Optimization Foundations of Reinforcement Learning

Jalaj Bhandari

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Optimization Foundations of Reinforcement Learning

Jalaj Bhandari

Reinforcement learning (RL) has attracted rapidly increasing interest in the machine learning and artificial intelligence communities in the past decade. With tremendous success already demonstrated for Game AI, RL offers great potential for applications in more complex, real world domains, for example in robotics, autonomous driving and even drug discovery. Although researchers have devoted a lot of engineering effort to deploy RL methods at scale, many state-of-the art RL techniques still seem *mysterious* - with limited theoretical guarantees on their behaviour in practice.

In this thesis, we focus on understanding convergence guarantees for two key ideas in reinforcement learning, namely *Temporal difference learning* and *policy gradient methods*, from an optimization perspective. In Chapter 2, we provide a *simple and explicit* finite time analysis of Temporal difference (TD) learning with linear function approximation. Except for a few key insights, our analysis mirrors standard techniques for analyzing stochastic gradient descent algorithms, and therefore inherits the simplicity and elegance of that literature. Our convergence results extend seamlessly to the study of TD learning with eligibility traces, known as TD(λ), and to Q-learning for a class of high-dimensional optimal stopping problems.

In Chapter 3, we turn our attention to policy gradient methods and present a simple and general understanding of their global convergence properties. The main challenge here is that even for simple control problems, policy gradient algorithms face non-convex optimization problems and

are widely understood to converge only to a stationary point of the objective. We identify structural properties – shared by finite MDPs and several classic control problems – which guarantee that despite non-convexity, any stationary point of the policy gradient objective is globally optimal. In the final chapter, we extend our analysis for finite MDPs to show linear convergence guarantees for many popular variants of policy gradient methods like projected policy gradient, Frank-Wolfe, mirror descent and natural policy gradients.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to extend my heartfelt gratitude to my advisor, Daniel Russo for sharing his exciting research ideas, constant encouragement and his willingness to help at any time. He has taught me to identify and work on the most impactful ideas, unblocked me when I was stuck, been kind when I made errors and has always insisted on simple but clear writing which has often helped me clarify my own thinking. I am incredibly lucky to have had the opportunity to work closely with him. I would also like to thank Garud Iyengar for his mentorship throughout my time at Columbia. He has always supported and guided me on acedemic and non-academic issues, recommended me for internships as well as pointed me toward exciting courses and research projects. While we have never published a paper together, I have truly enjoyed all of our research discussions on a couple of projects and one day we will definitely publish a paper together.

I would also like to thank Shipra Agrawal, Christian Kroer and Krzysztof Choromanski for serving on my thesis committee. I really enjoyed working closely with Shipra on a joint project with Amazon and taking her graduate class on Reinforcement learning. I interacted with Krzysztof for the first time at AISTATS, 2017. I thank him for fruitful discussions on reseach ideas and hope to pursue research together in the future sometime. I also owe a great deal to Profs. John Cunningham and David Blei for teaching excellent graduate level courses on machine learning and including me in their weekely group meethings which got me excited about research in statstical machine learning. John and I have also co-authored two papers which were led by Francois Fagan, and are not a part of this thesis. I learnt a lot from this colloboration and research experience.

I also wish to acknowledge constant support from the IEOR department which has been par-

ticuarly patient with me, specially during my on-going health issues (thanks to Garud, Vineet and Jay). Along with that, I want to thank my physios, Dr. Courtney Burdowski and Dr. Lea Noonan for helping me recover which allowed me to complete this work. My fellow students in IEOR have all been very helpful and supportive. A special thanks to Itai Feigenbaum and Praveen for their words of encouragement during initial years of my PhD. I would also like to thank my undergraduate professors, Nomesh Bolia and Kiran Seth at IIT Delhi as well as Sandeep Juneja at TIFR, for sparking my interest in scientific research. I stayed at International House, NYC from 2014-17 and met some of the most fun people who showered me with kindness and affection. Without naming everyone, I wish to thank them all, Emad Khan and Ms Lorraine Pirro in particular.

I credit this thesis, and whatever little I have accomplished in life to my parents, people name above and everyone else who helped me so far. The numerous errors, mistakes, and failures are all attributable to me and I'll hold onto them.

To my parents, Naresh and Anita, for their unconditional love and numerous sacrifices.

Chapter 1: Introduction

Reinforcement learning (RL) is a computational approach to automating sequential decision making where an agent learns by trial-and-error interactions with a complex, uncertain environment [1]. It uses the framework of Markov decision processes to define this interaction in terms of states of the environment, actions and reward signals. The agent's goal is to learn a decision rule (actions to take in a particular state) to maximize *long run sum of the reward signal*; as the action taken in the present state may not only affect the immediate reward but also change the next state. In the past decade, RL has attracted rapidly increasing interest in the machine learning and artificial intelligence communities where a primary goal is to come up with fully autonomous agents that are interactive and can continuously learn from past experiences.

With tremendous success already demonstrated on game environments like Alpha Go [2], RL algorithms offer great promise for applications to even more complex domains, for example in autonomous driving [3], robotics [4] and even drug discovery [5]. There is tremendous potential, as modeling the environments (to even reasonable approximations) in many modern applications is extremely challenging. Although researchers have devoted a lot of engineering effort to deploy RL methods at scale, many state-of-the-art RL techniques still seem *mysterious* – with limited theoretical guarantees on their behavior in practice. This poses a significant challenge in convincing practitioners to use these for real world tasks. The motivation of this thesis is to understand theoretical underpinnings of some of the key ideas in RL using tools from optimization theory. Doing this can help spur progress in principled approaches to algorithm design to tackle some key challenges, for example in statistical efficiency and robustness, in applications where collecting data is difficult and expensive.

We broadly focus on two different classes of RL techniques, namely, *Temporal difference learning* and *policy gradient methods*, both of which are widely adopted by practitioners and popular

among RL researchers. I considered a larger introductory chapter, but felt that it would be too repetitive in terms of context to be useful. Each chapter in this thesis is self-contained with an introduction to the problem, its motivation in context of the literature and our contribution. Therefore, we just give a brief summary of the main results below. The rest of this thesis is organized as follows.

- In Chapter 2, we consider the Temporal difference learning (TD) algorithm, which is a simple iterative method used to estimate the value function corresponding to a given policy in a Markov decision process. Although widely used in reinforcement learning, theoretical analysis of TD has proved challenging and few guarantees on its statistical efficiency are available. We provide a *simple and explicit finite time analysis* of temporal difference learning with linear function approximation. Even though TD updates are *not* stochastic gradient updates with respect to any fixed loss function, our analysis uncovers key insights which enable us to analyze TD mirroring standard techniques used for analyzing stochastic gradient descent algorithms. We also show how all of our convergence results extend seamlessly to the study of TD learning with eligibility traces, known as TD(λ), and to Q-learning for a class of high-dimensional optimal stopping problems.

- In Chapter 3, we consider policy gradients methods, which directly search for an optimal policy by performing stochastic gradient descent over a parameterized class of polices. Recently, these approaches have shown tremendous success with deep neural networks and Monte Carlo approximations to the true gradient. Unfortunately, even for simple control problems solvable by classical techniques, policy gradient algorithms face non-convex optimization problems and are widely understood to converge only to a local minima, assuming adequate smoothness properties. We present a simple and general understanding of global convergence properties of policy gradient methods. We focus on studying the optimization landscape and identify structural properties which guarantee despite non-convexity, any stationary point of the policy gradient objective is globally optimal. We then identify conditions under which the policy gradient objective is *gradient dominated*. For many first-order opti-

mization methods, this gradient dominance conditions guarantees fast convergence rates to globally optimal solutions for non-convex objectives. Our results apply to several classic dynamic programming problems including finite MDPs, linear quadratic control, optimal stopping and finite horizon inventory control, which provide an important benchmark for studying theoretical properties of model free reinforcement learning algorithms.

