $\mu_{\pi_*,c}$. Let $\mathcal{N}' \subset \mathbb{R}^d$ be a known set that contains the solution of SALP. The algorithm solves the following LP:

$$\begin{aligned} &\max_{w,s} c^\top \Psi w - \frac{2}{(1-\gamma)S} \sum_{x\in\mathcal{S}} s(x)\,, \\ &\text{s.t.} \quad \forall x \in \mathcal{S},\ (\Psi w)(x) \le (L\Psi w)(x) + s(x),\ s \ge \mathbf{0},\ w \in \mathcal{N}'\,. \end{aligned}$$

Let $\widehat{w}$ be the solution of this problem. Desai, Farias, and Moallemi [15] prove high probability bounds on the approximation error $\|J_* - \Psi\widehat{w}\|_{1,c}$. However, it is no longer clear if a performance bound on $\|J_* - J_{\pi_{\Psi\widehat{w}}}\|_{1,c}$ can be obtained from this approximation bound.

Next, we turn our attention to average cost ALP. Let ν be a distribution over states, $u : \mathcal{X} \to [1, \infty)$, $\eta > 0$, $\gamma \in [0,1]$, $P_\gamma^\pi = \gamma P^\pi + (1-\gamma)\mathbf{1}\nu^\top$, and $L_\gamma h = \min_\pi(\ell_\pi + P_\gamma^\pi h)$. Farias and Van Roy [16] propose the following optimization problem:

$$\begin{aligned} &\min_{w,s_1,s_2} s_1 + \eta s_2\,, \\ &\text{s.t.} \quad L_\gamma \Psi w - \Psi w + s_1 \mathbf{1} + s_2 u \ge \mathbf{0},\ s_2 \ge 0\,. \end{aligned} \tag{4.16}$$

Let $(w_*, s_{1,*}, s_{2,*})$ be the solution of this problem. Define the mixing time of policy π by

$$\tau_\pi = \inf \left\{ \tau \ : \ |\frac{1}{t}\sum_{t'=0}^{t-1} \nu^\intercal (P^\pi)^{t'} \ell_\pi - \lambda_\pi| \le \frac{\tau}{t}, \forall t \right\} \ .$$

Let $\tau_* = \liminf_{\delta\to 0}\{\tau_\pi : \lambda_\pi \le \lambda_* + \delta\}$. Let π_γ^* be the optimal policy when discount factor is γ. Let $\pi_{\gamma,w}$ be the greedy policy with respect to Ψw when discount factor is γ, $\mu_{\gamma,\pi}^\intercal = (1-\gamma)\sum_{t=0}^\infty \gamma^t \nu^\intercal (P^\pi)^t$ and $\mu_{\gamma,w} = \mu_{\gamma,\pi_{\gamma,w}}$. Farias and Van Roy [16] prove that if $\eta \ge (2-\gamma)\mu_{\gamma,\pi_\gamma^*}^\intercal u$,

$$\lambda_{w_*} - \lambda_* \le \frac{(1+\beta)\eta \max(D'',1)}{1-\gamma} \min_w \|h_\gamma^* - \Psi w\|_{\infty,1/u} + (1-\gamma)(\tau_* + \tau_{\pi_{w_*}})\,,$$

where $\beta = \max_\pi \|I - \gamma P^\pi\|_{\infty,1/u}$, $D'' = \mu_{\gamma,w_*}^\intercal V/(\nu^\intercal V)$ and $V = L_\gamma \Psi w_* - \Psi w_* + s_{1,*}\mathbf{1} + s_{2,*}u$. Similar results are obtained more recently by Veatch [51].

An appropriate choice for vector ν is $\nu = \mu_{\gamma,w_*}$. Unfortunately, w_* depends on ν. We should also note that solving (4.16) can be computationally expensive. Farias and Van Roy [16] propose constraint sampling techniques similar to [18], but no performance bounds are provided.

Wang et al. [54] study ALP (4.8) and show that there is a dual form for standard value function based algorithms, including on-policy and off-policy updating and policy improvement. They also study the convergence of these methods, but no performance bounds are shown.

In the primal form (4.3), an extra constraint $h = \Psi w$ is added to obtain

$$\begin{aligned} &\max_{\lambda,w} \lambda\,, \\ &\text{s.t.} \quad B(\lambda e + \Psi w) \ge \ell + P\Psi w\,. \end{aligned} \tag{4.17}$$

Let λ_* be the average loss of the optimal policy and $(\widetilde{\lambda}, \widetilde{w})$ be the solution of this LP. It turns out that the greedy policy with respect to $\widetilde{w}$ can be arbitrarily bad even if $|\lambda_* - \widetilde{\lambda}|$ was small [17]. Farias and Van Roy [17] propose a two stage procedure, where the above LP is the first stage and the second stage is

$$\begin{aligned} &\max_w c^\intercal \Psi w\,, \\ &\text{s.t.} \quad B(\widetilde{\lambda} e + \Psi w) \le \ell + P\Psi w\,, \end{aligned} \tag{4.18}$$

where c is a user specified weight vector. Let $\widehat{w}$ be the solution of the second stage. Let λ_w and μ_w be the average loss and the stationary distribution of the greedy policy with respect to Ψw. Farias and Van Roy [17] prove that

$$\lambda_w - \lambda_* \le \|h_* - \Psi w\|_{1,\mu_w}\ .$$

Further, it is shown that $\widehat{w}$ minimizes $\|h_{\widetilde{\lambda}} - \Psi w\|_{1,c}$ and that

$$\|h_* - \Psi\widehat{w}\|_{1,c} \le \|h_{\widetilde{\lambda}} - \Psi\widehat{w}\|_{1,c} + (\lambda_* - \widetilde{\lambda})c^\intercal (I - P^{\pi_*})^{-1} e\,,$$

which implies that $\|h_* - \Psi\widehat{w}\|_{1,c}$ is small. To get that $\lambda_{\widehat{w}} - \lambda_*$ is small, we need to use $c = \mu_{\widehat{w}}$. Value of $\mu_{\widehat{w}}$ is obtained only after solving the optimization problem (4.18). To fix this problem, Farias and Van Roy [17] propose to solve (4.18) iteratively, using $c = \mu_{\widehat{w}}$ from the solution of the last round.

The above approach has two problems. First, it is still not clear if the average loss of the resultant policy is close to λ_* (or the best policy in the policy class). Second, iteratively solving (4.18) is computationally expensive. Similar results are also obtained by Desai, Farias, and Moallemi [15] who also show that if we were able to sample from the stationary distribution of the optimal policy, then LP (4.17) can be solved efficiently.

Our Contributions

We prove that if we parameterize the policy space by using the approximate dual LPs, then we can solve the expanded efficient large-scale dual ALP problem for discounted cost and average cost under a (standard) assumption that the distribution of states under any policy converges quickly to its stationary distribution. We also show that it suffices to solve the approximate dual LPs by approximately minimizing a surrogate loss function equal to the sum of the objective and a scaled violation function

We begin with the average cost in Section 4.2 and prove that, for some parameter $H > 0$, we have the regret bound

$$\mu_{\widehat{\theta}}^\intercal \ell \leq \min_\theta \mu_\theta^\intercal \ell + HV(\theta) + O\left(\frac{1}{H}\log(1/\delta)\right) + O(\epsilon)$$

where $V(\theta) = \|[\mu_0 + \Phi\theta]_-\|_1 + \|(P-B)^\intercal(\mu_0 + \Phi\theta)\|_1$. The $V(\theta)$ term is zero for feasible points (points in the intersection of the feasible set of LP (4.8) and the span of the features). For points outside the feasible set, these terms measure the extent of constraint violations for the vector $\mu_0 + \Phi\theta$, which indicate how well stationary distributions can be represented by the chosen features.

In particular, setting $H = \epsilon^{-1}$ gives an $O\left(\frac{1}{\epsilon}V(\theta) + \epsilon\right)$ regret bound between the $\ell^\intercal \mu_{\widehat{\theta}}$ returned by the algorithm and the best $\ell^\intercal \mu_\theta$. We emphasize that this bound is on the loss of actually running the π_θ policy, which could differ from the surrogate used in the optimization, $\ell^\intercal(\mu_0 + \Phi\theta)$.

This regret bound is rather unwieldy as H needs to be set correctly to obtain a $O(\epsilon)$ regret bound. Section 4.3 addresses this shortcoming with a meta algorithm that solves the surrogate optimization for a carefully chosen set of H values. We show a regret bound of

$$\ell^\intercal \mu_{\widehat{\theta}_T}^{\leq} \ell^\intercal \mu_\theta + O\left(\sqrt{V(\theta)}\right) + O(\epsilon).$$

We then turn to the discounted cost problem in Section 4.4. We obtain a bound on the discounted cost of the form

$$\ell^\intercal \nu_{\widehat{\theta}} \leq \ell^\intercal \nu_\theta + O\left(\left(\frac{1}{1-\gamma} + \frac{1}{\epsilon}\right)V(\theta)\right) + O\left(\frac{\epsilon}{1-\gamma}\right)$$

where $V(\theta)$ is an analogous violation function for the constraints in the discounted dual ALP. We also show that, even with the $\frac{1}{1-\gamma}$ term, we can use a grid meta-algorithm to obtain a regret bound of the form

$$\ell^\intercal \nu_{\hat{\theta}_T}^\intercal \ell \le \ell^\intercal \mu_\theta + O\left(\sqrt{V(\theta)}\right) + O(\epsilon).$$

This analysis is presented in Section 4.5.

Section 4.6 then demonstrates the effectiveness of our method on a well studied example from queuing theory, the Rybko-Stolyar queue. We show that using two simple heuristic policies with a small number of simple features provides good performance.

4.2 The Dual ALP for Average Cost

Is this section, we propose and analyze our solution to the Expanded large-scale MDP problem for average cost. As discussed in the introduction, there are two main challenges for solving the planning problem in its LP formulation: the optimization is in dimension X, and there are $O(XA)$ constraints, which is intractable in the large state-space setting.

We solve the two challenges by projecting the dual LP onto a subspace and by approximately solving the optimization using stochastic gradient descent, respectively. Unlike previous approaches for the primal LP, we show that an approximate solution in the dual allows for a regret, i.e. one that controls the error between our approximate solution and the best solution in some approximate policy class, and thereby solve Equation (4.13). We also provide some interpretation of the approximations we make.

Recall that, for a matrix Φ and a known stationary distribution μ_0 (which may be set to zero if no distribution is known), we defined the dual ALP

$$\begin{aligned} &\min_\theta \theta^\intercal \Phi^\intercal \ell\,, \\ &\text{s.t.} \quad \theta^\intercal \Phi^\intercal \mathbf{1} = 1,\ \Phi\theta \ge \mathbf{0},\ \theta^\intercal \Phi^\intercal (P - B) = \mathbf{0} \end{aligned}$$

and associated every θ with the policy

$$\pi_\theta(a|x) = \frac{[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_+}{\sum_{a'} [\mu_0(x,a') + \Phi_{(x,a'),:}\theta]_+}\,.$$

We denote the stationary distribution of this policy μ_θ which is only equal to $\mu_0 + \Phi\theta$ if θ is in the feasible set.

A Reduction to Stochastic Convex Optimization

Unfortunately, the ALP (4.8) still has $O(XA)$ constraints and cannot be solved exactly. Instead, we will form an unconstrained convex optimization that will act as a surrogate for the original problem and show that it is a finite sum, e.g. equal to $\sum_{i=1}^N f_i(\theta)$. Therefore, we can apply the extensive literature of solving finite sum problems with stochastic gradient descent methods.

To this end, for a constant $H \geq 1$, define the following convex cost function by adding a multiple of the total constraint violations to the objective of the LP (4.8):

$$\begin{aligned} c(\theta) &:= \ell^\intercal(\mu_0 + \Phi\theta) + H \left\| [\mu_0 + \Phi\theta]_- \right\|_1 + H \left\| (P-B)^\intercal(\mu_0 + \Phi\theta) \right\|_1 \\ &= \ell^\intercal(\mu_0 + \Phi\theta) + H \left\| [\mu_0 + \Phi\theta]_- \right\|_1 + H \left\| (P-B)^\intercal \Phi\theta \right\|_1 \\ &= \ell^\intercal(\mu_0 + \Phi\theta) + H \sum_{(x,a)} \left| [\mu_0(x,a) + \Phi_{(x,a),:}\theta]_- \right| + H \sum_{x'} \left| (P-B)^\intercal_{:,x'} \Phi\theta \right| . \end{aligned} \tag{4.19}$$

We justify using this surrogate function as follows. Suppose we find a near optimal vector $\widehat{\theta}$ such that $c(\widehat{\theta}) \leq \min_{\theta\in\Theta} c(\theta) + O(\epsilon)$. We will prove

1. that $\left\| [\mu_0 + \Phi\widehat{\theta}]_- \right\|_1$ and $\left\| (P-B)^\intercal(\mu_0 + \Phi\widehat{\theta}) \right\|_1$ are small and $\mu_0 + \Phi\widehat{\theta}$ is close to $\mu_{\widehat{\theta}}$ (Lemma 56), and
2. that $\ell^\intercal(\mu_0 + \Phi\widehat{\theta}) \leq \min_{\theta\in\Theta} c(\theta) + O(\epsilon)$.

As we will show, these two facts imply that with high probability, for any $\theta \in \Theta$,

$$\mu_{\widehat{\theta}}^\intercal \ell \leq \mu_\theta^\intercal \ell + \frac{1}{\epsilon} \left\| [\mu_0 + \Phi\theta]_- \right\|_1 + \frac{1}{\epsilon} \left\| (P-B)^\intercal(\mu_0 + \Phi\theta) \right\|_1 + O(\epsilon).$$

Unfortunately, calculating the gradients of $c(\theta)$ is $O(XA)$. Instead, we construct unbiased estimators and use stochastic gradient descent. Let T be the number of iterations of our algorithm, q_1 and q_2 be distributions over the state-action and state space, respectively (we will later discuss how to choose them), and $((x_t, a_t))_{t=1\ldots T}$ and $(x'_t)_{t=1\ldots T}$ be i.i.d. samples from these distsributions. At round t, the algorithm estimates subgradient $\nabla c(\theta)$ by

$$g_t(\theta) = \ell^\intercal \Phi - H \frac{\Phi_{(x_t,a_t),:}}{q_1(x_t,a_t)} \mathbb{I}_{\{\mu_0(x_t,a_t) + \Phi_{(x_t,a_t),:}\theta < 0\}} + H \frac{(P-B)^\intercal_{:,x'_t}\Phi}{q_2(x'_t)} s((P-B)^\intercal_{:,x'_t}\Phi\theta). \tag{4.20}$$

This estimate is fed to the projected subgradient method, which in turn generates a vector θ_t. After T rounds, we average vectors $(\theta_t)_{t=1\ldots T}$ and obtain the final solution $\widehat{\theta}_T = \sum_{t=1}^T \theta_t / T$. Vector $\mu_0 + \Phi\widehat{\theta}_T$ defines a policy, which in turn defines a stationary distribution $\mu_{\widehat{\theta}_T}$. The algorithm is shown in Figure 4.1.

Regret bound

We now turn towards proving the main result of this section, Theorem 55. We begin with our assumptions.

We make a mixing assumption on the MDP so that any policy quickly converges to its stationary distribution.

Assumption A1 ***(Fast Mixing)*** For any policy π, there exists a constant $\tau(\pi) > 0$ such that for all distributions d and d' over the state space, $\left\| dP^\pi - d'P^\pi \right\|_1 \leq e^{-1/\tau(\pi)} \left\| d - d' \right\|_1$.

```
Input: Constant $S > 0$, number of rounds $T$, constant $H$.
Let $\Pi_\Theta$ be the Euclidean projection onto $\Theta$.
Initialize $\theta_1 = 0$.
for $t := 1, 2, \dots, T$ do
    Sample $(x_t, a_t) \sim q_1$ and $x'_t \sim q_2$.
    Compute subgradient estimate $g_t$ (4.20).
    Update $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta_t g_t)$.
end for
$\widehat{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \theta_t$.
Return policy $\pi_{\widehat{\theta}_T}$.
```

Figure 4.1: The Stochastic Subgradient Method for Markov Decision Processes

Define

$$C_1 = \max_{(x,a) \in \mathcal{X} \times \mathcal{A}} \frac{\|\Phi_{(x,a),:}\|}{q_1(x,a)} , \qquad C_2 = \max_{x \in \mathcal{X}} \frac{\|(P - B)^\top_{:,x} \Phi\|}{q_2(x)} .$$

These constants appear in our performance bounds. So we would like to choose distributions q_1 and q_2 such that C_1 and C_2 are small. For example, if there is $C' > 0$ such that for any (x, a) and i, $\Phi_{(x,a),i} \le C'/(XA)$ and each column of P has only N non-zero elements, then we can simply choose q_1 and q_2 to be uniform distributions. Then it is easy to see that

$$\frac{\|\Phi_{(x,a),:}\|}{q_1(x,a)} \le C' , \qquad \frac{\|(P - B)^\top_{:,x} \Phi\|}{q_2(x)} \le C'(N + A) .$$

As another example, if $\Phi_{:,i}$ are exponential distributions and feature values at neighboring states are close to each other, then we can choose q_1 and q_2 to be appropriate exponential distributions so that $\|\Phi_{(x,a),:}\|/q_1(x,a)$ and $\|(P - B)^\top_{:,x}\Phi\|/q_2(x)$ are always bounded. Another example is when there exists a constant $C'' > 0$ such that for any x, $\|P^\top_{:,x}\Phi\|/\|B^\top_{:,x}\Phi\| < C''$ (this condition requires that columns of Φ are close to their *one step look-ahead*) and we have access to an efficient algorithm that computes $Z_1 = \sum_{(x,a)} \|\Phi_{(x,a),:}\|$ and $Z_2 = \sum_x \|B^\top_{:,x}\Phi\|$ and can sample from $q_1(x,a) = \|\Phi_{(x,a),:}\|/Z_1$ and $q_2(x) = \|B^\top_{:,x}\Phi\|/Z_2$. In what follows, we assume that such distributions q_1 and q_2 are known.