- In Chapter 4, we extend our analysis of policy gradient methods for finite MDPs. We take a different perspective than that of Chapter 3, where we view this (and other problems) as an instance of smooth non-linear optimization and using gradient dominance, show sub-linear convergence with small stepsizes. Our insights in this chapter show how for tabular MDPs, different policy gradient algorithms with appropriately large step-sizes are in fact equivalent to a *soft-policy iteration* update and therefore enjoy linear convergence guarantees. Our analysis covers popular methods like projected policy gradient, Frank-Wolfe, mirror descent and natural policy gradient methods which are widely used in practice.

Chapter 2: A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation

Originally proposed by [6], temporal difference learning (TD) is one of the most widely used reinforcement learning algorithms and a foundational idea on which more complex methods are built. The algorithm operates on a stream of data generated by applying some policy to a poorly understood Markov decision process. The goal is to learn an approximate value function, which can then be used to track the net present value of future rewards as a function of the system's evolving state. TD maintains a parametric approximation to the value function, making a simple incremental update to the estimated parameter vector each time a state transition occurs.

While easy to implement, theoretical analysis of TD is subtle. Reinforcement learning researchers in the 1990s gathered both limited convergence guarantees [7] and examples of divergence [8]. Many issues were then clarified in the work of [9], which establishes precise conditions for the asymptotic convergence of TD with linear function approximation and gives examples of divergent behavior when key conditions are violated. With guarantees of asymptotic convergence in place, a natural next step is to understand the algorithm's statistical efficiency. How much data is required to guarantee a given level of accuracy? Can one give uniform bounds on this, or could data requirements explode depending on the problem instance? Twenty years after the work of [9], such questions remain largely unsettled.

2.1 Contributions

This chapter develops a *simple and explicit non-asymptotic analysis of TD with linear function approximation*. The resulting guarantees provide assurances of robustness. They explicitly bound the worst-case dependence on problem features like the discount factor, the conditioning of the

feature covariance matrix, and the mixing time of the underlying Markov chain. Our analysis reveals rigorous connections between TD and stochastic gradient descent algorithms, provides a template for finite time analysis of incremental algorithms with Markovian noise, and applies without modification to analyzing a class of high-dimensional optimal stopping problems. We elaborate on these contributions below.

- *Links with gradient descent:* Despite a cosmetic connection to stochastic gradient descent (SGD), incremental updates of TD are not (stochastic) gradient steps with respect to any fixed loss function. It is therefore difficult to show that it makes consistent, quantifiable, progress toward its asymptotic limit point. Nevertheless, Section 2.7 shows that expected TD updates obey crucial properties mirroring those of gradient descent on a particular quadratic loss function. In a model where the observations are corrupted by i.i.d. noise, these gradient-like properties of TD allow us to give state-of-the-art convergence bounds by essentially mirroring standard analyses of SGD. This approach may be of broader interest as SGD analyses are commonly taught in machine learning courses and serve as a launching point for a much broader literature on first-order optimization. Rigorous connections with the optimization literature can facilitate research on principled improvements to TD.

- *Non-asymptotic treatment with Markovian noise:* TD is usually applied online to a single Markovian data stream. However, to our knowledge, there has been no successful[1] non-asymptotic analysis in the setting with Markovian observation noise. Instead, many papers have studied such algorithms under the simpler i.i.d. noise model mentioned earlier [12, 13, 14, 15, 16, 17]. One reason is that the dependent nature of the data introduces a substantial technical challenge: the algorithm's updates are not only noisy, but can be severely biased. We use information theoretic techniques to control the magnitude of bias, yielding bounds that are essentially scaled by a factor of the mixing time of the underlying Markov process relative to those attained for the i.i.d. model. Our analysis in this setting applies only to a

[1]This was previously attempted by [10], but critical errors were shown by [11].

variant of TD that projects the iterates onto a norm ball. This projection step imposes a uniform bound on the noise of TD updates, which is needed for tractability. For similar reasons, projection operators are widely used throughout the stochastic approximation literature [18, Section 2].

- *An extendable approach:* Much of the paper focuses on analyzing the most basic temporal difference learning algorithm, known as TD(0). We also extend this analysis to other algorithms. First, we establish convergence bounds for temporal difference learning with eligibility traces, known as TD(λ). This is known to often outperform TD(0) [19], but a finite time analysis is more involved. Our analysis also applies without modification to Q-learning for a class of high-dimensional optimal stopping problems. Such problems have been widely studied due to applications in the pricing of financial derivatives [20, 21, 22, 23, 24]. For our purposes, this example illustrates more clearly the link between value prediction and decision-making. It also shows our techniques extend seamlessly to analyzing an instance of non-linear stochastic approximation. To our knowledge, no prior work has provided non-asymptotic guarantees for either TD(λ) or Q-learning with function approximation.

2.2 Related Literature

Non-asymptotic analysis of TD(0): There has been very little non-asymptotic analysis of TD(0). To our knowledge, [10] provided the first finite time analysis. However, several serious errors in their proofs were pointed out by [11]. A very recent work by [25] studies TD(0) with linear function approximation in an i.i.d. observation model, which assumes sequential observations used by the algorithm are drawn independently from their steady-state distribution. They focus on analysis with problem independent step-sizes of the form $1/T^\sigma$ for a fixed $\sigma \in (0, 1)$ and establish that mean-squared error converges at a rate[2] of $O(1/T^\sigma)$. Unfortunately, while the analysis is technically non-asymptotic, the constant factors in the bound display a complex dependence on

[2]In personal communication, the authors have told us their analysis also yields a $O(1/T)$ rate of convergence for problem dependent step-sizes, though we have not been able to easily verify this.

the problem instance and scale with some unusual quantities which can be very large in cases of practical interest.

Another interesting paper by [17] studies linear stochastic approximation algorithms under i.i.d. noise, including TD(0), with constant step-sizes and iterate averaging. This approach dates back to the works of [26, 27] and [28], which shows that the iterates of a constant step-size linear stochastic approximation algorithm form an ergodic Markov chain and, *in the case of i.i.d. observation noise*, their expectation in steady-state is equal to the true solution of the linear system. By a central limit theorem for ergodic sequences, the average iterate converges to the true solution, with mean-squared error decaying at rate $O(1/T)$. [29] give a sophisticated non-asymptotic analysis of the least-mean-squares algorithm with constant step-size and iterate-averaging. [17] aim to understand whether such guarantees extend to linear stochastic approximation algorithms more broadly. In the process, their work provides $O(1/T)$ bounds for iterate-averaged TD(0) with constant step-size. A remarkable feature of their approach is that the choice of step-size is independent of the conditioning of the features (although the bounds themselves do degrade if features become ill-conditioned). It is worth noting that these results rely critically on the assumption that noise is i.i.d. This is not due to any shortcoming in the techniques of [29] and [17]. Instead, under non-i.i.d. noise and a linear stochastic approximation algorithm applied with any constant step-size, the averaged-iterate might converge to the wrong limit as shown in a simple example by [28].

The recent works of [25] and [17] give bounds for TD(0) only under i.i.d. observation noise. Therefore their results are most comparable to what is presented in Section 2.8. For the i.i.d. noise model, the main argument in favor of our approach is that it allows for extremely simple proofs, interpretable constant terms, and illuminating connections with SGD. Moreover, it is worth emphasizing that our approach gracefully extends to more complex settings, including more realistic models with Markovian noise, the analysis of TD with eligibility traces, and the analysis of Q-learning for optimal stopping problems as shown in Sections 2.9, 2.10 and 2.11.

While not directly comparable to our results, we point the readers to the excellent work of [30]. To facilitate theoretical analysis, they consider a slightly modified version of the TD(λ) algorithm.

The authors provide a finite time analysis for this algorithm in an adversarial model where the goal is to predict the discounted sum of future rewards from each state. Performance is measured relative to the best fixed linear predictor in hindsight. The analysis is creative, but results depend on a several unknown constants and on the specific sequence of states and rewards on which the algorithm is applied. [30] also apply their techniques to study value function approximation in a Markov decision process. In that case, the bounds are much weaker than what is established here. Their bound scales with the size of the state space, which is enormous in most practical problems – and applies only to TD(1), a somewhat degenerate special case of TD(λ) in which it is equivalent to Monte Carlo policy evaluation [19].

Asymptotic analysis of stochastic approximation: There is a well developed theory around asymptotic analysis of stochastic approximation, a field that studies noisy recursive algorithms like TD [31, 32, 33]. Most asymptotic convergence proofs in reinforcement learning use a technique known as the ODE method [34]. Under some technical conditions and appropriate decaying step-sizes, this method ensures the almost-sure convergence of stochastic approximation algorithms to the invariant set of a certain 'mean' differential equation. The technique greatly simplifies asymptotic convergence arguments, since it completely circumvents issues with noise in the system and issues of step-size selection. But this also makes it a somewhat coarse tool, unable to generate insight into an algorithm's sensitivity to noise, ill-conditioning, or step-size choices. A more refined set of techniques begin to address these issues. Under fairly broad conditions, a central limit theorem for stochastic approximation algorithms characterizes their limiting variance. Such a central limit theorem has been specifically provided for TD by [35] and [36].

In addition to such asymptotic techniques, the modern literature on first-order stochastic optimization also focuses heavily on non-asymptotic analysis [37, 38, 39]. One reason is that such asymptotic analysis necessarily focuses on a regime where step-sizes are negligibly small relative to problem features and the iterates have already converged to a small neighborhood of the optimum. However, the use of a first-order method in the first place signals that a practitioner is mostly

interested in cheaply reaching a reasonably accurate solution, rather than the rate of convergence in the neighborhood of the optimum. In practice, it is common to use constant step-sizes, so iterates never truly converge to the optimum. A non-asymptotic analysis requires grappling with the algorithm's behavior in practically relevant regimes where step-sizes are still relatively large and iterates are not yet close to the true solution.

Analysis of related algorithms: A number of papers analyze algorithms related to and inspired by the classic TD algorithm. First, among others, [40, 41, 42, 43, 44] and [45] analyze least-squares temporal difference learning (LSTD). [46] study the related least-squares policy iteration algorithm. The asymptotic limit point of TD is a minimizer of a certain population loss, known as the mean-squared projected Bellman error. LSTD solves a least-squares problem, essentially computing the exact minimizer of this loss on the empirical data. It is easy to derive a central limit theorem for LSTD. Finite time bounds follow from establishing uniform convergence rates of the empirical loss to the population loss. Unfortunately, such techniques appear to be quite distinct from those needed to understand the online TD algorithms studied in this paper. Online TD has seen much wider use due to significant computational advantages [19].

Gradient TD methods are another related class of algorithms. These were derived by [13, 12] to address the issue that TD can diverge in so-called "off-policy" settings, where data is collected from a policy different from the one for which we want to estimate the value function. Unlike the classic TD(0) algorithm, gradient TD methods are designed to mimic gradient descent with respect to the mean squared projected Bellman error. [13, 12] propose asymptotically convergent two-time scale stochastic approximation schemes based on this and more recently [16] give a finite time analysis of two time scale stochastic approximation algorithms, including several variants of gradient TD algorithms. Alternatively, [47] and [14] propose to reformulate the original gradient TD optimization as a primal-dual saddle point problem and leverage convergence analysis form that literature to give a non-asymptotic analysis. This work was later revisited by [15], who established a faster rate of convergence. The works of [16, 14] and [15] all consider only

i.i.d. observation noise. One interesting open question is whether our techniques for treating the Markovian observation model will also apply to these analyses. Finally, it is worth highlighting that, to the best of our knowledge, substantial new techniques are needed to analyze the widely used TD(0), TD(λ) and the Q-learning studied in this paper. Unlike gradient TD methods, they do not mimic noisy gradient steps with respect to any fixed objective[3].

2.3 Problem formulation

Markov reward process. We consider the problem of evaluating the value function V_μ of a given policy μ in a Markov decision process (MDP). We work in the *on policy* setting, where data is generated by applying the policy μ in the MDP. Because the policy μ is applied automatically to select actions, such problems are most naturally formulated as value function estimation in a Markov reward process (MRP). A MRP[4] comprises of $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma)$ [19] where $\mathcal{S}$ is the set of finite states, $\mathcal{P}$ is the Markovian transition kernel, $\mathcal{R}$ is a reward function, and $\gamma < 1$ is the discount factor. For a discrete state-space $\mathcal{S}$, $\mathcal{P}(s'|s)$ specifies the probability of transitioning from a state s to another state s'. The reward function $\mathcal{R}(s, s')$ associates a reward with each state transition. We denote by $\mathcal{R}(s) = \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s)\mathcal{R}(s, s')$ the expected instantaneous reward generated from an initial state s.

The value function associated with this MRP, V_μ, specifies the expected cumulative discounted future reward as a function of the state of the system. In particular,

$$V_\mu(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t) \mid s_0 = s\right],$$

where the expectation is over sequences of states generated according to the transition kernel $\mathcal{P}$. This value function obeys the Bellman equation $T_\mu V_\mu = V_\mu$, where the Bellman operator T_μ asso-

[3]This can be formally verified for TD(0) with linear function approximation. If the TD step were a gradient with respect to a fixed objective, differentiating it should give the Hessian and hence a symmetric matrix. Instead, the matrix one attains is typically not a symmetric one.

[4]We avoid μ from notation for simplicity.

ciates a value function $V : \mathcal{S} \to \mathbb{R}$ with another value function $T_\mu V$ satisfying

$$(T_\mu V)(s) = \mathcal{R}(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s)V(s') \quad \forall\ s \in \mathcal{S}.$$

We assume rewards are bounded uniformly such that

$$|\mathcal{R}(s, s')| \leq r_{\max} \quad \forall\ s, s' \in \mathcal{S}.$$

Under this assumption, value functions are assured to exist and are the unique solution to Bellman's equation [48]. We also assume that the Markov reward process induced by following the policy μ is ergodic with a unique stationary distribution π. For any two states s, s': $\pi(s') = \lim_{t\to\infty} \mathbb{P}(s_t = s'|s_0 = s)$.

Following common references [48, 49, 50], we will simplify the presentation by assuming the state space $\mathcal{S}$ is a finite set of size $n = |\mathcal{S}|$. Working with a finite state space allows for the use of compact matrix notation, which is the convention in work on linear value function approximation. It also avoids measure theoretic notation for conditional probability distributions. Our proofs extend in an obvious way to problems with countably infinite state-spaces, as long the uniform ergodicity condition stated in Section 2.9, Assumption 1, continues to hold. For problems with general state-space, even the core results in dynamic programming hold only under suitable technical conditions [51].

Value function approximation. Given a fixed policy μ, the problem is to efficiently estimate the corresponding value function V_μ using only the observed rewards and state transitions. Unfortunately, due to the curse of dimensionality, most modern applications have intractably large state spaces, rendering exact value function learning hopeless. Instead, researchers resort to parametric approximations of the value function, for example by using a linear function approximator [19] or a non-linear function approximation such as a neural network [52]. In this work, we consider a

linear function approximation architecture where the true value-to-go $V_\mu(s)$ is approximated as

$$V_\mu(s) \approx V_\theta(s) = \phi(s)^\top \theta,$$

where $\phi(s) \in \mathbb{R}^d$ is a fixed feature vector for state s and $\theta \in \mathbb{R}^d$ is a parameter vector that is shared across states. When the state space is the finite set $\mathcal{S} = \{s_1, \ldots, s_n\}$, $V_\theta \in \mathbb{R}^n$ can be expressed compactly as

$$V_\theta = \begin{bmatrix} \phi(s_1)^\top \\ \vdots \\ \phi(s_n)^\top \end{bmatrix} \theta = \begin{bmatrix} \phi_1(s_1) & \phi_k(s_1) & \phi_d(s_1) \\ \vdots & \vdots & \vdots \\ \phi_1(s_n) & \phi_k(s_n) & \phi_d(s_n) \end{bmatrix} \theta = \Phi\theta,$$

where $\Phi \in \mathbb{R}^{n\times d}$ and $\theta \in \mathbb{R}^d$. We assume throughout that the d features vectors $\{\phi_k\}_{k=1}^d$, forming the columns of Φ are linearly independent.