Obviously, minimizing the convex surrogate function does not guarantee a feasible solution to the original dual LP. Therefore, we define the following non-feasibility penalties which roughly correspond to how far $\Phi\theta$ is from the simplex and how far $\Phi\theta$ is from a stationary distribution, respectively:

$$V_1(\theta) := \sum_{(x,a)} |[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_-| \text{ and} \tag{4.21}$$

$$V_2(\theta) := \|(P - B)^\top(\Phi\theta)\|_1 = \sum_{x'} |(P - B)^\top_{:,x'}\Phi\theta| . \tag{4.22}$$

The main theorem of the section is:

Theorem 55. *Consider an expanded efficient large-scale dual ALP problem and some error tolerance $\epsilon > 0$ and desired maximum probability of error $\delta > 0$. Then running the stochastic subgradient method (shown in Figure 4.1) with*

$$H = \frac{1}{\epsilon}, \qquad T \geq \max\left\{\frac{H^2}{\epsilon^2}, 40S^2 \log\frac{1}{\delta}\right\}, \quad \textit{and} \quad \eta = \left(\sqrt{d} + H(C_1 + C_2)\right)\frac{S}{\sqrt{T}},$$

yields a $\widehat{\theta}_T$ where

$$\mu_{\widehat{\theta}_T}^{\intercal}\ell \leq \min_{\theta\in\Theta}\left(\mu_\theta^{\intercal}\ell + \frac{1}{\epsilon}(V_1(\theta) + V_2(\theta)) + O(\epsilon)\right) , \tag{4.23}$$

holds with probability at least $1 - \delta$. Constants hidden in the big-O notation are polynomials in S, d, C_1, C_2, $\log(1/\delta)$, $\log(V_1(\theta) + V_2(\theta))$, $\tau(\mu_\theta)$, and $\tau(\mu_{\widehat{\theta}_T})$.

Functions V_1 and V_2 are bounded by small constants for any set of normalized features: for any $\theta \in \Theta$,

$$\begin{aligned} V_1(\theta) &\leq \|\mu_0\|_1 + \|\Phi\theta\|_1 \leq 1 + \sum_{(x,a)} |\Phi_{(x,a),:}\theta| \leq 1 + Sd\,, \\ V_2(\theta) &\leq \sum_{x'} |P_{:,x'}^{\intercal}(\mu_0 + \Phi\theta)| + \sum_{x'} |B_{:,x'}^{\intercal}(\mu_0 + \Phi\theta)| \\ &\leq \left(\sum_{x'} P_{:,x'}\right)^{\intercal} [\mu_0 + \Phi\theta]_+ + \left(\sum_{x'} B_{:,x'}\right)^{\intercal} [\mu_0 + \Phi\theta]_+ \\ &= 2[\mu_0 + \Phi\theta]_+^{\intercal}\mathbf{1} \\ &\leq 2|\mu_0 + \Phi\theta|^{\intercal}\mathbf{1} \\ &= 2 + 2S\ . \end{aligned}$$

Thus V_1 and V_2 can be very small given a carefully designed set of features. The output $\widehat{\theta}_T$ is a random vector as the algorithm is based on a stochastic convex optimization method. The above theorem shows that with high probability the policy implied by this output is near optimal.

The optimal choice for ϵ is $\epsilon = \sqrt{V_1(\theta_*) + V_2(\theta_*)}$, where θ_* is the minimizer of RHS of (4.23) and not known in advance. One could think of parameterizing the optimization problem by H, but the problem is not jointly convex in H and θ. Nevertheless, we present methods that recover a $O(\sqrt{V_1(\theta_*) + V_2(\theta_*)})$ error bound using a grid based method in Section 4.3.

Analysis

This section provides the necessary technical tools and a proof of the main result. We break the proof into two main ingredients. First, we demonstrate that a good approximation to the surrogate loss gives a feature vector that is almost a stationary distribution; this is Lemma 56. Second, we justify the use of unbiased gradients in Theorem 57 and Lemma 59. The section concludes with the proof of Theorem 55.

The first ingredient shows that we can relate the magnitude of the constraint violation of θ to the difference between $\Phi\theta$ and μ_θ, which quantifies how far $\Phi\theta$ is from a stationary distribution.

Lemma 56. *Let $u \in \mathbb{R}^{XA}$ be a vector, $\mathcal{N}$ be the set of points (x, a) where $u(x, a) < 0$, and $\mathcal{S}$ be the complement of $\mathcal{N}$. Assume*

$$\sum_{x,a} u(x,a) = 1, \sum_{(x,a)\in\mathcal{N}} |u(x,a)| \le \epsilon', \left\|u^\intercal(P-B)\right\|_1 \le \epsilon''.$$

The vector $[u]_+/\|[u]_+\|_1$ defines a policy, which in turn defines a stationary distribution μ_u. We have that

$$\|\mu_u - u\|_1 \le \tau(\mu_u)\log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' \ .$$

Proof. Let $f = u^\intercal(P-B)$. From $\left\|u^\intercal(P-B)\right\|_1 \le \epsilon''$, we get that for any $x' \in \mathcal{X}$,

$$\sum_{(x,a)\in\mathcal{S}} u(x,a)(P-B)_{(x,a),x'} = -\sum_{(x,a)\in\mathcal{N}} u(x,a)(P-B)_{(x,a),x'} + f(x')$$

such that $\sum_{x'}|f(x')| \le \epsilon''$. Let $h = [u]_+/\|[u]_+\|_1$. Let $H' = \left\|h^\intercal(B-P)\right\|_1$. We write

$$\begin{aligned}
H' &= \sum_{x'}\Big|\sum_{(x,a)\in\mathcal{S}} h(x,a)(B-P)_{(x,a),x'}\Big| \\
&= \frac{1}{1+\epsilon'}\sum_{x'}\Big|\sum_{(x,a)\in\mathcal{S}} u(x,a)(B-P)_{(x,a),x'}\Big| \\
&= \frac{1}{1+\epsilon'}\sum_{x'}\Big|-\sum_{(x,a)\in\mathcal{N}} u(x,a)(B-P)_{(x,a),x'} + f(x')\Big| \\
&\le \frac{1}{1+\epsilon'}\left(\sum_{x'}\Big|-\sum_{(x,a)\in\mathcal{N}} u(x,a)(B-P)_{(x,a),x'}\Big| + \sum_{x'}|f(x')|\right) \\
&\le \frac{1}{1+\epsilon'}\left(\epsilon'' + \sum_{(x,a)\in\mathcal{N}}\sum_{x'}|u(x,a)||(B-P)_{(x,a),x'}|\right) \\
&\le \frac{1}{1+\epsilon'}\left(\epsilon'' + \sum_{(x,a)\in\mathcal{N}} 2|u(x,a)|\right) \le \frac{2\epsilon'+\epsilon''}{1+\epsilon'} \\
&\le 2\epsilon' + \epsilon'' \ .
\end{aligned}$$

Vector h is almost a stationary distribution in the sense that

$$\left\|h^\intercal(B-P)\right\|_1 \le 2\epsilon' + \epsilon'' \ . \tag{4.24}$$

Let $\|w\|_{1,\mathcal{S}} = \sum_{(x,a)\in\mathcal{S}} |w(x,a)|$. First, we have that

$$\|h-u\|_1 \le \|h - \frac{u}{1+\epsilon'}\|_1 + \|u - \frac{u}{1+\epsilon'}\|_{1,\mathcal{S}} \le 2\epsilon' \ .$$

Next we bound $\|\mu_h - h\|_1$. Using $\nu_0 = h$ as the initial state distribution, we will show that as we run policy h (equivalently, policy μ_h), the state distribution converges to μ_h and this vector is close to h. From (4.24), we have $\mu_0^\intercal P = h^\intercal B + v_0$, where v_0 is such that $\|v_0\|_1 \le 2\epsilon' + \epsilon''$. Let M^h be a $X \times (XA)$ matrix that encodes policy h, $M^h_{(i,(i-1)A+1)\text{-}(i,iA)} = h(\cdot|x_i)$. Other entries of this matrix are zero. We have

$$h^\intercal P M^h = (h^\intercal B + v_0)M^h = h^\intercal B M^h + v_0 M^h = h^\intercal + v_0 M^h \ ,$$

where we used the fact that $h^\intercal B M^h = h^\intercal$. Let $\mu_1^\intercal = h^\intercal P M^h$ which is the state-action distribution after running policy h for one step. Let $v_1 = v_0 M^h P = v_0 P^h$ and notice that as $\|v_0\|_1 \le 2\epsilon' + \epsilon''$, we also have that $\|v_1\|_1 = \|P^{h\intercal} v_0^\intercal\|_1 \le \|v_0\|_1 \le 2\epsilon' + \epsilon''$. Thus,

$$\mu_1^\intercal P = h^\intercal P + v_1 = h^\intercal B + v_0 + v_1 \ .$$

By repeating this argument for k rounds, we obtain

$$\mu_k^\intercal = h^\intercal + (v_0 + v_1 + \cdots + v_{k-1})M^h$$

and it is easy to see that

$$\|(v_0 + v_1 + \cdots + v_{k-1})M^h\|_1 \le \sum_{i=0}^{k-1} \|v_i\|_1 \le k(2\epsilon' + \epsilon'').$$

Thus, $\|\mu_k - h\|_1 \le k(2\epsilon' + \epsilon'')$. Now, notice that μ_k is the state-action distribution after k rounds of policy μ_h. By the mixing assumption, $\|\mu_k - \mu_h\|_1 \le e^{-k/\tau(h)}$, so the choice of $k = \tau(h)\log(1/\epsilon')$ yields $\|\mu_h - h\|_1 \le \tau(h)\log(1/\epsilon')(2\epsilon' + \epsilon'') + \epsilon'$.

□

The second ingredient is the validity of using estimates of the subgradients. We assume access to estimates of the subgradient of a convex cost function. Error bounds can be obtained from results in the stochastic convex optimization literature; the following theorem, a high-probability version of Lemma 3.1 of Flaxman, Kalai, and McMahan [20] for stochastic convex optimization, is sufficient. We note that the variance reduced stochastic gradient descent literature (e.g. SAGA or SVGR) cannot be directly applied since a full gradient calculation is impossible, and most complexity upper bounds are at least $O(\sqrt{XA}/\epsilon)$ [55], which is inappropriate for out setting.

Theorem 57. *Consider a bounded set $\mathcal{Z} \subset \mathbb{R}^d$ of radius Z (i.e. $\|z\| \le Z$ for all $z \in \mathcal{Z}$) and a sequence of real-valued convex cost functions $(f_t)_{t=1,2,\ldots,T}$. Let $z_1, z_2, \ldots, z_T \in \mathcal{Z}$ be the stochastic gradient decent path defined by defined by $z_1 = 0$ and $z_{t+1} = \Pi_{\mathcal{Z}}(z_t - \eta f'_t)$, where $\Pi_{\mathcal{Z}}$ is the Euclidean projection onto $\mathcal{Z}$, $\eta > 0$ is a learning rate, and $f'_1, \ldots, f'_T$ are bounded unbiased* *subgradient estimates; that is,* $\mathbb{E}\left[f'_t|z_t\right] = \nabla f(z_t)$ *and* $\|f'_t\| \le F$ *for some* $F > 0$. *Then, for* $\eta = Z/(F\sqrt{T})$ *and any* $\delta \in (0,1)$,

$$\sum_{t=1}^{T} f_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} f_t(z) \le ZF\sqrt{T} + \sqrt{(1+4Z^2T)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2T}{d}\right)\right)} \tag{4.25}$$

with probability at least $1-\delta$.

Proof. Let $z_* = \operatorname{argmin}_{z\in\mathcal{Z}} \sum_{t=1}^{T} f_t(z)$ and $\eta_t = f'_t - \nabla f_t(z_t)$. Define function $h_t : \mathcal{Z} \to \mathbb{R}$ by $h_t(z) = f_t(z) + z\eta_t$. Notice that $\nabla h_t(z_t) = \nabla f_t(z_t) + \eta_t = f'_t$. By Theorem 1 of Zinkevich [56], we get that

$$\sum_{t=1}^{T} h_t(z_t) - \sum_{t=1}^{T} h_t(z_*) \le \sum_{t=1}^{T} h_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} h_t(z) \le ZF\sqrt{T}\ .$$

Thus,

$$\sum_{t=1}^{T} f_t(z_t) - \sum_{t=1}^{T} f_t(z_*) \le ZF\sqrt{T} + \sum_{t=1}^{T} (z_* - z_t)\eta_t\ .$$

Let $S_t = \sum_{s=1}^{t-1}(z_* - z_s)\eta_s$, which is a self-normalized sum [14]. By Corollary 3.8 and Lemma E.3 of Abbasi-Yadkori [1], we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\begin{aligned} |S_t| &\le \sqrt{\left(1 + \sum_{s=1}^{t-1}(z_t - z_*)^2\right)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2t}{d}\right)\right)} \\ &\le \sqrt{(1+4Z^2t)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2t}{d}\right)\right)}\ . \end{aligned}$$

Thus,

$$\sum_{t=1}^{T} f_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} f_t(z) \le ZF\sqrt{T} + \sqrt{(1+4Z^2T)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2T}{d}\right)\right)}\ .$$

□

Remark 58. *Let* B_T *denote the RHS of* (4.25). *If all cost functions are equal to* f, *then by convexity of* f *and an application of Jensen's inequality, we obtain that* $f(\sum_{t=1}^{T} z_t/T) - \min_{z\in\mathcal{Z}} f(z) \le B_T/T$.

The last step before giving the proof of Theorem 55 is to apply Theorem 57 to our convex surrogate function, $c(\theta)$.

Lemma 59. *Under the same conditions as in Theorem 55 and any $\delta \in (0,1)$*

$$c(\widehat{\theta}_T) - \min_{\theta \in \Theta} c(\theta) \le \frac{S(\sqrt{d} + H(C_1 + C_2))}{\sqrt{T}} + \sqrt{\frac{1 + 4S^2T}{T^2}\left(2\log\frac{1}{\delta} + d\log\left(1 + \frac{S^2T}{d}\right)\right)} \tag{4.26}$$

with probability at least $1-\delta$,

Proof. We prove the lemma by showing that conditions of Theorem 57 are satisfied. The assumptions allow an easy bound on the subgradient estimate:

$$\|g_t\| \le \|\ell^\intercal \Phi\| + H\frac{\|\Phi_{(x_t,a_t),:}\|}{q_1(x_t,a_t)} + H\frac{\|(P-B)^\intercal_{:,x'_t}\Phi\|}{q_2(x'_t)} \le \sqrt{d} + H(C_1 + C_2)\ .$$

Also, we show that the subgradient estimate is unbiased:

$$\begin{aligned}\mathbb{E}\left[g_t(\theta\right] &= \ell^\intercal\Phi - H\sum_{(x,a)} q_1(x,a)\frac{\Phi_{(x,a),:}}{q_1(x,a)}\mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\}} \\ &\qquad\qquad\qquad + H\sum_{x'} q_2(x')\frac{(P-B)^\intercal_{:,x'}\Phi}{q_2(x')}\operatorname{sgn}((P-B)^\intercal_{:,x'}\Phi\theta) \\ &= \ell^\intercal\Phi - H\sum_{(x,a)}\Phi_{(x,a),:}\mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\}} + H\sum_{x'}(P-B)^\intercal_{:,x'}\Phi\operatorname{sgn}((P-B)^\intercal_{:,x'}\Phi\theta) \\ &= \nabla_\theta c(\theta)\ .\end{aligned}$$

The result then follows from Theorem 57 and Remark 58.

It is also convenient to bound the norm of the gradient. If $\mu_0(x,a) + \Phi_{(x,a),:}\theta \ge 0$, then $\nabla_\theta|[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_-| = 0$. Otherwise, $\nabla_\theta|[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_-| = -\Phi_{(x,a),:}$. Calculating,

$$\begin{aligned}\nabla_\theta c(\theta) &= \ell^\intercal\Phi + H\sum_{(x,a)}\nabla_\theta|[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_-| + H\sum_{x'}\nabla_\theta|(P-B)^\intercal_{:,x'}\Phi\theta| \\ &= \ell^\intercal\Phi - H\sum_{(x,a)}\Phi_{(x,a),:}\mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\}} + H\sum_{x'}(P-B)^\intercal_{:,x'}\Phi\operatorname{sgn}((P-B)^\intercal_{:,x'}\Phi\theta)\,,\end{aligned} \tag{4.27}$$

where $\operatorname{sgn}(z) = \mathbb{I}_{\{z>0\}} - \mathbb{I}_{\{z<0\}}$ is the sign function. Let ± denote the plus or minus sign (the exact sign does not matter here). We have that

$$\|\nabla_\theta c(\theta)\| \le H\sqrt{\sum_{i=1}^d\left(\sum_{x'}\left(\pm\sum_{(x,a)}(P-B)_{(x,a),x'}\Phi_{(x,a),i}\right)\right)^2} + \|\ell^\intercal\Phi\| + H\sqrt{\sum_{i=1}^d\left(\sum_{(x,a)}|\Phi_{(x,a),i}|\right)^2}\ .$$

Thus,

$$\begin{aligned}\|\nabla_\theta c(\theta)\| &\le \sqrt{\sum_{i=1}^d (\ell^\top \Phi_{:,i})^2} + H\sqrt{d} + H\sqrt{\sum_{i=1}^d \left(\sum_{(x,a)} \left(\pm \sum_{x'} (P-B)_{(x,a),x'}\right) \Phi_{(x,a),i}\right)^2} \\ &\le \sqrt{d} + H\sqrt{d} + H\sqrt{\sum_{i=1}^d \left(2\sum_{(x,a)} |\Phi_{(x,a),i}|\right)^2} = \sqrt{d}(1+3H)\,,\end{aligned}$$

where we used $|\ell^\top \Phi_{:,i}| \le \|\ell\|_\infty \|\Phi_{:,i}\|_1 \le 1$. □

With both ingredients in place, we can prove our main result.

Proof of Theorem 55. Let b_T be the RHS of (4.26). Using the trivial fact that $\sqrt{a+b} \le 2\sqrt{a}+2\sqrt{b}$, we can easily derive

$$b_T \le \frac{S}{\sqrt{T}}\left((\sqrt{d} + H(C_1+C_2)) + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1+\frac{S^2T}{d}\right)}\right) + O\left(\frac{1}{T}\right). \quad (4.28)$$

Lemma 59 implies that with high probability for any $\theta \in \Theta$,

$$\ell^\top(\mu_0 + \Phi\widehat{\theta}_T) + H\,V_1(\widehat{\theta}_T) + H\,V_2(\widehat{\theta}_T) \le \ell^\top(\mu_0+\Phi\theta) + H\,V_1(\theta) + H\,V_2(\theta) + b_T\,. \quad (4.29)$$

From (4.29), we get that

$$V_1(\widehat{\theta}_T) \le \frac{1}{H}\left(2(1+S) + H\,V_1(\theta) + H\,V_2(\theta) + b_T\right) := \epsilon'\,, \quad (4.30)$$

$$V_2(\widehat{\theta}_T) \le \frac{1}{H}\left(2(1+S) + H\,V_1(\theta) + H\,V_2(\theta) + b_T\right) := \epsilon''\,. \quad (4.31)$$

Inequalities (4.30) and (4.31) and Lemma 56 give the following bound:

$$|\ell^\top \mu_{\widehat{\theta}_T} - \ell^\top(\mu_0 + \Phi\widehat{\theta}_T)| \le \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon'+\epsilon'') + 3\epsilon'\,, \quad (4.32)$$

and we can similarly bound

$$|\ell^\top\mu_\theta - \ell^\top(\mu_0+\Phi\theta)| \le \tau(\mu_\theta)\log(1/V_1(\theta))(2V_1(\theta)+V_2(\theta)) + 3V_1(\theta). \quad (4.33)$$

Combining these two equation with (4.29) gives the final result:

$$\begin{aligned}\ell^\top\mu_{\widehat{\theta}_T} &\le \ell^\top(\mu_0+\Phi\widehat{\theta}_T) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon'+\epsilon'') + 3\epsilon' \\ &\le \ell^\top(\mu_0+\Phi\theta_T) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon'+\epsilon'') + 3\epsilon' + HV_1(\theta) + HV_2(\theta) + b_T \\ &\le \ell^\top\mu_\theta + \tau(\mu_\theta)\log(1/V_1(\theta))(2V_1(\theta)+V_2(\theta)) + 3V_1(\theta) \\ &\quad + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon'+\epsilon'') + 3\epsilon' + HV_1(\theta) + HV_2(\theta) + b_T \\ &\le \ell^\top\mu_\theta + 2\left(V_1(\theta)+V_2(\theta)\right)\Big(3 + \tau(\mu_\theta)\log(1/V_1(\theta)) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + H\Big) \\ &\quad + \Big(2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + 3\Big)\frac{2(1+S)}{H} + (2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')+3)\frac{b_T}{H} + b_T.\end{aligned}$$

Using the form of b_T above, we find the regret bound

$$\begin{aligned}
\ell^\intercal \mu_{\widehat{\theta}_T} &\le \ell^\intercal \mu_\theta + 2\left(V_1(\theta)+V_2(\theta)\right)\Big(3 + \tau(\mu_\theta)\log(1/V_1(\theta)) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + H\Big) \\
&\quad + \Big(2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')+3\Big)\frac{2(1+S)}{H} + \frac{S}{\sqrt{T}}H(C_1+C_2) \\
&\quad + \left(\frac{2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')+3}{H}+2\right)\frac{S}{\sqrt{T}}\sqrt{10\log\frac{1}{\delta}} \\
&\quad + O\left(\frac{\log(T)}{\sqrt{T}}\right) + O\left(\frac{1}{\sqrt{T}H}\right) \qquad (4.34) \\
&\le \ell^\intercal \mu_\theta + \left(V_1(\theta)+V_2(\theta)\right)O(H) + O\left(\frac{1}{H}\right) + O\left(\frac{H}{\sqrt{T}}\right) \\
&\quad + O\left(\frac{1}{H\sqrt{T}}\right)\sqrt{\log\frac{1}{\delta}} + O\left(\frac{\log(T)}{\sqrt{T}}\right) \qquad (4.35)
\end{aligned}$$

Now, recall that we set

$$H = \frac{1}{\epsilon} \quad \text{and} \quad T = \max\left\{\frac{H^2}{\epsilon^2}, 40S^2\log\frac{1}{\delta}\right\},$$

which finally yields that with high probability, for any $\theta \in \Theta$,

$$\ell^\intercal \mu_{\widehat{\theta}_T} \le \ell^\intercal \mu_\theta + \frac{1}{\epsilon}(V_1(\theta)+V_2(\theta)) + O(\epsilon),$$

as claimed. □

Comparison with Previous results

With a precise statement of our main result, we return to compare Theorem 55 from Farias and Van Roy [16]. Their approach is to relate the original MDP to a perturbed version [1] and then analyze the corresponding ALP. (See Section 4.1 for more details.) Let Ψ be a feature matrix that is used to estimate value functions. Recall that λ_* is the average loss of the optimal policy and λ_w is the average loss of the greedy policy with respect to value function Ψw. Let h_γ^* be the differential value function when the restart probability in the perturbed MDP is $1-\gamma$. For vector v and positive vector u, define the weighted maximum norm $\|v\|_{\infty,u} = \max_x u(x)|v(x)|$. Farias and Van Roy [16] prove that for appropriate constants $C, C' > 0$ and weight vector u,

$$\lambda_{w_*} - \lambda_* \le \frac{C}{1-\gamma}\min_w \|h_\gamma^* - \Psi w\|_{\infty,u} + C'(1-\gamma)\ . \qquad (4.36)$$

[1] In a perturbed MDP, the state process restarts with a certain probability to a *restart distribution*. Such perturbed MDPs are closely related to discounted MDPs.

This bound has similarities to bound (4.23): tightness of both bounds depends on the quality of feature vectors in representing the relevant quantities (stationary distributions in (4.23) and value functions in (4.36)). Once again, we emphasize that the algorithm proposed by Farias and Van Roy [16] is computationally expensive and requires access to a distribution that depends on optimal policy.

4.3 Average Cost Grid Algorithm

The main regret bound of the previous section, Theorem 55, may be sufficient for many applications, especially if one has reason to believe that $V_1(\theta)$ and $V_2(\theta)$ are small. For example, the features could include many stationary distributions (hence the error terms are zero). However, the result may be unsatisfying since the complexity is a function of H, but running the stochastic gradient procedure for more iterations will not guarantee a better result.

In this section, we will present an algorithm that obtains the more natural regret bound

$$\ell^\intercal \mu_{\widehat{\theta}_T} \leq \min_{\theta \in \Theta} \ell^\intercal \mu_\theta + O\left(\sqrt{V_1(\theta) + V_2(\theta)}\right) + O(\epsilon).$$

Intuitively, we accomplish this by also optimizing over the violation scaling parameter H, thus approximately computing

$$\min_{\theta \in \Theta, H \in \mathbb{R}} \left(\ell^\intercal \mu_\theta + H(V_1(\theta) + V_2(\theta)) + \frac{\beta}{H}\right), \tag{4.37}$$

where β is some constant, but in particular can be chosen to match the coefficient on the $O(\epsilon)$ of the regret bound in Theorem 55:

$$\beta = S(C_1 + C_2 + 3\tau(\mu_{\widehat{\theta}_k}) + 3\tau(\mu_{\theta^*}) + 2). \tag{4.38}$$

However, if one does not require the dependence on S or $C_1 + C_2$ to be correct, $\beta = 1$ is acceptable.

Throughout the section, we use the following notation to simplify the presentation. We define $c(H, \theta) := \ell^\intercal \Phi\theta + H(V_1(\theta) + V_2(\theta))$ and $\theta^*_H := \operatorname{argmin}_\theta c(H, \theta)$. To obtain the $O\left(\sqrt{V_1(\theta^*) + V_2(\theta^*)}\right)$ regret, it suffices to solve the optimization problem

$$\min_{H,\theta} c(H, \theta) + \frac{\beta}{H}.$$

The objective is convex in H and θ individually but not jointly and hence we cannot use gradient descent. One might try alternating minimizing over θ and H, but this approach is difficult to analyze.

Instead, we will exploit what we already know how to do: approximate θ^*_H. If we define $F(H) = c(H, \theta^*_H) + \frac{\beta}{H}$, then we will show that approximating the minimum H is sufficient. Hence, we will analyze grid search, where we (precisely) select a grid $H_1, \ldots, H_K$ then use Algorithm 4.1 to approximate $\widehat{\theta}$ for every grid point.

The Grid Algorithm The *grid algorithm* takes as inputs a bound on the violation function $V_{\max}$, a desired error tolerance ϵ, and desired probability tolerance δ. The algorithm then carefully chooses a grid $H_1, \ldots, H_K$, and for each $i = 1, \ldots, K$, computes $\hat{\theta}_i$, the output of Algorithm 4.1, and $\widehat{V}_i$, an approximation to $V_1(\hat{\theta}_i) + V_2(\hat{\theta}_i)$. It then returns

$$\hat{k} := \operatorname*{argmin}_k \ell^\intercal \Phi \widehat{\theta}_k + H_k \widehat{V}_k + \frac{\beta}{H_k}. \tag{4.39}$$

Estimating the Error Functions

To run the Grid Algorithm, we need to be able to estimate the constraint violations $V_1(\theta) + V_2(\theta)$. Similar to the gradient estimate, we estimate $V_1 + V_2$ by importance-weighted sampling. For some n and samples $y_1, \ldots, y_n \sim q_1$ and $(x_1, a_1), \ldots, (x_n, a_n) \sim q_2$, define

$$\widehat{V}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \frac{[\mu_0(x_i, a_i) + \Phi_{(x_i,a_i),:}\theta]_-}{q_1(x, a)} + \frac{|(P - B)^\intercal_{:,y_i} \Phi \theta|}{q_2(y_i)}. \tag{4.40}$$

Since $V_1(\theta) = \sum_{(x,a)} |[\mu_0(x, a) + \Phi_{(x,a),:}\theta]_-|$ and $V_2(\theta) = \sum_{x'} |(P - B)^\intercal_{:,x'} \Phi \theta|$, this estimate is clearly unbiased. Also, we earlier assumed the existence of constants $C_1 = \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \frac{\|\Phi_{(x,a),:}\|}{q_1(x,a)}$ and $C_2 = \max_{x\in\mathcal{X}} \frac{\|(P-B)^\intercal_{:,x}\Phi\|}{q_2(x)}$, and so we can bound

$$\frac{[\mu_0(x_i, a_i) + \Phi_{(x_i,a_i),:}\theta]_-}{q_1(x, a)} + \frac{|(P - B)^\intercal_{:,y_i} \Phi \theta|}{q_2(y_i)} \le S(C_1 + 1) + SC_2$$

which gives us concentration of $\widehat{V}$ around V. In particular, applying Hoeffding's inequality yields:

Lemma 60. *Given $\epsilon > 0$ and $\delta \in [0, 1]$, for any θ, the violation function estimate $\widehat{V}_n(\theta)$ has*

$$|\widehat{V}_n(\theta) - (V_1(\theta) + V_2(\theta))| \le \epsilon$$

with probability at least $1 - \delta$ as long as we choose $n \ge \frac{(S(C_1+1)+SC_2)^2}{2\epsilon^2} \log\left(\frac{2}{\delta}\right)$.

Figure 4.2 provides a precise definition of the algorithm and specifies the grid, parameters for the SGD algorithm, and sample sizes for $\widehat{V}_k$ needed. The rest of Section 4.3 is devoted to analyzing this algorithm and proving that is does achieve a $O(\sqrt{V_1(\theta^*) + V_2(\theta^*)})$ regret.

Theorem 61. *For some $\epsilon > 0$ and $\delta \in [0, 1]$, the Grid Algorithm specified in Figure 4.2 has regret*

$$\mu_{\widehat{\theta}_T}^\intercal \ell \le \mu_\theta^\intercal \ell + O\left(\sqrt{V_1(\theta) + V_2(\theta)}\right) + O(\epsilon) \tag{4.44}$$

with probability at least $1 - \delta$.

The proof is delayed until Section 4.3, but a rough outline is as follows. We construct the grid $H_1, \ldots, H_K$ such that $\max_{H_k \le H \le H_{k+1}} F(H) - F(H_k)$ is always $\frac{\epsilon}{2}$ by reasoning about the Lipschitz constant of $F(H)$. This lets us conclude that $H_{\hat{k}}, \widehat{\theta}_{\hat{k}}$ produces an objective value close to that of H^*, θ^*, which has the desired $O\left(\sqrt{V_1(\theta^*) + V_2(\theta^*)}\right)$ regret.

```
Input: Upper bound $V_1(\theta) + V_2(\theta) \le V_{\max}$, error tolerance $\epsilon > 0$,
probability $\delta > 0$
Compute grid
    $H_k := \begin{cases} \beta \left( \frac{1}{\sqrt{V_{\max}}} - \frac{k-1}{4}\epsilon \right)^{-1} & \text{if } k \le \overline{k} \\ H_{\overline{k}} + \frac{\epsilon(k-\overline{k})}{4V_{\max}} & \text{otherwise} \end{cases} \qquad (4.41)$
where
    $\overline{k} := \min \left\{ i : V_{\max}(H_{i+1} - H_i) \ge \frac{\epsilon}{4} \right\}, \quad \text{and} \qquad (4.42)$
    $K := \min\{k : H_k \ge 4\frac{\beta}{\epsilon}\}. \qquad (4.43)$

for $k := 1, 2, \ldots, K$ do
    Obtain $\widehat{\theta}_k$ from Algorithm 4.1 with $T = \max\left\{ \frac{H^2}{\epsilon^2}, 40S^2 \log \frac{K}{\delta} \right\}$
    Set $n = \frac{8(S(C_1+1)+SC_2)^2}{\epsilon^2} \log\left(\frac{4K}{\delta}\right)$
    Set $\widehat{V}_k = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{[\mu_0(x_i,a_i) + \Phi_{(x_i,a_i),:}\hat{\theta}_k]_-}{q_1(x,a)} + \frac{|(P-B)^\intercal_{:,y_i}\Phi\hat{\theta}_k|}{q_2(y_i)} \right]$
end for
Set $k^* = \operatorname{argmin}_k \ell^\intercal \Phi \hat{\theta}_k + H_k \widehat{V}_k + \frac{\beta}{H_k}$
Return policy $\pi_{\widehat{\theta}_{\hat{k}}}$
```

Figure 4.2: The Grid algorithm

Analysis

The main idea of the proof is to show that $F(H)$ (the optimization in H only after the optimal θ is used) is well behaved and Lipshitz; we construct the sequence H_k such that $\max_{H_k \le H \le H_{k+1}} F(H)$ is always $\frac{\epsilon}{2}$. Finally, we show that the error from finding the minimum H_k and the error for the approximate optimization of θ given a fixed H add up to an order ϵ error.