Norms in value function and parameter space. For a symmetric positive definite matrix A, define the inner product $\langle x, y\rangle_A = x^\top A y$ and the associated norm $\|x\|_A = \sqrt{x^\top A x}$. If A is positive semi-definite rather than positive definite then $\|\cdot\|_A$ is called a semi-norm. Let $D = \mathrm{diag}(\pi(s_1), \ldots, \pi(s_n)) \in \mathbb{R}^{n\times n}$ denote the diagonal matrix whose elements are given by the entries of the stationary distribution $\pi(\cdot)$. Then, for two value functions V and V',

$$\|V - V'\|_D = \sqrt{\sum_{s\in\mathcal{S}} \pi(s)\,(V(s) - V'(s))^2},$$

measures the mean-square difference between the value predictions under V and V', in steady-state. This suggests a natural norm on the space of parameter vectors. In particular, for any $\theta, \theta' \in \mathbb{R}^d$,

$$\|V_\theta - V_{\theta'}\|_D = \sqrt{\sum_{s\in\mathcal{S}} \pi(s)\,(\phi(s)^\top(\theta - \theta'))^2} = \|\theta - \theta'\|_\Sigma$$

where

$$\Sigma := \Phi^\top D \Phi = \sum_{s \in \mathcal{S}} \pi(s) \phi(s) \phi(s)^\top$$

is the steady-state feature covariance matrix.

Feature regularity. We assume that feature vectors are uniformly bounded, that is $\sup_{s \in \mathcal{S}} \|\phi(s)\|_2 < \infty$. For notational convenience, we assume features are normalized so that $\|\phi(s)\|_2 \leq 1$ for all $s \in \mathcal{S}$. This is without loss of generality because the TD algorithm is invariant to feature re-scaling. Precisely, TD applied with feature mapping $\phi(\cdot)$ and initial parameter θ_0 produces an identical sequence of value functions to the TD algorithm with feature mapping $\tilde{\phi}(\cdot) = k\phi(\cdot)$ and initial parameter $\tilde{\theta}_0 = \theta_0/k$, for any scalar $k > 0$. Note that all our results bound the mean-squared gap between value predictions. We also assume that any entirely redundant or irrelevant features have been removed, so Σ has full rank. Let $\omega > 0$ be the minimum eigenvalue of Σ. From our bound on the feature vectors, the maximum eigenvalue of Σ is less than 1, so $1/\omega$ bounds the condition number of the feature covariance matrix[5]. The following lemma is an immediate consequence of our assumptions.

Lemma 1 (Norm equivalence)**.** *For all* $\theta \in \mathbb{R}^d$, $\sqrt{\omega}\|\theta\|_2 \leq \|V_\theta\|_D \leq \|\theta\|_2$.

One typical style of result in the study of strongly convex optimization gives fast rates of convergence in terms of the number of iterations T. But these bounds degrade when ω is very small and generally require apriori knowledge of some good lower bound on ω. We give some results in that style, but also give results in the style of [53], where bounds and stepsizes have no dependence on ω.

2.4 Temporal difference learning

We consider the classic temporal difference learning algorithm [6]. The algorithm starts with an initial parameter estimate θ_0 and at every time step t, it observes one data tuple $O_t = (s_t, r_t =$

[5]Let $\lambda_{\max}(A) = \max_{\|x\|_2=1} x^\top A x$ denote the maximum eigenvalue of a symmetric positive-semidefinite matrix. Since this is a convex function, $\lambda_{\max}(\Sigma) \leq \sum_{s \in \mathcal{S}} \pi(s) \lambda_{\max}(\phi(s)\phi(s)^\top) \leq \sum_{s \in \mathcal{S}} \pi(s) = 1$.

$\mathcal{R}(s_t, s_t'), s_t')$ consisting of the current state, the current reward and the next state reached by playing policy μ in the current state. This tuple is used to define a loss function, which is taken to be the squared sample Bellman error. The algorithm then proceeds to compute the next iterate θ_{t+1} by taking a gradient-like update. Some of our bounds guarantee accuracy of the average iterate, denoted by $\bar{\theta}_t = t^{-1} \sum_{i=0}^{t-1} \theta_i$. The version of TD presented in Algorithm 1 below also makes online updates to the averaged iterate.

TD is not a true stochastic gradient method with respect to any fixed loss function, which makes its analysis challenging. The TD update can be written as $g_t(\theta) = (y_t - V_\theta(s_t)) \frac{d}{d\theta} V_\theta(s_t)$, where $y_t = r_t + \gamma V_\theta(s_t')$ is a sample based estimate of the Bellman update to V_{θ_t}. Then $g_t(\theta_t) = -\frac{\partial}{\partial \theta} \frac{1}{2} (y_t - V_\theta(s_t))^2 \big|_{\theta=\theta_t}$ can be interpreted as the negative gradient of a certain squared loss function, *but this calculation treats the target y_t as fixed and ignores its implicit dependence on θ_t.* To emphasize the contrast with stochastic gradient methods, [19] refer to TD as a *semi-gradient* method. Accordingly, we will refer to $g_t(\cdot)$ as *negative semi-gradient* throughout the paper.

We present in Algorithm 1 the simplest variant of TD, which is known as TD(0). It is also worth highlighting that here we study online temporal difference learning, which makes incremental semi-gradient updates to the parameter estimate based on the most recent data observations only. Such algorithms are widely used in practice, but harder to analyze than so-called batch TD methods like the LSTD algorithm of [54].

Algorithm 1: TD(0) with linear function approximation

Input : initial guess θ_0, step-size sequence $\{\alpha_t\}_{t \in \mathbb{N}}$.
Initialize: $\bar{\theta}_0 \leftarrow \theta_0$.
for $t = 0, 1, \ldots$ **do**
  Observe tuple: $O_t = (s_t, r_t = \mathcal{R}(s_t, s_t'), s_t')$
  Define target: $y_t = \mathcal{R}(s_t, s_t') + \gamma V_{\theta_t}(s_t')$ `/* sample Bellman operator */`
  Loss function: $\frac{1}{2}(y_t - V_\theta(s_t))^2$ `/* sample Bellman error squared */`
  Compute negative semi-gradient: $g_t(\theta_t) = -\frac{\partial}{\partial \theta} \frac{1}{2} (y_t - V_\theta(s_t))^2 \big|_{\theta=\theta_t}$
  Take a semi-gradient step: $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$ `/* `α_t`: step-size */`
  Update averaged iterate: $\bar{\theta}_{t+1} \leftarrow \left(\frac{t}{t+1}\right) \bar{\theta}_t + \left(\frac{1}{t+1}\right) \theta_t$ `/* `$\bar{\theta}_{t+1} = \frac{1}{t+1} \sum_{\ell=0}^{t} \theta_\ell$ `*/`
end

At time t, TD takes a step in the direction of the negative semi-gradient $g_t(\theta_t)$ evaluated at

the current parameter. As a general function of θ and the tuple $O_t = (s_t, r_t, s'_t)$, the negative semi-gradient can be written as

$$g_t(\theta) = \left(r_t + \gamma\phi(s'_t)^\top\theta - \phi(s_t)^\top\theta\right)\phi(s_t). \tag{2.1}$$

The long-run dynamics of TD are closely linked to the expected negative semi-gradient step when the tuple $O_t = (s_t, r_t, s'_t)$ follows its *steady-state* behavior:

$$\bar{g}(\theta) := \sum_{s,s'\in\mathcal{S}} \pi(s)\mathcal{P}(s'|s)\left(\mathcal{R}(s,s') + \gamma\phi(s')^\top\theta - \phi(s)^\top\theta\right)\phi(s) \quad \forall\ \theta \in \mathbb{R}^d.$$

This can be rewritten more compactly in several useful ways. One such way is,

$$\bar{g}(\theta) = \mathbb{E}\left[\phi r\right] + \mathbb{E}\left[\phi(\gamma\phi' - \phi)^\top\right]\theta, \tag{2.2}$$

where $\phi = \phi(s)$ is the feature vector of a random initial state $s \sim \pi$, $\phi' = \phi(s')$ is the feature vector of a random next state drawn according to $s' \sim \mathcal{P}(\cdot \mid s)$, and $r = \mathcal{R}(s, s')$. In addition, since $\sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s)\left(\mathcal{R}(s,s') + \gamma\phi(s')^\top\theta\right) = (T_\mu\Phi\theta)(s)$, we can recognize that

$$\bar{g}(\theta) = \Phi^\top D(T_\mu\Phi\theta - \Phi\theta). \tag{2.3}$$

See [9] for a derivation of this fact.