Lemma 62. *Let $\epsilon > 0$ be some desired error tolerance. Let $V_{\max}$ be some upper bound on $V_1(\theta) + V_2(\theta)$; we can always take $V_{\max} = 3 + S(d+2)$. The sequence defined by*

$$H_k := \begin{cases} \beta \left(\frac{1}{\sqrt{V_{\max}}} - \frac{k-1}{2}\epsilon \right)^{-1} & \textit{if } k \le \overline{k} \\ H_{\overline{k}} + \frac{\epsilon(k-\overline{k})}{2V_{\max}} & \overline{k} < k \le K \end{cases} \tag{4.45}$$

for

$$\overline{k} := \min\left\{i : V_{\max}(H_{i+1} - H_i) \geq \frac{\epsilon}{2}\right\} \quad \textit{and} \quad K = \min\left\{k : H_k \geq \frac{2\beta}{\epsilon}\right\} \tag{4.46}$$

guarantees that, for all $1 \leq k \leq K$*,*

$$\max_{H_k \leq H \leq H_{k+1}} |F(H_k) - F(H)| \leq \epsilon. \tag{4.47}$$

Proof. Since $c(H, \theta)$ is linear in H, $c(H, \theta^*_H)$ is concave in H. Also, using δ to denote some positive number, we can show that $c(H, \theta^*_H)$ is increasing

$$\begin{aligned} c(H, \theta^*_H) &= \min_\theta \ell^\intercal \Phi\theta + H(V_1(\theta) + V_2(\theta)) \\ &\leq \min_\theta \ell^\intercal \Phi\theta + (H + \delta)(V_1(\theta) + V_2(\theta)) \\ &= c(H + \delta, \theta^*_{H+\delta}) \end{aligned}$$

and sub-linear

$$\begin{aligned} c(H + \delta, \theta^*_{H+\delta}) &= \min_\theta \ell^\intercal \Phi\theta + (H + \delta)(V_1(\theta) + V_2(\theta)) \\ &\leq \ell^\intercal \Phi\theta^*_H + (H + \delta)(V_1(\theta^*_H) + V_2(\theta^*_H)) \\ &= c(H, \theta^*_H) + \delta(V_1(\theta^*_H) + V_2(\theta^*_H)) \\ &\leq c(H, \theta^*_H) + \delta V_{\max}. \end{aligned}$$

Using the monotonicity property of $c(H, \theta^*_H)$, we have

$$\begin{aligned} \max_{H_i \leq H \leq H_{i+1}} |F(H_i) - F(H)| &\leq \max_{H_i \leq H \leq H_{i+1}} c(H_{i+1}, \theta^*_{H_{i+1}}) - c(H_i, \theta^*_{H_i}) + \beta\left(\frac{1}{H_i} - \frac{1}{H_{i+1}}\right) \\ &\leq (H_{i+1} - H_i)V_{\max} + \beta\left(\frac{1}{H_i} - \frac{1}{H_{i+1}}\right). \end{aligned}$$

First, consider the case where $k \leq \overline{k}$. Then $H_k = \beta\left(\sqrt{V_{\max}} - \frac{k-1}{2}\epsilon\right)^{-1}$ so that $\left(\frac{1}{H_k} - \frac{1}{H_{k+1}}\right) = \frac{\epsilon}{2}$. By definition of $\overline{k}$, we also have that $V_{\max}(H_{i+1} - H_i) \leq \frac{\epsilon}{2}$. In the case where $k > \overline{k}$, we have $\beta\left(\frac{1}{H_k} - \frac{1}{H_{k+1}}\right) < (H_{k+1} - H_k)V_{\max}$ and $(H_{k+1} - H_k)V_{\max} = \frac{\epsilon}{2}$. In either case,

$$|F(H_{k+1}) - F(H_k)| \leq (H_{k+1} - H_k)V_{\max} + \beta\left(\frac{1}{H_k} - \frac{1}{H_{k+1}}\right) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \leq \epsilon$$

as desired.

Finally, we check that if $H^* > H_K$, then $|F(H^*) - F(H_K)| \leq \epsilon$. Note that

$$F(H^*) = \min_{H,\theta} \ell^\intercal \Phi\theta + H(V_1(\theta) + V_2(\theta)) + \frac{\beta}{H} = \min_\theta \ell^\intercal \Phi\theta + 2\sqrt{\beta(V_1(\theta) + V_2(\theta))}$$

and

$$F(H_K) = \min_\theta \ell^\intercal \Phi\theta + \sqrt{\beta}\left(\frac{V_1(\theta)+V_2(\theta)}{\epsilon} + \epsilon\right),$$

where we solved for the optimum H for a given θ as $H^*_\theta = \sqrt{\beta/(V_1(\theta)+V_2(\theta))}$. Therefore, $H^* > H_K$ implies that $V_1(\theta^*) + V_2(\theta^*) \le \frac{\epsilon^2}{4\beta}$. Then

$$\begin{aligned} F(H_K) - F(H^*) &= \min_{\theta'}\left(\ell^\intercal\Phi\theta' + \frac{2\beta}{\epsilon}\left(V_1(\theta')+V_2(\theta')\right) + \frac{\epsilon}{2}\right) - \left(\min_\theta \ell^\intercal\Phi\theta + 2\sqrt{\beta(V_1(\theta)+V_2(\theta))}\right) \\ &\le \max_\theta \ell^\intercal\Phi\theta + \frac{2\beta}{\epsilon}\left(V_1(\theta)+V_2(\theta)\right) + \frac{\epsilon}{2} - \left(\ell^\intercal\Phi\theta + 2\sqrt{\beta(V_1(\theta)+V_2(\theta))}\right) \\ &\le \max_\theta \frac{2\beta}{\epsilon}\left(V_1(\theta)+V_2(\theta)\right) + \frac{\epsilon}{2} \\ &\le \epsilon, \end{aligned}$$

completing the last case. □

We can now formally state and prove Theorem 61.

Proof. Running Algorithm 4.1 for $H_1, \ldots, H_K$ with $T = 16\frac{H_k^2}{\epsilon^2}$, produces a sequence $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ such that

$$c(H_k, \widehat{\theta}_K) \le c(H_k, \theta^*_K) + H_k V(\theta^*) + \frac{\beta}{H_k} + \frac{\epsilon}{4}$$

holds for all k simultaneously with probability at least $1 - \frac{\delta}{2}$, which is easily argued by noting that the probability of error for any single k is δ/K and applying the union bound. Let θ^*_k be the true optimal θ of $C(H_k, \theta)$. Additionally, Lemma 60, along with our choice of $n = \frac{8(S(C_1+1)+SC_2)^2}{\epsilon^2}\log\left(\frac{4K}{\delta}\right)$, guarantees that, with probability at least $1 - \frac{\delta}{2K}$, $|V_1(\widehat{\theta}_k) + V_2(\widehat{\theta}_k) - \widehat{V}_k| \le \frac{\epsilon}{4}$, and hence the statement holds for all $\widehat{V}_k$ with probability at most $1 - \frac{\delta}{2}$.

The first step is bounding the suboptimality of the objective. Recalling that $\hat{k}$ is the minimizer of $\ell^\intercal\Phi\widehat{\theta}_k + H_k\widehat{V}_k + \frac{\beta}{H_k}$, and using k^* as the minimizer of $c(H_k, \theta^*_k) + \frac{\beta}{H_k}$, we have

$$\begin{aligned} \ell^\intercal\Phi\widehat{\theta}_{\hat{k}} + H_{\hat{k}}\widehat{V}_{\hat{k}} + \frac{\beta}{H_{\hat{k}}} &= \min_k \ell^\intercal\Phi\widehat{\theta}_k + H_k\widehat{V}_k + \frac{\beta}{H_k} \\ &\le \ell^\intercal\Phi\widehat{\theta}_{k^*} + H_{k^*}\widehat{V}_{k^*} + \frac{\beta}{H_{k^*}} \\ &\le c(H_{k^*}, \widehat{\theta}_{k^*}) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{4} && \text{(Lemma 60)} \\ &\le c(H_{k^*}, \theta^*_{k^*}) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{2} && \text{(Lemma 59)} \\ &= \min_k c(H_k, \theta^*_k) + \frac{\beta}{H_k} + \frac{\epsilon}{2} \\ &\le \min_{H,\theta} c(H,\theta) + \frac{\beta}{H} + \epsilon && \text{(Lemma 62).} \end{aligned}$$

The statement holds with probability at least $\frac{\delta}{2} + \frac{\delta}{2}$, where the first term is from estimating $\widehat{V}_k$ (Lemma 60) and the second term is from bounding the SGD error (Lemma 59). Hence, the Gird Algorithm minimizes the objective to within ϵ.

A bound of the suboptimality of the objective is not enough. We wish to prove a statement about the regret of $\ell^\top \mu_{\theta_{\hat{k}}}$, and hence we must relate this quantity to $\ell^\top \Phi \hat{\theta}_{\hat{k}}$. Since all quantities are non-negative, this implies that $|\frac{\beta}{H_{\hat{k}}} - \frac{\beta}{H^*}| \le \epsilon$. Finally, we can put together the regret bound. To apply Lemma 3 and bound the distance between $\ell^\top \Phi \mu_{\widehat{\theta}_{\hat{k}}}$ and $\ell^\top \Phi \widehat{\theta}_{\hat{k}}$, we first need to bound $V_1(\widehat{\theta}_{\hat{k}})$ and $V_2(\widehat{\theta}_{\hat{k}})$. Using the bounded suboptimality of $\widehat{\theta}_{\hat{k}}$ as an optimizer of $c(H_{\hat{k}}, \theta)$, we have

$$\begin{aligned}
\ell^\top \Phi \widehat{\theta}_{\hat{k}} + H_{\hat{k}} \left(V_1(\widehat{\theta}_{\hat{k}}) + V_2(\widehat{\theta}_{\hat{k}}) \right) &\le \ell^\top \Phi \theta^*_{\hat{k}} + H_{\hat{k}} \left(V_1(\theta^*_{\hat{k}}) + V_2(\theta^*_{\hat{k}}) \right) + \frac{\epsilon}{2} \\
&\le \ell^\top \Phi \theta^* + H^* \left(V_1(\theta^*) + V_2(\theta^*) \right) + \epsilon
\end{aligned}$$

and can conclude that

$$\begin{aligned}
V_1(\widehat{\theta}_{\hat{k}}) &\le \frac{1}{H_{\hat{k}}} \left(2(S+1) + \sqrt{V_1(\theta^*) + V_2(\theta^*)} \right) \\
&\le \left(\frac{1}{H^*} + \epsilon \right) \left(2(S+1) + \sqrt{V_1(\theta^*) + V_2(\theta^*)} \right) \\
&= (2(S+1) + \epsilon) \sqrt{V_1(\theta^*) + V_2(\theta^*)} + (V_1(\theta^*) + V_2(\theta^*)) + 2(S+1)\epsilon.
\end{aligned}$$

Completely analogous reasoning gives the same bound on $V_2(\widehat{\theta}_{\hat{k}})$.

Then, applying Lemma 56, we have

$$\begin{aligned}
\ell^\top \Phi \mu_{\theta_{\hat{k}}} &\le \ell^\top \Phi \widehat{\theta}_{\hat{k}} + 4\tau(\mu_{\theta_{\hat{k}}}) \log(1/\epsilon') \left((2(S+1) + \epsilon) \sqrt{V_1(\theta^*) + V_2(\theta^*)} + (V_1(\theta^*) + V_2(\theta^*)) + 2(S+1)\epsilon \right) \\
&\le \ell^\top \Phi \widehat{\theta}^* + 4\tau(\mu_{\theta_{\hat{k}}}) \log(1/\epsilon') \left((2(S+1) + \epsilon) \sqrt{V_1(\theta^*) + V_2(\theta^*)} + (V_1(\theta^*) + V_2(\theta^*)) + 2(S+1)\epsilon \right) \\
&\quad + H^*(V_1(\theta^*) + V_2(\theta^*)) + \frac{\beta}{H^*} + \epsilon \\
&\le \ell^\top \mu_{\theta^*} + 4\tau(\mu_{\theta_{\hat{k}}}) \log(1/\epsilon') \left((2(S+1) + \epsilon) \sqrt{V_1(\theta^*) + V_2(\theta^*)} + (V_1(\theta^*) + V_2(\theta^*)) + 2(S+1)\epsilon \right) \\
&\quad + H^*(V_1(\theta^*) + V_2(\theta^*)) + \frac{\beta}{H^*} + \epsilon + (V_1(\theta^*) + V_2(\theta^*)).
\end{aligned}$$

Now, using that $H^* = \left(\sqrt{V_1(\theta) + V_2(\theta)} \right)^{-1}$, the previous statement implies

$$\ell^\top \mu_{\theta_{\hat{k}}} \le \min_\theta \ell^\top \mu_\theta + O\left(\sqrt{V_1(\theta) + V_2(\theta)} \right) + O\left(V_1(\theta) + V_2(\theta) \right) + O(\epsilon),$$

which produces the theorem statement, because $V_1(\theta) + V_2(\theta) \le 1$. □

4.4 The Dual ALP for Discounted Cost

We now change settings to discounted cost but the problem remains: find a policy with discounted loss almost as low as the best in the class. Fortunately, most of the tools from the average cost carry over with small modifications. Hence, this section studies how to approximately solve θ given an H, and the next section studies how to use a grid algorithm to optimize H as well.

Recall that the LP we intend to approximately solve is

$$\begin{aligned} &\min_{\theta\in\mathbb{R}^d} \ \ell^\intercal \Phi\theta\,, \\ &\text{s.t.} \quad (B-\gamma P)^\intercal \Phi\theta = \alpha, \quad \Phi\theta \geq 0. \end{aligned} \tag{4.10}$$

This LP has another interpretation. The dual of the approximate dual is

$$\begin{aligned} \max\nolimits_{J\in\mathbb{R}^X} \quad & \alpha^\intercal J \\ \text{s.t.} \quad & \Phi^\intercal\left(\ell + (\gamma P - B)J - z\right) = 0, \\ & z \geq 0, \end{aligned}$$

which is the original primal with a form of constraint aggregation.

Approximately solving the LP Analogous to V_1 and V_2, we define, relative to a feature matrix Φ, the constraint violation functions

$$V_3(\theta) := \|\left[\Phi\theta\right]_-\|_1 \quad \text{and} \tag{4.48}$$

$$V_4(\theta) := \|(B-\gamma P)^T\Phi\theta - \alpha\| \tag{4.49}$$

so that we can approximate the solution of the LP by minimizing the convex surrogate

$$\begin{aligned} c^\gamma(\theta) &:= \ell^\intercal\Phi\theta + H\left(V_3(\theta) + V_4(\theta)\right) \\ &= \ell^\intercal\Phi\theta + H\|\left[\Phi\theta\right]_-\|_1 + H\|(B-\gamma P)^\intercal\Phi\theta - \alpha\|_1 \\ &= \ell^\intercal\Phi\theta + H\sum_{(x,a)}\left[\Phi_{(x,a),:}\theta\right]_- + H\sum_{x'}\left|(B-\gamma P)^\intercal_{:,x'}\Phi\theta - \alpha\right| \end{aligned} \tag{4.50}$$

with some constant H and the constraint set $\Theta = \{\theta : \|\theta\|_2 \leq S\}$.

We will minimize (4.50) through stochastic gradient descent. We will sample constraints for V_3 and V_4 with distributions $q_3 \in \triangle_{\mathcal{X}\times\mathcal{A}}$ and $q_4 \in \triangle_{\mathcal{X}}$, respectively, and calculate the unbiased subgradient

$$g_t^\gamma(\theta) = \ell^\intercal\Phi - H\frac{\Phi_{(x_t,a_t),:}}{q_3(x_t,a_t)}\mathbb{I}_{\{\Phi_{(x_t,a_t),:}\theta<0\}} + H\frac{(P-\gamma B)^\intercal_{:,x'_t}\Phi}{q_4(x'_t)}\operatorname{sgn}((P-\gamma B)^\intercal_{:,x'_t}\Phi\theta). \tag{4.51}$$

Then, the algorithm is exactly the same as Figure 4.1 with g_t^γ instead of g_t. Recall that we are using the shorthand

$$J_\theta = J_{\pi_{\Phi\theta}} \quad \text{and} \quad \nu_\theta = \nu_{\pi_{\Phi\theta}}. \tag{4.52}$$

Thus, our objective is to show that $\alpha^\intercal J_{\widehat{\theta}_T}$ is small.