2.5 Asymptotic convergence of temporal difference learning

The main challenge in analyzing TD is that the semi-gradient steps $g_t(\theta)$ are not true stochastic gradients with respect to any fixed objective. The semi-gradient step taken at time t pulls the value prediction $V_{\theta_{t+1}}(s_t)$ closer to y_t, but y_t itself depends on V_{θ_t}. So does this circular process converge? The key insight of [9] was to interpret this as a stochastic approximation scheme for solving a fixed point equation known as the projected Bellman equation. Contraction properties together

with general results from stochastic approximation theory can then be used to show convergence.

Should TD converge at all, it should be to a stationary point. Because the feature covariance matrix Σ is full rank there is a unique[6] vector θ^* with $\bar{g}(\theta^*) = 0$. We briefly review results that offer insight into θ^* and proofs of the asymptotic convergence of TD.

Understanding the TD limit point. Tsitsiklis and Van Roy [9] give an interesting characterization of the limit point θ^*. They show it is the unique solution to the *projected* Bellman equation

$$\Phi\theta = \Pi_D T_\mu \Phi\theta, \tag{2.4}$$

where $\Pi_D(\cdot)$ is the projection operator onto the subspace $\{\Phi x \mid x \in \mathbb{R}^d\}$ spanned by these features in the inner product $\langle\cdot,\cdot\rangle_D$. To see why this is the case, note that by using $\bar{g}(\theta^*) = 0$ along with Equation (2.3),

$$0 = x^\top \bar{g}(\theta^*) = \langle \Phi x \, , \, T_\mu \Phi\theta^* - \Phi\theta^* \rangle_D \quad \forall \; x \in \mathbb{R}^d.$$

That is, the Bellman error at θ^*, given by $(T_\mu\Phi\theta^* - \Phi\theta^*)$, is orthogonal to the space spanned by the features in the inner product $\langle\cdot,\cdot\rangle_D$. By definition, this means $\Pi_D \left(T_\mu\Phi\theta^* - \Phi\theta^*\right) = 0$ and hence θ^* must satisfy the projected Bellman equation.

The following lemma shows the projected Bellman operator, $\Pi_D T_\mu(\cdot)$ is a contraction, and so in principle, one could converge to the approximate value function $\Phi\theta^*$ by repeatedly applying it. TD appears to serve as a simple stochastic approximation scheme for solving the projected-Bellman fixed point equation.

Lemma 2. ***[Tsitsiklis and Van Roy [9]]*** *$\Pi_D T_\mu(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus γ, that is,*

$$\left\|\Pi_D T_\mu V_\theta - \Pi_D T_\mu V_{\theta'}\right\|_D \leq \gamma \|V_\theta - V_{\theta'}\|_D \quad \forall\, \theta, \theta' \in \mathbb{R}^d.$$

Finally, the limit of convergence comes with some competitive guarantees. From Lemma 2, a short

[6]This follows formally as a consequence of Lemma 3 in this paper.

argument shows

$$\left\|V_{\theta^*} - V_\mu\right\|_D \leq \frac{1}{\sqrt{1-\gamma^2}}\left\|\Pi_D V_\mu - V_\mu\right\|_D. \tag{2.5}$$

See Chapter 6 of [48] for a proof. The left hand side of Equation (2.5) measures the root-mean-squared deviation between the value predictions of the limiting TD value function and the true value function. On the right hand side, the projected value function $\Pi_D V_\mu$ minimizes root-mean-squared prediction errors among all value functions in the span of Φ. If V_μ actually falls within the span of the features, there is no approximation error at all and TD converges to the true value function.

Asymptotic convergence via the ODE method. Like many analyses in reinforcement learning, the convergence proof of [9] appeals to a powerful technique from the stochastic approximation literature known as the "ODE method". Under appropriate conditions, and assuming a decaying step-size sequence satisfying the Robbins-Monro conditions, this method establishes the asymptotic convergence of the stochastic recursion $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$ as a consequence of the global asymptotic stability of the deterministic ODE: $\dot{\theta}_t = \bar{g}(\theta_t)$. The critical step in the proof of [9] is to use the contraction properties of the Bellman operator to establish this ODE is globally asymptotically stable with the equilibrium point θ^*.

The ODE method vastly simplifies convergence proofs. First, because the continuous dynamics can be easier to analyze than discretized ones, and more importantly, because it avoids dealing with stochastic noise in the problem. At the same time, by side-stepping these issues, the method offers little insight into the critical effect of step-size sequences, problem conditioning, and mixing time issues on algorithm performance.

2.6 Outline of analysis

The remainder of this chapter focuses on a finite time analysis of TD. Broadly, we establish two types of finite time bounds on $\mathbb{E}\left[\|V_{\bar{\theta}_T} - V_{\theta^*}\|_D^2\right]$, which measures the mean-squared gap between the value predictions under the averaged-iterate $\bar{\theta}_T$ and under the TD limit point θ^*. We first derive

bounds that depend on the condition number of the feature covariance matrix. These mirror what one might expect from the literature on stochastic optimization of strongly convex functions: results showing that TD with constant step-sizes converges to within a radius of V_{θ^*} at an exponential rate, and $O(1/T)$ convergence rates with appropriate decaying step-sizes.

These results establish fast rates of convergence, but only if the problem is well conditioned. The choice of step-sizes is also very sensitive to problem conditioning. Work on robust stochastic approximation [53] argues instead for the use of comparatively large step-sizes together with iterate averaging[7]. Following the spirit of this work, we also give explicit bounds on $\mathbb{E}\left[\|V_{\bar{\theta}_T} - V_{\theta^*}\|_D^2\right]$ with a slower $O(1/\sqrt{T})$ convergence rates, but importantly both the bounds and step-sizes are completely independent of problem conditioning.

Our approach is to start by developing insights from simple, stylized settings, and then incrementally extend the analysis to more complex settings. The analysis is outlined below.

Noiseless case: Drawing inspiration from the ODE method discussed above, we start by analyzing the Euler discretization of the ODE $\dot{\theta}_t = \bar{g}(\theta_t)$, which is the deterministic recursion $\theta_{t+1} = \theta_t + \alpha\bar{g}(\theta_t)$. We call this method "mean-path TD". As motivation, the section first considers a fictitious gradient descent algorithm designed to converge to the TD fixed point. We then develop striking analogues for mean-path TD of the key properties underlying the convergence of gradient descent. Easy proofs then yield two bounds mirroring those given for gradient descent.

Independent noise: Section 2.8 studies TD under an i.i.d. observation model, where the data-tuples used by TD are drawn i.i.d. from the stationary distribution. The techniques used to analyze mean-path TD(0) extend easily to this setting, and the resulting bounds mirror standard guarantees for stochastic gradient descent.

Markov noise: In Section 2.9, we analyze TD in the more realistic setting where the data is col-

[7]This approach argues for using step-sizes of the order of $1/\sqrt{t}$, where t is the current iteration. These are much larger than the stepsizes, on the order of $1/t$, that are suggested in the classical stochastic approximation literature. This should not be confused with the approach of using even larger stepsizes that do not depend on t or the total number of iterations T (for example see [17] as well as the related works of [26, 27] and [28]).

lected from a single sample path of an ergodic Markov chain. This setting introduces significant challenges due to the highly dependent nature of the data. For tractability, we assume the Markov chain satisfies a certain uniform bound on the rate at which it mixes, and study a variant of TD that uses a projection step to ensure uniform boundedness of the iterates. In this case, our results essentially scale by a factor of the mixing time relative to the i.i.d. case.

Extension to TD(λ): In Section 2.10, we extend the analysis under the Markov noise to TD with eligibility traces, popularly known as TD(λ). Eligibility traces are known to often provide performance gains in practice, but theoretical analysis is more complex. Such analysis offers some insight into the subtle tradeoffs in the selection of the parameter $\lambda \in [0, 1]$.

Approximate optimal stopping: A final section extends our results to a class of high dimensional optimal stopping problems. We analyze Q-learning with linear function approximation. Building on observations of [20], we show the key properties used in our analysis of TD continue to hold for Q-learning in this setting. The convergence bounds shown in Sections 2.8 and 2.9 therefore apply *without any modification*.

2.7 Analysis of mean-path TD

All practical applications of TD involve observation noise. However, a great deal of insight can be gained by investigating a natural deterministic analogue of the algorithm. Here we study the recursion

$$\theta_{t+1} = \theta_t + \alpha \bar{g}(\theta_t) \qquad t \in \mathbb{N}_0 = \{0, 1, 2, \ldots\},$$

which is the Euler discretization of the ODE described in Section 2.5. We will refer to this iterative algorithm as *mean-path TD*. In this section, we develop key insights into the dynamics of mean-path TD that allow for a remarkably simple finite time analysis of its convergence. Later sections of the paper show how these ideas extend gracefully to analyses with observation noise.

The key to our approach is to develop properties of mean-path TD that closely mirror those of gradient descent on a particular quadratic loss function. To this end, in the next subsection, we

review a simple analysis of gradient descent. In Subsection 2.7.2, we establish key properties of mean-path TD mirroring those used to analyze this gradient descent algorithm. Finally, Subsection 2.7.3 gives convergence rates of mean-path TD, with proofs and rates mirroring those given for gradient descent except for a constant that depends on the discount factor, γ.

2.7.1 Gradient descent on a value function loss

Consider the cost function

$$f(\theta) = \frac{1}{2}\|V_{\theta^*} - V_\theta\|_D^2 = \frac{1}{2}\|\theta^* - \theta\|_\Sigma^2,$$

which measures the mean-squared gap between the value predictions under θ and those under the stationary point of TD, θ^*. Consider as well a hypothetical algorithm that performs gradient descent on f, iterating $\theta_{t+1} = \theta_t - \alpha\nabla f(\theta_t)$ for all $t \in \mathbb{N}_0$. Of course, this algorithm is not implementable, as one does not know the limit point θ^* of TD. However, reviewing an analysis of such an algorithm will offer great insights into our eventual analysis of TD.

To start, a standard decomposition characterizes the evolution of the error at iterate θ_t:

$$\|\theta^* - \theta_{t+1}\|_2^2 = \|\theta^* - \theta_t\|_2^2 + 2\alpha\nabla f(\theta_t)^\top(\theta^* - \theta_t) + \alpha^2\|\nabla f(\theta_t)\|_2^2.$$

To use this decomposition, we need two things. First, some understanding of $\nabla f(\theta_t)^\top(\theta^* - \theta_t)$, capturing whether the gradient points in the direction of $(\theta^* - \theta_t)$. And second, we need an upper bound on the norm of the gradient $\|\nabla f(\theta_t)\|_2^2$. In this case, $\nabla f(\theta) = \Sigma(\theta - \theta^*)$, from which we conclude

$$\nabla f(\theta)^\top(\theta^* - \theta) = -\|\theta^* - \theta\|_\Sigma^2 = -\|V_{\theta^*} - V_\theta\|_D^2. \tag{2.6}$$

In addition, one can show[8]

$$\|\nabla f(\theta)\|_2 \leq \|V_{\theta^*} - V_\theta\|_D. \tag{2.7}$$

Now, using (2.6) and (2.7), we have that for step-size $\alpha = 1$,

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - \|V_{\theta^*} - V_{\theta_t}\|_D^2. \tag{2.8}$$

The distance to θ^* decreases in every step, and does so more rapidly if there is a large gap between the value predictions under θ and θ^*. Combining this with Lemma 1 gives

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq (1-\omega)\|\theta^* - \theta_t\|_2^2 \leq \ldots \leq (1-\omega)^{t+1}\|\theta^* - \theta_0\|_2^2. \tag{2.9}$$

Recall that ω denotes the minimum eigenvalue of Σ. This shows that error converges at a fast geometric rate. However the rate of convergence degrades if the minimum eigenvalue ω is close to zero. Such a convergence rate is therefore only meaningful if the feature covariance matrix is well conditioned.

By working in the space of value functions and performing iterate averaging, one can also give a guarantee that is independent of ω. Recall the notation $\bar{\theta}_T = T^{-1}\sum_{t=0}^{T-1}\theta_t$ for the averaged iterate. A simple proof from (2.8) shows

$$\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\|V_{\theta^*} - V_{\theta_t}\|_D^2 \leq \frac{\|\theta^* - \theta_0\|_2^2}{T}. \tag{2.10}$$

2.7.2 Key properties of mean-path TD

This subsection establishes analogues for mean-path TD of the key properties (2.6) and (2.7) used to analyze gradient descent. First, to characterize the semi-gradient update, our analysis

[8]This can be seen from the fact that for any vector u with $\|u\|_2 \leq 1$,

$$u^\top \nabla f(\theta) = \langle u\,,\, \theta - \theta^*\rangle_\Sigma \leq \|u\|_\Sigma\|\theta^* - \theta\|_\Sigma \leq \|\theta^* - \theta\|_\Sigma = \|V_{\theta^*} - V_\theta\|_D.$$

builds on Lemma 7 of [9], which uses the contraction properties of the projected Bellman operator to conclude that

$$\bar{g}(\theta)^\top(\theta^* - \theta) > 0 \quad \forall\ \theta \neq \theta^*. \tag{2.11}$$

That is, the expected update of TD always forms a positive angle with $(\theta^* - \theta)$. Though only Equation (2.11) was stated in their lemma, [9] actually reach a much stronger conclusion in their proof itself. This result, given in Lemma 3 below, establishes that the expected updates of TD point in a descent direction of $\|\theta^* - \theta\|_2^2$, and do so more strongly when the gap between value functions under θ and θ^* is large. We will show that this more quantitative form of (2.11) allows for elegant finite time-bounds on the performance of TD.

Note that this lemma mirrors the property in Equation (2.6), but with a smaller constant of $(1 - \gamma)$. This reflects that expected TD must converge to θ^* by bootstrapping [6] and may follow a less direct path to θ^* than the fictitious gradient descent method considered in the previous subsection. Recall that the limit point θ^* solves $\bar{g}(\theta^*) = 0$.

Lemma 3. *For any $\theta \in \mathbb{R}^d$,*

$$\bar{g}(\theta)^\top(\theta^* - \theta) \geq (1 - \gamma)\|V_{\theta^*} - V_\theta\|_D^2.$$

Proof. We use the notation described in Equation (2.2) of Section 2.4. Consider a stationary sequence of states with random initial state $s \sim \pi$ and subsequent state s', which, conditioned on s, is drawn from $\mathcal{P}(\cdot|s)$. Set $\phi = \phi(s)$, $\phi' = \phi(s')$ and $r = \mathcal{R}(s, s')$. Define $\xi = V_{\theta^*}(s) - V_\theta(s) = (\theta^* - \theta)^\top\phi$ and $\xi' = V_{\theta^*}(s') - V_\theta(s') = (\theta^* - \theta)^\top\phi'$. By stationarity, ξ and ξ' are two correlated random variables with the same marginal distribution. By definition, $\mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$ since s is drawn from π.