A key difference between the average and discounted cases is the interpretation for the dual variables, μ and ν. In the average case, the feasible μ exactly corresponded to stationary distributions and therefore the average loss was precisely $\ell^\intercal \mu$. However, in the discounted case, the dual variables ν correspond to the expected discounted number of visits to each state and $\ell^\intercal \nu = \alpha^\intercal J$, where J is the value function corresponding to policy π_ν.

A Regret Bound for the Discounted Case

Unlike the average cost case, the discounted cost case does not need a fast mixing assumption. However, we do need to assume that the operator 1-norm of Φ is upper bounded by some constant C:

$$\|\Phi\|_1 = \max_{x:\|x\|_1=1} \|\Phi x\|_1 = \max_{1\le j\le d} \sum_{(x,a)} |\Phi_{(x,a),j}| \le C. \tag{4.53}$$

As before, we will need to choose constraint sampling distributions such that we can bound

$$C_3 = \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \frac{\|\Phi_{(x,a),:}\|}{q_3(x,a)}\,, \qquad C_4 = \max_{x\in\mathcal{X}} \frac{\|(P-\gamma B)^\intercal_{:,x}\Phi\|}{q_4(x)}\,.$$

While special structure may suggest natural choices of sampling distributions to ensure small C_3 and C_4, there are general conditions that allow these constants to be bounded. For example, if there is $C' > 0$ such that for any (x,a) and i, $\Phi_{(x,a),i} \le C'/(XA)$ and each column of P has only N non-zero elements, we can choose q_3 and q_4 to be uniform distributions and we can bound

$$\frac{\|\Phi_{(x,a),:}\|}{q_3(x,a)} \le C'\,, \qquad \frac{\|(P-\gamma B)^\intercal_{:,x}\Phi\|}{q_4(x)} \le C'(N+A)\,.$$

Finally, note that we can always upper bound the constraint violation functions. For any $\theta \in \Theta$,

$$\begin{aligned}
V_3(\theta) &\le \|\Phi\theta\|_1 \le \sum_{j=1}^{d}\sum_{(x,a)} |\Phi_{(x,a),j}||\theta_j| \le C\|\theta\|_1 \le C\sqrt{d}\|\theta\|_2 \le \sqrt{d}\,CS \\
V_4(\theta) &\le \sum_{x'} |B^\intercal_{:,x'}(\Phi\theta)| + \gamma \sum_{x'} |P^\intercal_{:,x'}(\Phi\theta)| + \|\alpha\|_1 \\
&\le \sum_{(x,a)} \left(\sum_{x'} B_{(x,a),x'}\right) |(\Phi\theta)_{(x,a)}| + \gamma \sum_{(x,a)} \left(\sum_{x'} P_{(x,a),x'}\right) |(\Phi\theta)_{(x,a)}| + 1 \\
&= (1+\gamma)\|\Phi\theta\|_1 + 1 \\
&\le (1+\gamma)\sqrt{d}\,CS + 1
\end{aligned}$$

and hence

$$V_3(\theta) + V_4(\theta) \le 1 + \sqrt{d}CS(2+\gamma) \le 4\sqrt{d}CS \tag{4.54}$$

as long as $C, S \ge 1$.

We now present the regret bound for discounted cost and a fixed H.

Theorem 63. *Consider an expanded efficient large-scale dual ALP problem and some error tolerance $\epsilon > 0$ and desired maximum probability of error $\delta > 0$. Running the stochastic subgradient method (shown in Figure 4.1) with $H \geq 1$, $T = O\left(\frac{S^2 \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$ and constant learning rate $\eta = S/(G'\sqrt{T})$, where $G' = \sqrt{d} + H(C_3 + C_4)$, yields a $\widehat{\theta}_T$ with*

$$\ell^\intercal \nu_{\widehat{\theta}_T} \leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)(V_3(\theta) + V_4(\theta)) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + O(\epsilon).$$

Constants hidden in the big-O notation are polynomials in S, d, C_3, C_4, and C.

Analysis

Here, we present the proof of the main result of this section, beginning with showing that if some vector ν is close to a feasible point of the LP, then it almost equals the expected frequencies of visits of the policy π_ν (when the system runs under the policy π_h with the initial distribution α), i.e.,

$$\nu_{\pi_\nu}(x, a) = \sum_{x'} \alpha(x') \sum_{t=1}^{\infty} \gamma^{t-1} P^{\pi_h}\left(x_t = x, a_t = a | x_1 = x'\right). \tag{4.55}$$

Lemma 64. *For any vector $\nu \in \mathbb{R}^{XA}$, let $\mathcal{N}$ be the set of points (x, a) where $\nu(x, a) \leq 0$ and $\mathcal{S} = \mathcal{N}^c$ and define the constants $\sum_{(x,a)\in\mathcal{N}} |\nu(x, a)| = \epsilon'$ and $\|(B - \gamma P)^\intercal \nu - \alpha\|_1 = \epsilon''$. Further assume that for each x, there exists an a such that $(x, a) \in \mathcal{S}$. Then, for the policy π_ν define by*

$$\pi_\nu(a|x) = \frac{[\nu(x, a)]_+}{\sum_{a'}[\nu(x, a')]_+}, \tag{4.56}$$

the expected frequencies of visits under the policy is close to ν:

$$\|\nu_{\pi_\nu} - \nu\|_1 \leq \frac{3\epsilon' + \epsilon''}{1-\gamma}.$$

Proof. First, we notice that,

$$\|\,[\nu]_+ - \nu\|_1 \leq \sum_{(x,a)\in\mathcal{N}} |\nu(x, a)|_1 = \epsilon'. \tag{4.57}$$

Let $\xi = (B - \gamma P)^\intercal \nu - \alpha \in \mathbb{R}^X$ with $\|\xi\|_1 = \epsilon''$ according to the assumption. For any $x' \in \mathcal{X}$, we have,

$$\sum_{(x,a)\in\mathcal{S}} \nu(x, a)(B - \gamma P)_{(x,a),x'} - \alpha(x') = - \sum_{(x,a)\in\mathcal{N}} \nu(x, a)(B - \gamma P)_{(x,a),x'} + \xi(x').$$

Let $v_0 = (B - \gamma P)^\intercal h - \alpha$, we have

$$\begin{aligned}
\|v_0\|_1 &= \sum_{x'} \left| \sum_{(x,a)} h(x,a)(B-\gamma P)_{(x,a),x'} - \alpha(x') \right| \\
&= \sum_{x'} \left| \sum_{(x,a)\in\mathcal{S}} \nu(x,a)(B-\gamma P)_{(x,a),x'} - \alpha(x') \right| \\
&= \sum_{x'} \left| - \sum_{(x,a)\in\mathcal{N}} \nu(x,a)(B-\gamma P)_{(x,a),x'} + \xi(x') \right| \qquad (4.58)
\end{aligned}$$

with the upper bound

$$\begin{aligned}
\|v_0\|_1 &\leq \sum_{x'} \left| - \sum_{(x,a)\in\mathcal{N}} \nu(x,a)(B-\gamma P)_{(x,a),x'} \right| + \|\xi\|_1 \\
&\leq \sum_{(x,a)\in\mathcal{N}} \left(|\nu(x,a)| \sum_{x'} \left|(B-\gamma P)_{(x,a),x'}\right| \right) + \epsilon'' \\
&\leq 2 \sum_{(x,a)\in\mathcal{N}} |\nu(x,a)| + \epsilon'' \\
&\leq 2\epsilon' + \epsilon''. \qquad (4.59)
\end{aligned}$$

Let M^h be a $X \times (XA)$ matrix that encodes the policy π_ν, where $M^h_{(i,(i-1)A+1)-(i,iA)} = \pi_\nu\left(\cdot|x_i\right)$. As a concrete example, when the state space $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{A} = \{a_1, a_2\}$, then

$$M^h = \begin{pmatrix} \pi_\nu(a_1|x_1) & \pi_\nu(a_2|x_1) & 0 & 0 \\ 0 & 0 & \pi_\nu(a_1|x_2) & \pi_\nu(a_2|x_2) \end{pmatrix}.$$

By the definition of π_ν in (4.56), it is easy to check that $h^\intercal B M^h = h^\intercal$.

With M^h, the ν_{π_h} defined in (4.55) can be written as,

$$\nu_{\pi_h}^\intercal = \sum_{t=1}^{\infty} \gamma^{t-1} \alpha^\intercal M^h (PM^h)^{t-1} \qquad (4.60)$$

Now, we are ready to bound $\|\nu_{\pi_\nu} - \nu\|_1$. By the definition of v_0 (i.e., $v_0 = (B - \gamma P)^\intercal h - \alpha$), we have,

$$\alpha^\intercal M^h = h^\intercal B M^h - \gamma h^\intercal P M^h - v_0^\intercal M^h = h^\intercal - \gamma h^\intercal P M^h - v_0^\intercal M^h,$$

where the last equality is due to $h^\intercal B M^h = h^\intercal$. Therefore,

$$\alpha^\intercal M^h (PM)^{t-1} = h^\intercal (PM^h)^{t-1} - \gamma h^\intercal (PM^h)^t - v_0^\intercal M^h (PM)^{t-1},$$

By (4.60), we have,

$$\nu_{\pi_h}^{\intercal} = h^{\intercal} - \sum_{t=1}^{\infty} \gamma^{t-1} v_0^{\intercal} M_h (PM^h)^{t-1}. \tag{4.61}$$

Let $z_t = v_0^{\intercal} M_h (PM^h)^t$. By (4.59), we have

$$\|z_0\| = \|v_0^{\intercal} M_h\|_1 = \sum_{x,a} |v_0(x)\pi_\nu(a|x)| \le \sum_x \left(|v_0(x)| \sum_a |\pi_\nu(a|x)| \right) = \|v_0\|_1 \le 2\epsilon' + \epsilon''.$$

Further,

$$\begin{aligned}\|z_{t+1}\|_1 = \|z_t PM^h\|_1 &= \sum_{x,a} \sum_{x',a'} |z_t(x',a') P(x|x',a') \pi_\nu(a|x)| \\ &\le \sum_{x,a} \left(|z_t(x',a')| \sum_{x',a'} |P_{\pi_\nu}(x,a|x',a')| \right) = \|z_t\|_1.\end{aligned}$$

By the induction, we know that $\|z_t\|_1 \le 2\epsilon' + \epsilon''$ for all t. By (4.61),

$$\|\nu_{\pi_h} - h\|_1 \le \sum_{t=1}^{\infty} \gamma^{t-1} \|z_{t-1}\|_1 \le \frac{2\epsilon' + \epsilon''}{1-\gamma}. \tag{4.62}$$

Combining this with (4.57) and the triangle inequality,

$$\|\nu_{\pi_h} - \nu\|_1 \le \frac{2\epsilon' + \epsilon''}{1-\gamma} + \epsilon' \le \frac{3\epsilon' + \epsilon''}{1-\gamma}. \tag{4.63}$$

□

Next, we need the analog of Lemma 59 for the discounted case, which is again a direct application of Theorem 57.

Lemma 65. *Given some error tolerance $\epsilon > 0$ and desired maximum probability of error $\delta > 0$, running the stochastic subgradient method (shown in Figure 4.1) on $c^\gamma(\theta)$ with $T \ge 1/\epsilon^4$, $H = 1/\epsilon$, and constant learning rate $\eta = \frac{S}{\sqrt{T}}\left(\sqrt{d} + H(C_3 + C_4)\right)$ produces a $\widehat{\theta}_T$ such that, with probability at least $1-\delta$,*

$$c^\gamma(\widehat{\theta}_T) - \min_{\theta\in\Theta} c^\gamma(\theta) \le S\frac{\sqrt{d} + H(C_3 + C_4)}{\sqrt{T}} + \sqrt{\frac{1 + 4S^2T}{T^2}\left(2\log\frac{1}{\delta} + d\log\left(1 + \frac{S^2T}{d}\right)\right)}. \tag{4.64}$$

Proof. We (once again) prove the lemma by showing that conditions of Theorem 57 are satisfied. First, the subgradient norms have the easy bound

$$\|g_t^\gamma\| \le \|\ell^\intercal \Phi\| + H\frac{\|\Phi_{(x_t,a_t),:}\|}{q_3(x_t,a_t)} + H\frac{\|(P-\gamma B)^\intercal_{:,x'_t}\Phi\|}{q_4(x'_t)} \le \sqrt{d} + H(C_3+C_4)\ .$$

Finally, we show that the subgradient estimate is unbiased:

$$\begin{aligned}\mathbb{E}\left[g_t^\gamma(\theta)\right] &= \ell^\intercal\Phi - H\sum_{(x,a)} q_3(x,a)\frac{\Phi_{(x,a),:}}{q_3(x,a)}\mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\}} \\ &\qquad + H\sum_{x'} q_4(x')\frac{(P-\gamma B)^\intercal_{:,x'}\Phi}{q_4(x')}\operatorname{sgn}((P-\gamma B)^\intercal_{:,x'}\Phi\theta) \\ &= \ell^\intercal\Phi - H\sum_{(x,a)}\Phi_{(x,a),:}\mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\}} + H\sum_{x'}(P-\gamma B)^\intercal_{:,x'}\Phi\operatorname{sgn}((P-\gamma B)^\intercal_{:,x'}\Phi\theta) \\ &= \nabla_\theta c^\gamma(\theta)\ .\end{aligned}$$

□

With this lemma in hand, the proof of Theorem 63] proceeds in much the same way as the proof of Theorem 55].

Proof of Theorem 63. Recall that the convex surrogate for the discounted cost is

$$c^\gamma(\theta) = \ell^\intercal\Phi\theta + H\|\left[\Phi\theta\right]_-\|_1 + H\|(B-\gamma P)^\intercal\Phi\theta - \alpha\|_1, \tag{4.65}$$

with the constraint set $\Theta = \{\theta : \|\theta\|_2 \le S\}$.