Using the expression for $\bar{g}(\theta)$ in Equation (2.2),

$$\bar{g}(\theta) = \bar{g}(\theta) - \bar{g}(\theta^*) = \mathbb{E}[\phi(\gamma\phi' - \phi)^\top(\theta - \theta^*)] = \mathbb{E}[\phi(\xi - \gamma\xi')]. \tag{2.12}$$

Therefore

$$(\theta^* - \theta)^\top \bar{g}(\theta) = \mathbb{E}\left[\xi(\xi - \gamma\xi')\right] = \mathbb{E}[\xi^2] - \gamma\mathbb{E}[\xi'\xi] \geq (1-\gamma)\mathbb{E}[\xi^2] = (1-\gamma)\|V_{\theta^*} - V_\theta\|_D^2.$$

The inequality above uses Cauchy-Schwartz inequality together with the fact that ξ and ξ' have the same marginal distribution to conclude $\mathbb{E}[\xi\xi'] \leq \sqrt{\mathbb{E}[\xi^2]}\sqrt{\mathbb{E}[(\xi')^2]} = \mathbb{E}[\xi^2]$. □

Lemma 4 is the other key ingredient to our results. It upper bounds the norm of the expected negative semi-gradient, providing an analogue of Equation (2.7).

Lemma 4. $\|\bar{g}(\theta)\|_2 \leq 2\|V_\theta - V_{\theta^*}\|_D \ \forall\ \theta \in \mathbb{R}^d$.

Proof. Beginning from (2.12) in the Proof of Lemma 3, we have

$$\|\bar{g}(\theta)\|_2 = \|\mathbb{E}[\phi(\xi - \gamma\xi')]\|_2 \leq \sqrt{\mathbb{E}\left[\|\phi\|_2^2\right]}\sqrt{\mathbb{E}\left[(\xi - \gamma\xi')^2\right]} \leq \sqrt{\mathbb{E}[\xi^2]} + \gamma\sqrt{\mathbb{E}[(\xi')^2]} = (1+\gamma)\sqrt{\mathbb{E}[\xi^2]},$$

where the second inequality uses the assumption that $\|\phi\|_2 \leq 1$ and the final equality uses that ξ and ξ' have the same marginal distribution. We conclude by recalling that $\mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$ and $1 + \gamma \leq 2$. □

Lemmas 3 and 4 are quite powerful when used in conjunction. As in the analysis of gradient descent reviewed in the previous subsection, our analysis starts with a recursion for the error term, $\|\theta_t - \theta^*\|^2$. See Equation (2.13) in Theorem 1 below. Lemma 3 shows the first order term in this recursion reduces the error at each time step, while using the two lemmas in conjunction shows the first order term dominates a constant times the second order term. Precisely,

$$\bar{g}(\theta)^\top(\theta^* - \theta) \geq (1-\gamma)\|V_{\theta^*} - V_\theta\|_D^2 \geq \frac{(1-\gamma)}{4}\|\bar{g}(\theta)\|_2^2.$$

This leads immediately to conclusions like Equation (2.14), from which finite time convergence bounds follow. It is also worth pointing out that as TD(0) is an instance of linear stochastic approximation, these two lemmas can be interpreted as statements about the eigenvalues of the matrix

driving its behavior[9].

2.7.3 Finite time analysis of mean-path TD

We now combine the insights of the previous subsection to establish convergence rates for mean-path TD. These mirror the bounds for gradient descent given in Equations (2.9) and (2.10), except for an additional dependence on the discount factor. The first result bounds the distance between the value function under an averaged iterate and under the TD stationary point. This gives a comparatively slow $O(1/T)$ convergence rate, but does not depend at all on the conditioning of the feature covariance matrix. When this matrix is well conditioned, so the minimum eigenvalue ω of Σ is not too small, the geometric convergence rate given in the second part of the theorem dominates.

Theorem 1. *Consider a sequence of parameters* $(\theta_0, \theta_1, \ldots)$ *obeying the recursion*

$$\theta_{t+1} = \theta_t + \alpha \bar{g}(\theta_t) \qquad t \in \mathbb{N}_0 = \{0, 1, 2, \ldots\},$$

where $\alpha = (1-\gamma)/4$. *Then,*

$$\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \;\leq \frac{4\|\theta^* - \theta_0\|_2^2}{T(1-\gamma)^2}$$

and

$$\|V_{\theta^*} - V_{\theta_T}\|_D^2 \;\leq \exp\left\{-\left(\frac{(1-\gamma)^2\omega}{4}\right)T\right\}\|\theta^* - \theta_0\|_2^2.$$

Proof. With probability 1, for every $t \in \mathbb{N}_0$, we have

$$\|\theta^* - \theta_{t+1}\|_2^2 \;= \|\theta^* - \theta_t\|_2^2 - 2\alpha(\theta^* - \theta_t)^\top \bar{g}(\theta_t) + \alpha^2\|\bar{g}(\theta_t)\|_2^2. \tag{2.13}$$

[9]Recall from Section 2.4 that $\bar{g}(\theta)$ is an affine function. That is, it can be written as $A\theta - b$ for some $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. Lemma 3 shows that $A \preceq -(1-\gamma)\Sigma$, i.e. that $A + (1-\gamma)\Sigma$ is negative definite. It is easy to show that $\|\bar{g}(\theta)\|_2^2 = (\theta - \theta^*)^\top(A^\top A)(\theta - \theta^*)$, so Lemma 4 shows that $A^\top A \preceq \Sigma$. Taking this perspective, the important part of these lemmas is that they allow us to understand TD in terms of feature covariance matrix Σ and the discount factor γ rather than the more mysterious matrix A.

Applying Lemmas 3 and 4 and using a constant step-size of $\alpha = (1-\gamma)/4$, we get

$$\begin{aligned}\|\theta^* - \theta_{t+1}\|_2^2 &\leq \|\theta^* - \theta_t\|_2^2 - \left(2\alpha(1-\gamma) - 4\alpha^2\right)\|V_{\theta^*} - V_{\theta_t}\|_D^2 \\ &= \|\theta^* - \theta_t\|_2^2 - \left(\frac{(1-\gamma)^2}{4}\right)\|V_{\theta^*} - V_{\theta_t}\|_D^2. \end{aligned} \tag{2.14}$$

Then,

$$\left(\frac{(1-\gamma)^2}{4}\right)\sum_{t=0}^{T-1}\|V_{\theta^*} - V_{\theta_t}\|_D^2 \leq \sum_{t=0}^{T-1}\left(\|\theta^* - \theta_t\|_2^2 - \|\theta^* - \theta_{t+1}\|_2^2\right) \leq \|\theta^* - \theta_0\|_2^2.$$

Applying Jensen's inequality gives the first result:

$$\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\|V_{\theta^*} - V_{\theta_t}\|_D^2 \leq \frac{4\|\theta^* - \theta_0\|_2^2}{(1-\gamma)^2 T}.$$

Now, returning to (2.14), and applying Lemma 1 implies

$$\begin{aligned}\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - \left(\frac{(1-\gamma)^2}{4}\right)\omega\|\theta^* - \theta_t\|_2^2 &= \left(1 - \frac{\omega(1-\gamma)^2}{4}\right)\|\theta^* - \theta_t\|_2^2 \\ &\leq \exp\left\{-\frac{\omega(1-\gamma)^2}{4}\right\}\|\theta^* - \theta_t\|_2^2,\end{aligned}$$

where the final inequality uses that $\left(1 - \frac{\omega(1-\gamma)^2}{4}\right) \leq e^{\frac{-\omega(1-\gamma)^2}{4}}$. Repeating this inductively and using that $\|V_{\theta^*} - V_{\theta_T}\|_D^2 \leq \|\theta^* - \theta_T\|_2^2$ as shown in Lemma 1 gives the desired result. □

2.8 Analysis for the i.i.d. observation model

This section studies TD under an i.i.d. observation model, and establishes three explicit guarantees that mirror standard finite time bounds available for SGD. Specifically, we study a model where the random tuples observed by the TD algorithm are sampled i.i.d. from the stationary distribution of the Markov reward process. This means that for all states s and s',

$$\mathbb{P}\left[(s_t, r_t, s_t') = (s, \mathcal{R}(s,s'), s')\right] = \pi(s)\mathcal{P}(s'|s), \tag{2.15}$$

and the tuples $\{(s_t, r_t, s'_t)\}_{t\in\mathbb{N}}$ are drawn independently across time. Note that the probabilities in Equation (2.15) correspond to a setting where the first state s_t is drawn from the stationary distribution, and then s'_t is drawn from $\mathcal{P}(\cdot|s_t)$. This model is widely used for analyzing RL algorithms. See for example [12], [13], [10], and [25].