Now, obtain $\widehat{\theta}_T$ from the stochastic gradient descent algorithm. By Lemma 65, the error bound must be less than

$$b_T = \frac{S}{\sqrt{T}}\left((\sqrt{d}+H(C_3+C_4)) + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1+\frac{S^2T}{d}\right)}\right) + O\left(\frac{1}{T}\right).$$

Then with high probability, we have for any $\theta\in\Theta$,

$$\ell^\intercal\Phi\widehat{\theta}_T + H\,V_3(\widehat{\theta}_T) + H\,V_4(\widehat{\theta}_T) \le \ell^\intercal\Phi\theta + H\,V_3(\theta) + H\,V_4(\theta) + b_T\ . \tag{4.66}$$

Since we can bound

$$\ell^\intercal\Phi\theta \le \|\ell\|_\infty\|\Phi\theta\|_1 \le \sqrt{d}\,CS,$$

rearranging Equation (4.66) yields

$$V_3(\widehat{\theta}_T) \le \frac{1}{H}\left(2\sqrt{d}\,CS + H\,V_3(\theta) + H\,V_4(\theta) + b_T\right) := \epsilon'\text{ , and} \tag{4.67}$$

$$V_4(\widehat{\theta}_T) \le \frac{1}{H}\left(2\sqrt{d}\,CS + H\,V_3(\theta) + H\,V_4(\theta) + b_T\right) := \epsilon''\ . \tag{4.68}$$

Using these bounds on $V_3(\widehat{\theta}_T)$ and $V_3(\widehat{\theta}_T)$ with Lemma 64 gives

$$|\ell^\intercal \nu_{\widehat{\theta}_T} - \ell^\intercal \Phi \widehat{\theta}_T| \leq \|\nu_{\widehat{\theta}_T} - \Phi \widehat{\theta}_T\|_1 \leq \frac{3\epsilon' + \epsilon''}{1-\gamma}. \tag{4.69}$$

Lemma 64, applied to ν_θ, implies that

$$|\ell^\intercal \nu_\theta - \ell^\intercal \Phi \theta| \leq \|\nu_\theta - \Phi\theta\|_1 \leq \frac{3V_3(\theta) + V_4(\theta)}{1-\gamma}, \tag{4.70}$$

and so

$$\begin{aligned}
\ell^\intercal \nu_{\widehat{\theta}_T} &\leq \ell^\intercal \Phi \widehat{\theta}_T + \frac{3\epsilon' + \epsilon''}{1-\gamma} \\
&\leq \ell^\intercal \Phi \theta + H\, V_3(\theta) + H\, V_4(\theta) + b_T + \frac{3\epsilon' + \epsilon''}{1-\gamma} \\
&\leq \ell^\intercal \nu_\theta + \frac{3V_3(\theta) + V_4(\theta)}{1-\gamma} + H\, V_3(\theta) + H\, V_4(\theta) + b_T + \frac{3\epsilon' + \epsilon''}{1-\gamma}\ .
\end{aligned}$$

First, we simplify

$$\begin{aligned}
\frac{3\epsilon' + \epsilon''}{1-\gamma} &= \frac{3}{H(1-\gamma)}\left(2\sqrt{d}CS + HV_3(\theta) + HV_4(\theta) + b_T\right) \\
&= \frac{3}{(1-\gamma)}\left(V_3(\theta) + V_4(\theta)\right) + \frac{3}{H(1-\gamma)} 2\sqrt{d}CS + \frac{4S(\sqrt{d} + C_3 + C_4)}{\sqrt{T}H(1-\gamma)} \\
&\quad + \frac{3S}{\sqrt{T}H(1-\gamma)} 2\sqrt{10 \log\frac{1}{\delta}} + \frac{3S}{\sqrt{T}H(1-\gamma)} 2\sqrt{5d \log\left(1 + \frac{S^2 T}{d}\right)} + O\left(\frac{1}{T^{3/2}(1-\gamma)H}\right) \\
&= \frac{3}{(1-\gamma)}\left(V_3(\theta) + V_4(\theta)\right) + \frac{6}{H(1-\gamma)}\sqrt{d}CS + O\left(\frac{\log(T)}{(1-\gamma)H\sqrt{T}}\right).
\end{aligned}$$

Plugging in this expression and b_T, we have

$$\begin{aligned}
\ell^\intercal \nu_{\widehat{\theta}_T} &\leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)\left(V_3(\theta) + V_4(\theta)\right) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + O\left(\frac{\log(T)}{(1-\gamma)H\sqrt{T}}\right) + b_T \\
&\leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)\left(V_3(\theta) + V_4(\theta)\right) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + \frac{S}{\sqrt{T}} H(C_3 + C_4) \\
&\quad + \frac{S}{\sqrt{T}}\left(C_3 + C_4 + \sqrt{d} + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1 + \frac{S^2T}{d}\right)}\right) \qquad (4.71) \\
&\quad + O\left(\frac{\log(T)}{(1-\gamma)H\sqrt{T}}\right) + O\left(\frac{1}{T}\right). \qquad (4.72)
\end{aligned}$$

Thus, setting T such that

$$T \geq \frac{S^2}{\epsilon^2}\left(C_3 + C_4 + \sqrt{d} + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1+\frac{S^2T}{d}\right)}\right)^2 \tag{4.73}$$
$$\Rightarrow T = O\left(\frac{S^2\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$$

yields

$$\ell^\intercal \nu_{\widehat{\theta}_T} \leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)(V_3(\theta) + V_4(\theta)) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + 2\epsilon + O\left(\frac{\epsilon}{H}\right) + O(\epsilon^2), \tag{4.74}$$

where, as usual, the O hides log factors. This statement holds with probability at least $1-\delta$ and for any $\theta \in \Theta$.

□

Error Bound

Previous ADP literature concentrated on showing that the optimal value function is well approximated if the feature space contains elements close to the optimum; i.e. $|\alpha^\intercal J_{\widehat{\theta}_T} - \alpha^\intercal J^*|$ was bounded in terms of $\min_\theta \|\Phi\theta - \nu^*\|_1$. Theorem 63 is certainly more general, as it remains non-trivial even if $\min_\theta \|\Phi\theta - \nu^*\|_1$ is large. For completeness, we provide a corollary of this form.

Corollary 66. *Under the same conditions as Theorem 63,*

$$\alpha^\intercal J_{\widehat{\theta}_T} - \alpha^\intercal J^* \leq C_3\left(\frac{1}{1-\gamma} + \frac{1}{\epsilon}\right)\min_\theta \|\Phi\theta - \nu^*\|_1 + C_2\frac{\epsilon}{1-\gamma}. \tag{4.75}$$

Proof. Let θ^* be one of the vectors minimizing $\|\Phi\theta - \nu^*\|_1$. Theorem 63 gives

$$\alpha^\intercal J_{\theta_T} - \alpha^\intercal J_{\theta^*} \leq C_1\left(\frac{1}{1-\gamma} + \frac{1}{\epsilon}\right)(V_3(\theta^*) + V_4(\theta^*)) + C_2\frac{\epsilon}{1-\gamma},$$

Since $\nu^* \geq 0$ and by the simple fact that $[x]_- \leq |y-x|$ for any $y \geq 0$, we have

$$V_3(\theta^*) \leq \|\Phi\theta - \nu^*\|_1. \tag{4.76}$$

For the term $V_4(\theta^*)$, since ν^* is feasible (i.e., $(B-\gamma P)^\intercal \nu^* = \alpha$)

$$\begin{aligned} V_4(\theta^*) &\leq \|(B-\gamma P)^\intercal(\Phi\theta^* - \nu^*)\|_1 + \|(B-\gamma P)^\intercal \nu^* - \alpha\|_1 = \|(B-\gamma P)^\intercal(\Phi\theta^* - \nu^*)\|_1 \\ &\leq \|(B-\gamma P)^\intercal\|_1 \|\Phi\theta - \nu^*\|_1 \leq (\|B^\intercal\|_1 + \gamma\|P^\intercal\|_1)\|\Phi\theta - \nu^*\|_1 \\ &= (1+\gamma)\|\Phi\theta - \nu^*\|_1, \end{aligned} \tag{4.77}$$

where $\|\cdot\|_1$ is the matrix operator 1-norm. Therefore, we have,

$$\alpha^\intercal J_{\pi_{[\Phi\widehat{\theta}_T]_+}} - \alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} \leq C_1\left(\frac{1}{1-\gamma}+\frac{1}{\epsilon}\right)(2+\gamma)\|\Phi\theta^*-\nu^*\|_1 + C_2\frac{\epsilon}{1-\gamma}. \tag{4.78}$$

Next, we bound $\alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} - \alpha^\intercal J^*$. Since $\alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} = \ell^\intercal \nu_{\pi_{[\Phi\theta^*]_+}}$ and $\alpha^\intercal J^* = \ell^\intercal \nu^*$ and by Lemma 64,

$$\begin{aligned}\alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} - \alpha^\intercal J^* &\leq \|\ell\|_\infty \|\nu_{\pi_{[\Phi\theta^*]_+}} - \nu^*\|_1 \leq \|\nu_{\pi_{[\Phi\theta^*]_+}} - \Phi\theta^*\|_1 + \|\Phi\theta^*-\nu^*\|_1 \\ &\leq \frac{3V_3(\theta^*)+V_4(\theta^*)}{1-\gamma} + \|\Phi\theta^*-\nu^*\|_1 \leq \frac{5}{1-\gamma}\|\Phi\theta^*-\nu^*\|_1. \end{aligned}\tag{4.79}$$

where the last inequality is due to (4.76) and (4.77).

The result follows by combining (4.78) and (4.79). □

4.5 Discounted Cost Grid Algorithm

For the average cost case, we saw that setting H correctly (or optimizing for it) could allow us to recover a regret bound that scaled with the square root of the optimal violation function. Theorem 63 showed that, with high probability,

$$\ell^\intercal \nu_{\widehat{\theta}_T} \leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma}+H\right)(V_3(\theta)+V_4(\theta)) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + O(\epsilon)$$

as long as T was sufficiently large.

This suggests that we want H and θ to optimize

$$\ell^\intercal\Phi\theta + \left(\frac{6}{1-\gamma}+H\right)(V_3(\theta)+V_4(\theta)) + \frac{\beta}{H},$$

where we have defined $\beta := \frac{6\sqrt{d}CS}{(1-\gamma)}$.

As before, we will approximately minimize the regret bound over H by using a grid algorithm.

It is important to note that the optimum H^* can never be larger than $1/\sqrt{V_{\max}}$, where $V_{\max}$ is some bound on $V_3(\theta^*)+V_4(\theta^*)$. Even though we cannot compute this quantity, we may still restrict the domain of H to

$$H \geq \min_\theta 1/\sqrt{V_3(\theta)+V_4(\theta)} \geq \left(1+\sqrt{d}CS(2+\gamma)\right)^{-\frac{1}{2}} \geq \left(4\sqrt{d}CS\right)^{-\frac{1}{2}}. \tag{4.80}$$

where the bound on $V_3(\theta)+V_4(\theta)$ is taken from (4.54).

For convenience, we will reuse some notation from the average cost section. Define

$$c^\gamma(H,\theta) := \ell^\intercal\Phi\theta + \left(H+\frac{6}{1-\gamma}\right)(V_3(\theta)+V_4(\theta)),$$

$\theta^*_H := \operatorname{argmin}_\theta c(H, \theta)$, and $F(H) = c(H, \theta^*_H) + \frac{\beta}{H}$. The *grid algorithm for the discounted case* takes as inputs a bound on the violation function $V_{\max}$, discount factor γ, an error tolerance ϵ, and desired probability tolerance δ. The algorithm then carefully chooses a grid $H_1, \dots, H_K$, and for each $i = 1, \dots, K$, computes $\hat{\theta}_i$, the output of Algorithm 4.1, and $\widehat{V}_i$, an approximation to $V_3(\hat{\theta}_i) + V_4(\hat{\theta}_i)$. It then returns

$$\hat{k} := \operatorname*{argmin}_k \ell^\intercal \Phi \widehat{\theta}_k + \left(H_k + \frac{1}{1-\gamma} \right) \widehat{V}_k + \frac{\beta}{H_k}. \tag{4.81}$$

Estimating the Violation Functions

Given some θ, we can estimate the violation function $V_3(\theta) + V_4(\theta)$ in much the same way as the average cost case.

For some n and samples $y_1, \dots, y_n \sim q_3$ and $(x_1, a_1), \dots, (x_n, a_n) \sim q_4$, define

$$\widehat{V}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \frac{[\Phi_{(x_i,a_i),:}\theta]_-}{q_3(x,a)} + \frac{|(B - \gamma P)^\intercal_{:,y_i} \Phi\theta - \alpha|}{q_4(y_i)}. \tag{4.82}$$

Since $V_1(\theta) = \sum_{(x,a)} |[\Phi_{(x,a),:}\theta]_-|$ and $V_2(\theta) = \sum_{x'} |(B - \gamma P)^\intercal_{:,x'} \Phi\theta - \alpha|$, this estimate is clearly unbiased. Also, we earlier assumed the existence of constants

$$C_3 = \max_{(x,a) \in \mathcal{X} \times \mathcal{A}} \frac{\|\Phi_{(x,a),:}\|}{q_3(x,a)}, \qquad C_4 = \max_{x \in \mathcal{X}} \frac{\|(P - \gamma B)^\intercal_{:,x} \Phi\|}{q_4(x)},$$

and so we can bound

$$\frac{[\Phi_{(x_i,a_i),:}\theta]_-}{q_3(x,a)} + \frac{|(B - \gamma P)^\intercal_{:,y_i} \Phi\theta - \alpha|}{q_4(y_i)} \le S(C_3 + 2C_4).$$

Therefore, we have concentration of $\widehat{V}$ around V. In particular, applying Hoeffding's inequality yields the following lemma.

Lemma 67. *Given $\epsilon > 0$ and $\delta \in [0, 1]$, for any θ, the violation function estimate $\widehat{V}_n(\theta)$ has*

$$|\widehat{V}_n(\theta) - (V_3(\theta) + V_4(\theta))| \le \epsilon$$

with probability at least $1 - \delta$ as long as we choose $n \ge \frac{(S(C_1 + 2C_4))^2}{2\epsilon^2} \log\left(\frac{2}{\delta}\right)$.

Figure 4.3 provides a precise definition of the algorithm and specifies the grid, parameters for the SGD algorithm, and sample sizes for $\widehat{V}_k$ needed. This algorithm has the following regret bound.

Theorem 68. *For some $\epsilon > 0$ and $\delta \in [0, 1]$, the Grid Algorithm for discounted cost specified in Figure 4.3 has regret*

$$\ell^\intercal \nu_{\theta_{\hat{k}}} \le \min_\theta \ell^\intercal \nu_\theta + O\left(\sqrt{V_1(\theta) + V_2(\theta)}\right) + O(\epsilon),$$

with probability at least $1 - \delta$.

```
**Input:** Upper bound $V_3(\theta) + V_4(\theta) \leq V_{\max}$, error tolerance $\epsilon > 0$, probability $\delta > 0$
Compute grid
    $H_k := \begin{cases} \beta \left( \frac{1}{\sqrt{V_{\max}}} - \frac{k-1}{4}\epsilon \right)^{-1} & \text{if } k \leq \overline{k} \\ H_{\overline{k}} + \frac{\epsilon(k-\overline{k})}{4V_{\max}} & \text{otherwise} \end{cases} \tag{4.83}$
where
    $\overline{k} := \min \left\{ i : V_{\max}(H_{i+1} - H_i) \geq \frac{\epsilon}{4} \right\}, \quad \text{and} \tag{4.84}$
    $K := \min\{k : H_k \geq 4\frac{\beta}{\epsilon\sqrt{1-\gamma}}\}. \tag{4.85}$

**for** $k := 1, 2, \ldots, K$ **do**
    Obtain $\widehat{\theta}_k$ from Algorithm 4.1 with $T = O\left(\frac{S^2 \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$
    Set $n \geq \frac{(S(C_1+2C_4))^2}{2\epsilon^2} \log\left(\frac{4}{\delta}\right)$
    Set $\widehat{V}_k = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{[\Phi_{(x_i,a_i),:}\hat{\theta}_k]_-}{q_3(x,a)} + \frac{|(B-\gamma P)^\intercal_{:,y_i}\Phi\hat{\theta}_k - \alpha|}{q_4(y_i)}\right]$
**end for**
Set $k^* = \text{argmin}_k\, \ell^\intercal \Phi\hat{\theta}_k + H_k\widehat{V}_k + \frac{\beta}{H_k(1-\gamma)}$
Return policy $\pi_{\widehat{\theta}_{\hat{k}}}$
```

Figure 4.3: The Grid algorithm for discounted cost

Analysis

As before, let $\epsilon > 0$ be some desired error tolerance and $V_{\max}$ be some upper bound on $V_3(\theta)+V_4(\theta)$; e.g. we can always take $V_{\max} = 4\sqrt{d}CS$. Define the grid

$$H_k := \begin{cases} \beta \left(\frac{1}{\sqrt{V_{\max}}} - \frac{k-1}{2}\epsilon \right)^{-1} & \text{if } k \leq \overline{k} \\ H_{\overline{k}} + \frac{\epsilon(k-\overline{k})}{2V_{\max}} & \overline{k} < k \leq K \end{cases} \tag{4.86}$$

where

$$\overline{k} := \min \left\{ i : V_{\max}(H_{i+1} - H_i) \geq \frac{\epsilon}{2} \right\} \tag{4.87}$$

and $K = \min\{k : H_k \geq \frac{2\beta}{\epsilon}\}$.

Lemma 69. *Let $\epsilon > 0$ be some desired error tolerance. Let $V_{\max}$ be some upper bound on $V_1(\theta) + V_4(\theta)$; we can always take $V_{\max} = 3 + S(d+2)$. The sequence defined by*

$$H_k := \begin{cases} \beta\left(\frac{1}{\sqrt{V_{\max}}} - \frac{k-1}{2}\epsilon\right)^{-1} & \textit{if } k \leq \overline{k} \\ H_{\overline{k}} + \frac{\epsilon(k-\overline{k})}{2V_{\max}} & \overline{k} < k \leq K \end{cases} \tag{4.88}$$

for

$$\overline{k} := \min\left\{i : V_{\max}(H_{i+1} - H_i) \geq \frac{\epsilon}{2}\right\} \quad \textit{and} \quad K = \min\left\{k : H_k \geq \frac{2\beta}{\epsilon\sqrt{1-\gamma}}\right\} \tag{4.89}$$

guarantees that, for all $1 \leq k \leq K$,

$$\max_{H_k \leq H \leq H_{k+1}} |F(H_k) - F(H)| \leq \epsilon. \tag{4.90}$$

Proof. Since $c(H,\theta)$ is linear in H, $c(H,\theta^*_H)$ is concave in H. Also, using δ to denote some positive number, we can show that $c(H,\theta^*_H)$ is increasing

$$\begin{aligned} c(H,\theta^*_H) &= \min_\theta \ell^\intercal \Phi\theta + \left(H + \frac{1}{1-\gamma}\right)(V_3(\theta) + V_4(\theta)) \\ &\leq \min_\theta \ell^\intercal \Phi\theta + \left(H + \frac{1}{1-\gamma} + \delta\right)(V_3(\theta) + V_4(\theta)) \\ &= c(H+\delta, \theta^*_{H+\delta}) \end{aligned}$$

and sublinear

$$\begin{aligned} c(H+\delta,\theta^*_{H+\delta}) &= \min_\theta \ell^\intercal \Phi\theta + \left(H + \frac{1}{1-\gamma} + \delta\right)(V_3(\theta) + V_4(\theta)) \\ &\leq \ell^\intercal \Phi\theta^*_H + \left(H + \frac{1}{1-\gamma} + \delta\right)(V_3(\theta^*_H) + V_4(\theta^*_H)) \\ &= c(H,\theta^*_H) + \delta(V_3(\theta^*_H) + V_4(\theta^*_H)) \\ &\leq c(H,\theta^*_H) + \delta V_{\max}. \end{aligned}$$

Using the monotonicity property of $c(H,\theta^*_H)$, we have

$$\begin{aligned} \max_{H_i \leq H \leq H_{i+1}} |F(H_i) - F(H)| &\leq \max_{H_i \leq H \leq H_{i+1}} c(H_{i+1},\theta^*_{H_i+1}) - c(H_i,\theta^*_{H_i}) + \beta\left(\frac{1}{H_i} - \frac{1}{H_{i+1}}\right) \\ &\leq (H_{i+1} - H_i)V_{\max} + \beta\left(\frac{1}{H_i} - \frac{1}{H_{i+1}}\right). \end{aligned}$$

First, consider the case where $k \leq \overline{k}$. Then $H_k = \beta\left(\sqrt{V_{\max}} - \frac{k-1}{2}\epsilon\right)^{-1}$ so that $\left(\frac{1}{H_k} - \frac{1}{H_{k+1}}\right) = \frac{\epsilon}{2}$. By definition of $\overline{k}$, we also have that $V_{\max}(H_{i+1} - H_i) \leq \frac{\epsilon}{2}$. In the case where $k > \overline{k}$, we have

$\frac{\beta}{1-\gamma}\left(\frac{1}{H_k}-\frac{1}{H_{k+1}}\right) < (H_{k+1}-H_k)V_{\max}$ and $(H_{k+1}-H_k)V_{\max}=\frac{\epsilon}{2}$. In either case,

$$|F(H_{k+1})-F(H_k)| \le (H_{k+1}-H_k)V_{\max}+\beta\left(\frac{1}{H_k}-\frac{1}{H_{k+1}}\right) \le \frac{\epsilon}{2}+\frac{\epsilon}{2} \le \epsilon,$$

as desired. Finally, we check that if $H^* > H_K$, then $|F(H^*)-F(H_K)| \le \epsilon$. Since

$$F(H^*) = \min_\theta \ell^\intercal\Phi\theta + \left(H+\frac{1}{1-\gamma}\right)(V_3(\theta)+V_4(\theta))+\frac{\beta}{H},$$

we can solve for the optimal H as a function of θ as $H^*_\theta = \sqrt{\beta(V_3(\theta)+V_4(\theta))^{-1}}$, yielding

$$F(H^*) = \min_{H,\theta} \ell^\intercal\Phi\theta + \frac{1}{1-\gamma}(V_3(\theta)+V_4(\theta))+2\sqrt{\beta(V_3(\theta)+V_4(\theta))}.$$

Therefore, $H^* > H_K$ implies that $V_3(\theta^*)+V_4(\theta^*) \le \frac{\epsilon^2(1-\gamma)}{4\beta}$. Then

$$\begin{aligned}
F(H_K)-F(H^*) &= \min_{\theta'}\left(\ell^\intercal\Phi\theta' + \left(\frac{1}{1-\gamma}+\frac{2\beta}{\epsilon}\right)(V_3(\theta')+V_4(\theta'))+\frac{\epsilon}{2}\right) \\
&\quad - \left(\min_\theta \ell^\intercal\Phi\theta + \frac{1}{1-\gamma}(V_3(\theta)+V_4(\theta))+2\sqrt{\beta(V_3(\theta)+V_4(\theta))}\right) \\
&\le \max_\theta \ell^\intercal\Phi\theta + \frac{2\beta}{\epsilon(1-\gamma)}(V_3(\theta)+V_4(\theta))+\frac{\epsilon}{2} - \left(\ell^\intercal\Phi\theta + 2\sqrt{\beta(V_3(\theta)+V_4(\theta))}\right) \\
&\le \max_\theta \frac{2\beta}{\epsilon(1-\gamma)}(V_3(\theta)+V_4(\theta))+\frac{\epsilon}{2} \\
&\le \epsilon,
\end{aligned}$$

completing the last case. □

Proof of Theorem 63. Running Algorithm 4.1 for $H_1,\ldots,H_K$ with $4T$ set as in Theorem 63 produces a sequence $\widehat{\theta}_1,\ldots,\widehat{\theta}_K$ such that

$$c(H_k,\widehat{\theta}_K) \le c(H_k,\theta^*_K) + H_kV(\theta^*)+\frac{\beta}{H_k}+\frac{\epsilon}{4}$$

holds for all k simultaneously with probability at least $1-\frac{\delta}{2}$, which is easily argued by noting that the probability of error for any single k is δ/K and applying the union bound.

Lemma 67, along with our choice of

$$n \ge \frac{(S(C_1+2C_4))^2}{2\epsilon^2}\log\left(\frac{4K}{\delta}\right)$$

guarantees that $|V_1(\widehat{\theta}_k)+V_2(\widehat{\theta}_k)-\widehat{V}_k| \le \frac{\epsilon}{4}$ holds with probability at least $1-\frac{\delta}{2K}$, and hence the statement holds for all $\widehat{V}_k$ with probability at most $1-\frac{\delta}{2}$.

We now turn to bounding the suboptimality of the objective. Recalling that $\hat{k}$ is the minimizer of $\ell^\intercal \Phi \widehat{\theta}_k + H_k \widehat{V}_k + \frac{\beta}{H_k}$, and using k^* as the minimizer of $c(H_k, \theta^*_k) + \frac{\beta}{H_k}$, we have

$$\begin{aligned}
\ell^\intercal \Phi \widehat{\theta}_{\hat{k}} + H_{\hat{k}} \widehat{V}_{\hat{k}} + \frac{\beta}{H_{\hat{k}}} &= \min_k \ell^\intercal \Phi \widehat{\theta}_k + H_k \widehat{V}_k + \frac{\beta}{H_k} \\
&\le \ell^\intercal \Phi \widehat{\theta}_{k^*} + H_{k^*} \widehat{V}_{k^*} + \frac{\beta}{H_{k^*}} \\
&\le c(H_{k^*}, \widehat{\theta}_{k^*}) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{4} && \text{(Lemma 67)} \\
&\le c(H_{k^*}, \theta^*_{k^*}) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{2} && \text{(Theorem 63)} \\
&= \min_k c(H_k, \theta^*_k) + \frac{\beta}{H_k} + \frac{\epsilon}{2} \\
&\le \min_{H,\theta} c(H, \theta) + \frac{\beta}{H} + \epsilon && \text{(Lemma 69).}
\end{aligned}$$

The statement holds with probability at least $\frac{\delta}{2} + \frac{\delta}{2}$, where the first term is from estimating $\widehat{V}_k$ (Lemma 67) and the second term is from bounding the SGD error (Theorem 63). Hence, the Gird Algorithm minimizes the objective to within ϵ.

Next, we use Lemma 64 to bound the discrepency between $\Phi\theta$ and ν_θ. Therefore, we need to bound $V_3(\widehat{\theta}_{\hat{k}})$ and $V_4(\widehat{\theta}_{\hat{k}})$. Since all quantities are non-negative, this implies that $|\frac{\beta}{H_{\hat{k}}} - \frac{\beta}{H^*}| \le \epsilon$. Using the bounded suboptimality of $\widehat{\theta}_{\hat{k}}$ as an optimizer of $c(H_{\hat{k}}, \theta)$, we have

$$\begin{aligned}
\ell^\intercal \Phi \widehat{\theta}_{\hat{k}} + \left(\frac{1}{1-\gamma} + H_{\hat{k}}\right)\left(V_1(\widehat{\theta}_{\hat{k}}) + V_2(\widehat{\theta}_{\hat{k}})\right) &\le \ell^\intercal \Phi \theta^*_{\hat{k}} + \left(\frac{1}{1-\gamma} + H_{\hat{k}}\right)\left(V_1(\theta^*_{\hat{k}}) + V_2(\theta^*_{\hat{k}})\right) + \frac{\epsilon}{2} \\
&\le \ell^\intercal \Phi \theta^* + \left(\frac{1}{1-\gamma} + H^*\right)\left(V_1(\theta^*) + V_2(\theta^*)\right) + \epsilon \\
&= \ell^\intercal \Phi \theta^* + \frac{1}{1-\gamma}\left(V_1(\theta^*) + V_2(\theta^*)\right) \\
&\quad + \sqrt{V_1(\theta^*) + V_2(\theta^*)} + \epsilon.
\end{aligned}$$

Next, we crudely bound $\ell^\intercal \Phi \theta \le \sqrt{d} CS$ and use $\left(\frac{1}{1-\gamma} + H_{\hat{k}}\right)^{-1} \le \frac{1}{H_{\hat{k}}}$ to obtain

$$\begin{aligned}
V_1(\widehat{\theta}_{\hat{k}}) + V_2(\widehat{\theta}_{\hat{k}}) &\le \frac{1}{H_{\hat{k}}}\left(2\sqrt{d}CS + \sqrt{V_1(\theta^*) + V_2(\theta^*)} + \frac{1}{(1-\gamma)}\left(V_1(\theta^*) + V_2(\theta^*)\right) + \epsilon\right) \\
&\le \left(\frac{1}{H^*} + \beta\epsilon\right)\left(2\sqrt{d}CS + \sqrt{V_1(\theta^*) + V_2(\theta^*)} + \frac{1}{(1-\gamma)}\left(V_1(\theta^*) + V_2(\theta^*)\right) + \epsilon\right) \\
&\le 2\sqrt{d}CS\sqrt{V_1(\theta^*) + V_2(\theta^*)} + \left(V_1(\theta^*) + V_2(\theta^*)\right) + \frac{\left(V_1(\theta^*) + V_2(\theta^*)\right)^{\frac{3}{2}}}{(1-\gamma)} + O(\epsilon).
\end{aligned}$$

Then, applying Lemma 64, we have

$$
\begin{aligned}
\ell^\intercal \Phi \mu_{\theta_{\hat{k}}} &\leq \ell^\intercal \Phi \widehat{\theta}_{\hat{k}} + \frac{3}{1-\gamma}\left(2\sqrt{d}CS\sqrt{V_1(\theta^*)+V_2(\theta^*)} + (V_1(\theta^*)+V_2(\theta^*)) + \frac{(V_1(\theta^*)+V_2(\theta^*))^{\frac{3}{2}}}{(1-\gamma)}\right) \\
&\quad + O\left(\frac{\epsilon}{1-\gamma}\right) \\
&\leq \ell^\intercal \Phi \theta^* + \frac{3}{1-\gamma}\left(2\sqrt{d}CS\sqrt{V_1(\theta^*)+V_2(\theta^*)} + (V_1(\theta^*)+V_2(\theta^*)) + \frac{(V_1(\theta^*)+V_2(\theta^*))^{\frac{3}{2}}}{(1-\gamma)}\right) \\
&\quad + \left(\frac{1}{1-\gamma} + H^*\right)(V_1(\theta^*)+V_2(\theta^*)) + O\left(\frac{\epsilon}{1-\gamma}\right) \\
&\leq \ell^\intercal \nu_{\theta^*} + \frac{3}{1-\gamma}\left(2\sqrt{d}CS\sqrt{V_1(\theta^*)+V_2(\theta^*)} + (V_1(\theta^*)+V_2(\theta^*)) + \frac{(V_1(\theta^*)+V_2(\theta^*))^{\frac{3}{2}}}{(1-\gamma)}\right) \\
&\quad + \left(\frac{1}{1-\gamma} + H^*\right)(V_1(\theta^*)+V_2(\theta^*)) + \epsilon + \frac{3}{1-\gamma}(V_1(\theta^*)+V_2(\theta^*)) + O\left(\frac{\epsilon}{1-\gamma}\right) \\
&\leq \ell^\intercal \nu_{\theta^*} + \left(1 + \frac{3}{1-\gamma}2\sqrt{d}CS\right)\sqrt{V_1(\theta^*)+V_2(\theta^*)} + \frac{3}{1-\gamma}\frac{(V_1(\theta^*)+V_2(\theta^*))^{\frac{3}{2}}}{(1-\gamma)} \\
&\quad + \frac{7}{1-\gamma}(V_1(\theta^*)+V_2(\theta^*)) + O\left(\frac{\epsilon}{1-\gamma}\right).
\end{aligned}
$$

All in all, this simplifies to

$$
\ell^\intercal \nu_{\theta_{\hat{k}}} \leq \min_\theta \ell^\intercal \nu_\theta + O\left(\sqrt{V_1(\theta)+V_2(\theta)}\right) + O\left((V_1(\theta)+V_2(\theta))^{\frac{3}{2}}\right) + O\left(\frac{\epsilon}{1-\gamma}\right).
$$

Using our assumption that $(V_1(\theta)+V_2(\theta)) < 1$, we obtain the theorem statement. □

4.6 Experiments

In this section, we apply both algorithms to the four-dimensional discrete-time queuing network illustrated in Figure 4.4. This network has a relatively long history; see, e.g. [42] and more recently [19] (c.f. Section 6.2). There are four queues, $\mu_1, \dots, \mu_4$, each with state $0, \dots, B$. Since the cardinality of the state space is $X = (1+B)^4$, even a modest B results in huge state-spaces. For time t, let $X_t \in X$ be the state and $s_{i,t} \in \{0,1\}$, $i = 1,2,3,4$ denote whether queue i is being served. Server 1 only serves queue 1 or 4, server 2 only serves queue 2 or 3, and neither server can idle. Thus, $s_{1,t} + s_{4,t} = 1$ and $s_{2,t} + s_{3,t} = 1$. The dynamics are as follows. At each time t, the following random variables are sampled independently: $A_{1,t} \sim$ Bernoulli(a_1), $A_{3,t} \sim$ Bernoulli(a_3), and $D_{i,t} \sim$ Bernoulli$(d_i * s_{i,t})$ for $i = 1,2,3,4$. Using $e_1, \dots, e_4$ to denote the standard basis vectors,

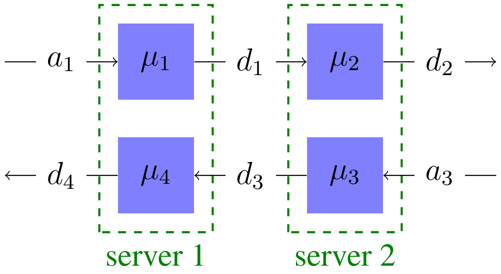

Figure 4.4: The 4D queuing network. Customers arrive at queue μ_1 or μ_3 then are referred to queue μ_2 or μ_4, respectively. Server 1 can either process queue 1 or 4, and server 2 can only process queue 2 or 3.

the dynamics are:

$$X'_{t+1} = X_t + A_{1,t}e_1 + A_{3,t}e_3 + D_{1,t}(e_2 - e_1) - D_{2,t}e_2 + D_{3,t}(e_4 - e_3) - D_{4,t}e_4,$$

and $X_{t+1} = \max(\mathbf{0}, \min(\mathbf{B}, X'_{t+1}))$ (i.e. all four states are thresholded from below by 0 and above by B). The loss function is the total queue size: $\ell(X_t) = ||X_t||_1$. We compared our method against two common heuristics. In the first, denoted LONGER, each server operates on the queue that is longer with ties broken uniformly at random (e.g. if queue 1 and 4 had the same size, they are equally likely to be served). In the second, denoted LBFS (last buffer first served), the downstream queues always have priority (server 1 will serve queue 4 unless it has length 0, and server 2 will serve queue 2 unless it has length 0). These heuristics are common and have been used as benchmarks for queuing networks (e.g. [19]).

We used $a_1 = a_3 = .08$, $d_1 = d_2 = .12$, and $d_3 = d_4 = .28$, and buffer sizes $B_1 = B_4 = 38$, $B_2 = B_3 = 25$ as the parameters of the network.. The asymmetric size was chosen because server 1 is the bottleneck and tend to have longer queues. The first two features are features of the stationary distributions corresponding to two heuristics. We also included two types of non-stationary-distribution features. For every interval $(0, 5], (6, 10], \ldots, (45, 50]$ and action A, we added a feature ψ with $\varphi(x, a) = 1$ if $\ell(x, a)$ is in the interval and $a = A$. To define the second type, consider the three intervals $I_1 = [0, 10]$, $I_2 = [11, 20]$, and $I_3 = [21, 25]$. For every 4-tuple of intervals $(J_1, J_2, J_3, J_4) \in \{I_1, I_2, I_3\}^4$ and action A, we created a feature ψ with $\psi(x, a) = 1$ only if $x_i \in J_i$ and $a = A$. Every feature was normalized to sum to 1. In total, we had 372 features which is about a 10^4 reduction in dimension from the original problem.

Stochastic Gradient Descent

We ran our stochastic gradient descent algorithm with $I = 1000$ sampled constraints and constraint gain $H = 2$. Our learning rate began at 10^{-4} and halved every 2000 iterations. The results of our algorithm are plotted in Figure 4.5, where $\widehat{\theta}_t$ denotes the running average of θ_t. The left plot is of the LP objective, $\ell^\top(\mu_0 + \Phi\widehat{\theta}_t)$. The middle plot is of the sum of the constraint violations, $\|[\mu_0 + \Phi\widehat{\theta}_t]_-\|_1 + \|(P - B)^\top\Phi\widehat{\theta}_t\|_1$. Thus, $c(\widehat{\theta}_t)$ is a scaled sum of the first two plots. Finally, the

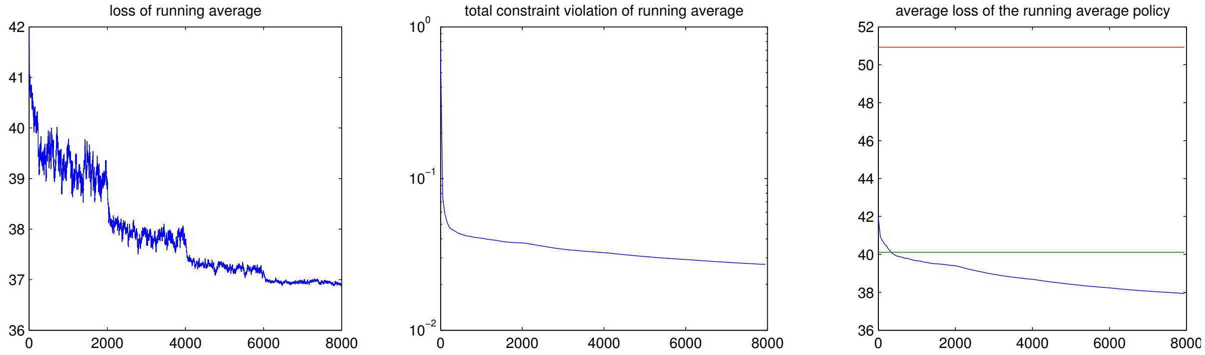

Figure 4.5: The left plot is of the linear objective of the running average, i.e. $\ell^\mathsf{T}\Phi\widehat{\theta}_t$. The center plot is the sum of the two constraint violations of $\widehat{\theta}_t$, and the right plot is $\ell^\mathsf{T}\tilde{\mu}_{\widehat{\theta}_t}$ (the average loss of the derived policy). The two horizontal lines correspond to the loss of the two heuristics, LONGER and LBFS.

right plot is of the average losses, $\ell^\mathsf{T}\mu_{\widehat{\theta}_t}$ and the two horizontal lines correspond to the loss of the two heuristics, LONGER and LBFS. The right plot demonstrates that, as predicted by our theory, minimizing the surrogate loss $c(\theta)$ does lead to lower average losses.

All previous algorithms (including [19]) work with value functions, while our algorithm works with stationary distributions. Due to this difference, we cannot use the same feature vectors to make a direct comparison. The solution that we find in this different approximating set is slightly worse than the solution of Farias and Van Roy [19].

4.7 Conclusion

This chapter demonstrated the feasibility of solving the MDP planning problem with a parametric policy class based on an approximate dual LP. Unlike previous approaches, we were able to prove *regret bounds*, that is, bounds relative to the best policy in our parametric class. We obtained results for both the average cost and discounted cost settings as well as empirical justification.

There are several promising directions. First, are such regret bounds possible in the primal formulation? The primal has the advantage that feature vectors are more intuitive to design.

Another drawback to our methods is that we need a backwards simulator, that is, access to every state with positive probability of transitioning into a state x. Are their alternative formulations that remove this requirement?

Chapter 5

Conclusion

5.1 Summary

This thesis examined two parallel approaches toward efficient sequential decision making. In the first, we examined minimax strategies in online learning for two different games. In the prediction game, for each round $t = 1, \ldots, T$, the learner played $\boldsymbol{a}_t$ in an attempt to predict the true label $\boldsymbol{x}_t \in \mathcal{X}$. We saw that the minimum enclosing ball for $\mathcal{X}$ is the central object in determining the minimax strategy and regret. If the minimum enclosing ball has center $\boldsymbol{c}$ and radius ρ, then the minimax strategy is

$$\boldsymbol{a}_n \;\coloneqq\; \boldsymbol{c} + \alpha_n \sum_{t=1}^{n-1} (\boldsymbol{x}_t - \boldsymbol{c}).$$

and the minimax regret is

$$\mathcal{R} \leq \rho^2 \sum_{t=1}^{T} \alpha_t.$$

This algorithm is efficient and scalable.

We also studied the online linear regression game, in both the fixed-design and adversarial-design settings. Under a variety of constraints on the learner, the minimax strategy for online linear regression is

$$\hat{y}_{t+1} \;=\; \boldsymbol{x}_{t+1}^{\intercal} \boldsymbol{P}_{t+1} \boldsymbol{s}_t$$

and the strategy does not need to know T in advance. In fact, the strategy is minimax against all other strategies, even those that know the game length in advance. We have also provided an intuitive view of the algorithm as follow-the-regularized-leader with a specific data-dependent regularizer which automatically adjusts to the scale of the data and to how much data budget remains.

Finally, we studied the MDP planning problem with a large state-space. We demonstrated that solving this problem with a parametric policy class based on an approximate dual LP allowed us to prove regret bounds between the loss of our policy and the best policy in the class; that is, we showed

$$\ell^{\intercal} \xi_{\widehat{\theta}} \leq \min_{\theta \in \Theta} \ell^{\intercal} \xi_{\theta} + V(\theta) + O(\epsilon),$$

where $\xi_{\widehat{\theta}}$ was the appropriate dual variable for either the average cost or the discounted cost case. We also saw empirical results for the average cost case on a 4-dimensional queuing example.

5.2 Future Directions

The minimax technique is powerful but often computationally unwieldy. The prediction and linear regression games relied on the quadratic structure of the losses and value-to-go to arrive at closed form expressions for the strategy and regret. One obvious direction is exploring other loss functions and constraint sets that allow for a precise minimax analysis. An orthogonal direction would be to extend the minimax tools to other scenarios with squared loss. Examples include estimation under changing dynamics (e.g. hidden Markov models) or control problems with linear dynamics and square loss (e.g. linear quadratic regulators).

Additionally, one could instead use the minimax analysis as a tool to obtain precise regret guarantees of more common regret-minimization algorithms. For example, can we analyze follow the regularized leader by comparing it to the follow the regularized leader algorithm with the data dependent regularization studied in Chapter 3?

There are also a few pressing questions for the large-scale MDP planning. Is a regret bound possible in the primal formulation? Such a result would allow one to import the literature on designing features in the primal, as it is less intuitive to design features in the dual. What is the impact of the feature matrix to the performance? Finally, the algorithm required knowledge of all states that transition into an arbitrary state x. In practice, having knowledge of the forward dynamics (through e.g. a forwards simulator) is more common that having knowledge of the backward dynamics, which is required by our method. Is it possible to remove this requirement?

Bibliography

[1] Y. Abbasi-Yadkori. “Online Learning for Linearly Parametrized Control Problems”. PhD thesis. University of Alberta, 2012.

[2] Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. “Optimal strategies and minimax lower bounds for online convex games”. In: *COLT*. Ed. by Rocco A. Servedio and Tong Zhang. Omnipress, 2008, pp. 415–423.

[3] Jacob Abernethy, Manfred K. Warmuth, and Joel Yellin. “When Random Play is Optimal Against an Adversary”. In: *COLT*. Ed. by Rocco A. Servedio and Tong Zhang. Omnipress, 2008, pp. 437–446.

[4] Katy S Azoury and Manfred K Warmuth. “Relative loss bounds for on-line density estimation with the exponential family of distributions”. In: *Machine Learning* 43.3 (2001), pp. 211–246.

[5] Peter L. Bartlett, Peter Grunwald, Peter Harremoes, Fares Hedayati, and Wojciech Kotlowski. “Horizon-Independent Optimal Prediction with Log-Loss in Exponential Families”. In: *CoRR* abs/1305.4324 (2013).

[6] Peter L. Bartlett, Wouter M. Koolen, Alan Malek, Manfred K. Warmuth, and Eiji Takimoto. “Minimax Fixed-design Linear Regression”. In: *Proceedings of The 28th Annual Conference on Learning Theory (COLT)*. Ed. by P. Grunwald, E. Hazan, and S. Kale. 2015, pp. 226–239.

[7] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[8] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2007.

[9] D. P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena scientific optimization and computation series. Athena Scientific, 1996.

[10] Glenn W Brier. “Verification of forecasts expressed in terms of probability”. In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[11] Nicolo Cesa-Bianchi, Philip M Long, and Manfred K Warmuth. “Worst-case quadratic loss bounds for prediction using linear functions and gradient descent”. In: *Neural Networks, IEEE Transactions on* 7.3 (1996), pp. 604–619.

[12] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[13] Nicolò Cesa-Bianchi and Ohad Shamir. "Efficient Online Learning via Randomized Rounding". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger. 2011, pp. 343–351.

[14] V. H. de la Peña, T. L. Lai, and Q-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

[15] V. V. Desai, V. F. Farias, and C. C. Moallemi. "Approximate Dynamic Programming via a Smoothed Linear Program". In: *Operations Research* 60.3 (2012), pp. 655–674.

[16] Daniela Pucci de Farias and Benjamin Van Roy. "A Cost-Shaping Linear Program for Average-Cost Approximate Dynamic Programming with Performance Guarantees". In: *Mathematics of Operations Research* 31 (2006).

[17] Daniela Pucci de Farias and Benjamin Van Roy. "Approximate Linear Programming for Average-Cost Dynamic Programming". In: *Advances in Neural Information Processing Systems (NIPS)*. 2003.

[18] Daniela Pucci de Farias and Benjamin Van Roy. "On Constraint Sampling in the Linear Programming Approach to Approximate Dynamic Programming". In: *Mathematics of Operations Research* 29 (2004).

[19] Daniela Pucci de Farias and Benjamin Van Roy. "The linear programming approach to approximate dynamic programming". In: *Operations Research* 51 (2003).

[20] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*. 2005.

[21] Jürgen Forster. "On relative loss bounds in generalized linear regression". In: *Fundamentals of Computation Theory*. Springer. 1999, pp. 269–280.

[22] Dean P. Foster. "Prediction in the worst case". In: *Annals of Statistics* 19.2 (1991), pp. 1084–1090.

[23] C. Guestrin, M. Hauskrecht, and B. Kveton. "Solving factored MDPs with continuous and discrete variables". In: *Twentieth Conf. Uncertainty in Artificial Intelligence*. 2004.

[24] David A. Harville. "MATRIX ALGEBRA FROM A STATISTICIAN'S PERSPECTIVE". In: (1997).

[25] M. Hauskrecht and B. Kveton. "Linear program approximations to factored continuous-state Markov decision processes". In: *Advances in Neural Information Processing Systems*. 2003.

[26] Fares Hedayati and Peter L. Bartlett. "Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior". In: *International Conference on Artificial Intelligence and Statistics*. 2012, pp. 504–510.

[27] R. A. Howard. *Dynamic Programming and Markov Processes*. MIT, 1960.

[28] Jyrki Kivinen and Manfred K Warmuth. "Exponentiated gradient versus gradient descent for linear predictors". In: *Information and Computation* 132.1 (1997), pp. 1–63.

[29] Petri Kontkanen and Petri Myllymäki. “A Fast Normalized Maximum Likelihood Algorithm for Multinomial Data”. In: *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*. 2005, pp. 1613–1616.

[30] Wouter M Koolen, Alan Malek, and Peter L Bartlett. “Efficient Minimax Strategies for Square Loss Games”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 3230–3238.

[31] Wouter M Koolen, Alan Malek, Peter L Bartlett, and Yasin Abbasi. “Minimax Time Series Prediction”. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett. Curran Associates, Inc., 2015, pp. 2548–2556. URL: `http://papers.nips.cc/paper/5730-minimax-time-series-prediction.pdf`.

[32] Wouter M. Koolen, Jiazhong Nie, and Manfred K. Warmuth. “Learning a set of directions”. In: *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*. Ed. by Shai Shalev-Shwartz and Ingo Steinwart. June 2013.

[33] H. R. Maei, Cs. Szepesvári, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton. “Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation”. In: *Advances in Neural Information Processing Systems*. 2009.

[34] H. R. Maei, Cs. Szepesvári, S. Bhatnagar, and R. S. Sutton. “Toward off-policy learning control with function approximation”. In: *Proceedings of the 27th International Conference on Machine Learning*. 2010.

[35] Alan Malek, Yasin Abbasi-Yadkori, and Peter Bartlett. “Linear Programming for Large-Scale Markov Decision Problems”. In: *Proceedings of The 31st International Conference on Machine Learning*. 2014, pp. 496–504.

[36] A. S. Manne. “Linear programming and sequential decisions”. In: *Management Science* 6.3 (1960), pp. 259–267.

[37] Brendan McMahan and Jacob Abernethy. “Minimax optimal algorithms for unconstrained linear optimization”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 2724–2732.

[38] Nimrod Megiddo. “Linear-time algorithms for linear programming in Rˆ3 and related problems”. In: *SIAM journal on computing* 12.4 (1983), pp. 759–776.

[39] Edward Moroshko and Koby Crammer. “Weighted last-step min–max algorithm with improved sub-logarithmic regret”. In: *Theoretical Computer Science* 558 (2014), pp. 107–124.

[40] M. Petrik and S. Zilberstein. “Constraint relaxation in approximate linear programs”. In: *Proc. 26th Internat. Conf. Machine Learning (ICML)*. 2009.

[41] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1st. New York, NY, USA: John Wiley & Sons, Inc., 1994.

[42] A. N. Rybko and A. L. Stolyar. “Ergodicity of stochastic processes describing the operation of open queueing networks”. In: *Problemy Peredachi Informatsii* 28.3 (1992), pp. 3–26.

[43] P. Schweitzer and A. Seidmann. “Generalized polynomial approximations in Markovian decision processes”. In: *Journal of Mathematical Analysis and Applications* 110 (1985), pp. 568–582.

[44] Rocco A. Servedio and Tong Zhang, eds. *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*. Omnipress, 2008.

[45] Yurii Mikhailovich Shtar’kov. “Universal sequential coding of single messages”. In: *Problemy Peredachi Informatsii* 23.3 (1987), pp. 3–17.

[46] Maurice Sion. “On general minimax theorems”. In: *Pacific J. Math.* 8.1 (1958), pp. 171–176.

[47] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. Bradford Book. MIT Press, 1998.

[48] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, Cs. Szepesvári, and E. Wiewiora. “Fast gradient-descent methods for temporal-difference learning with linear function approximation”. In: *Proceedings of the 26th International Conference on Machine Learning*. 2009.

[49] R. S. Sutton, Cs. Szepesvári, and H. R. Maei. “A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation”. In: *Advances in Neural Information Processing Systems*. 2009.

[50] Eiji Takimoto and Manfred K. Warmuth. “The minimax strategy for Gaussian density estimation”. In: *13th COLT*. 2000, pp. 100–106.

[51] M. H. Veatch. “Approximate Linear Programming for Average Cost MDPs”. In: *Mathematics of Operations Research* 38.3 (2013).

[52] Volodimir G Vovk. “Aggregating strategies”. In: *Proc. Third Workshop on Computational Learning Theory*. Morgan Kaufmann. 1990, pp. 371–383.

[53] Volodya Vovk. “Competitive on-line linear regression”. In: *Advances in Neural Information Processing Systems* (1998), pp. 364–370.

[54] T. Wang, D. Lizotte, M. Bowling, and D. Schuurmans. “Dual Representations for Dynamic Programming”. In: *Journal of Machine Learning Research* (2008), pp. 1–29.

[55] Lin Xiao and Tong Zhang. “A proximal stochastic gradient method with progressive variance reduction”. In: *SIAM Journal on Optimization* 24.4 (2014), pp. 2057–2075.

[56] M. Zinkevich. “Online Convex Programming and Generalized Infinitesimal Gradient Ascent”. In: *ICML*. 2003.