Theorem 2 follows from a unified analysis that combines the techniques of the previous section with typical arguments used in the SGD literature. All bounds depend on $\sigma^2 = \mathbb{E}[\|g_t(\theta^*)\|_2^2] = \mathbb{E}[\|g_t(\theta^*) - \bar{g}(\theta^*)\|_2^2]$, which roughly captures the variance of TD updates at the stationary point θ^*. The bound in part (a) follows the spirit of work on so-called *robust stochastic approximation* [53]. It applies to TD with iterate averaging and relatively large step-sizes. The result is a simple bound on the mean-squared gap between the value predictions under the averaged iterate and the TD fixed point. The main strength of this result is that the step-sizes and the bound do not depend at all on the condition number of the feature covariance matrix. Note that the requirement that $\sqrt{T} \geq 8/(1-\gamma)$ is not critical; one can carry out analysis using the step-size $\alpha_0 = \min\{(1-\gamma)/8, \sqrt{T}\}$, but the bounds we attain only become meaningful in the case where T is sufficiently large, so we chose to simplify the exposition.

Parts (b) and (c) provide faster convergence rates in the case where the feature covariance matrix is well conditioned. Part (b) studies TD applied with a constant step-size, which is common in practice. In this case, the value function V_{θ_t} will never converge to the TD fixed point, but our results show the expected distance to V_{θ^*} converges at an exponential rate below some level that depends on the choice of step-size. This is sometimes referred to as the rate at which the initial point V_{θ_0} is "forgotten". Bounds like this justify the common practice of starting with large step-sizes, and sometimes dividing the step-sizes in half once it appears error is no-longer decreasing. Part (c) attains an $O(1/T)$ convergence rate for a carefully chosen decaying step-size sequence. This step-size sequence requires knowledge of the minimum eigenvalue of the feature covariance matrix Σ, which plays a role similar to a strong convexity parameter in the optimization literature. In practice, this would need to be estimated, possibly by constructing a sample average approximation to the feature covariance matrix. The proof of part (c) closely follows an inductive argument

presented in [37]. Note that the bound in part (c) is only meaningful when T is large relative to $1/\omega$ and $(1-\gamma)^{-1}$. We suspect this is due to fundamental challenges in applying TD to problems with poor conditioning or long time horizons, but it would be interesting to formally validate this.

We should note that stepsizes were chosen to enable a convenient finite time analysis. Alternative choices may lead to stronger bounds and better practical performance. As in [37], our results in parts (b) and (c) could be modified so that the stepsizes and final bound depend on some underestimate $\omega' < \omega$, of the true minimum eigenvalue ω. However, the challenge of setting such stepsizes is one of the major reasons [53] advocate instead for results like those in part (a) of Theorem 2. It is also worth noting that our analysis in part (a) can be extended to decreasing stepsizes of the form $\alpha_t = \min\{(1-\gamma)/8,\ 1/\sqrt{t}\}$, at the expense of slightly worse constants. Such extensions are common in the optimization literature. See for example Corollary 3.2.8 of [55]. Recall that $\bar{\theta}_T = T^{-1}\sum_{t=0}^{T-1}\theta_t$ denotes the averaged iterate. We show the following result.

Theorem 2. *Suppose TD is applied under the i.i.d. observation model and set* $\sigma^2 = \mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right]$.

(a) For any $T \geq (8/(1-\gamma))^2$ *and a constant step-size sequence* $\alpha_0 = \cdots = \alpha_T = \frac{1}{\sqrt{T}}$,

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + 2\sigma^2}{\sqrt{T}(1-\gamma)}.$$

(b) For any constant step-size sequence $\alpha_0 = \cdots = \alpha_T \leq \omega(1-\gamma)/8$,

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_T}\|_D^2\right] \leq \left(e^{-\alpha_0(1-\gamma)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \alpha_0\left(\frac{2\sigma^2}{(1-\gamma)\omega}\right).$$

(c) For a decaying step-size sequence $\alpha_t = \frac{\beta}{\lambda+t}$ *with* $\beta = \frac{2}{(1-\gamma)\omega}$ *and* $\lambda = \frac{16}{(1-\gamma)^2\omega}$,

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_T}\|_D^2\right] \leq \frac{\nu}{\lambda+T} \quad \textit{where} \quad \nu = \max\left\{\frac{8\sigma^2}{(1-\gamma)^2\omega^2},\ \frac{16\|\theta^* - \theta_0\|_2^2}{(1-\gamma)^2\omega}\right\}.$$

Our proof is able to directly leverage Lemma 3, but the analysis requires the following extension of Lemma 4 which gives an upper bound to the expected norm of the semi-gradient.

Lemma 5. *For any fixed* $\theta \in \mathbb{R}^d$, $\mathbb{E}\left[\|g_t(\theta)\|_2^2\right] \leq 2\sigma^2 + 8\|V_\theta - V_{\theta^*}\|_D^2$ *where* $\sigma^2 = \mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right]$.

Proof. For brevity of notation, set $\phi = \phi(s_t)$ and $\phi' = \phi(s_t')$. Define $\xi = (\theta^* - \theta)^\top \phi$ and $\xi' = (\theta^* - \theta)^\top \phi'$. By stationarity, ξ and ξ' have the same marginal distribution and $\mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$, following the same argument as in Lemma 3. Using the formula for $g_t(\theta)$ in Equation (2.1), we have

$$\begin{aligned}
\mathbb{E}\left[\|g_t(\theta)\|_2^2\right] &\leq \mathbb{E}\left[\left(\|g_t(\theta^*)\|_2 + \|g_t(\theta) - g_t(\theta^*)\|_2\right)^2\right] \\
&\leq 2\mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right] + 2\mathbb{E}\left[\|g_t(\theta) - g_t(\theta^*)\|_2^2\right] \\
&= 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\phi - \gamma\phi'\right)^\top (\theta^* - \theta)\right\|_2^2\right] \\
&= 2\sigma^2 + 2\mathbb{E}\left[\|\phi(\xi - \gamma\xi')\|_2^2\right] \\
&\leq 2\sigma^2 + 2\mathbb{E}\left[|\xi - \gamma\xi'|^2\right] \\
&\leq 2\sigma^2 + 4\left(\mathbb{E}\left[|\xi|^2\right] + \gamma^2\mathbb{E}\left[|\xi'|^2\right]\right) \\
&\leq 2\sigma^2 + 8\|V_{\theta^*} - V_\theta\|_D^2,
\end{aligned}$$

where we used the assumption that $\|\phi\|_2^2 \leq 1$. The second inequality uses the basic algebraic identity $(x+y)^2 \leq 2\max\{x,y\}^2 \leq 2x^2 + 2y^2$, along with the linearity of expectation operators. □

Using this we give a proof of Theorem 2 below. Let us remark here on a consequence of the i.i.d noise model that considerably simplifies the proof. Until now, we have often developed properties of the TD updates $g_t(\theta)$ applied to an arbitrary, but fixed, vector $\theta \in \mathbb{R}^d$. For example, we have given an expression for $\bar{g}(\theta) := \mathbb{E}[g_t(\theta)]$, where this expectation integrates over the random tuple $O_t = (s_t, r_t, s_t')$ influencing the TD update. In the i.i.d noise model, the current iterate, θ_t, is independent of the tuple O_t, and so $\mathbb{E}[g_t(\theta_t)|\theta_t] = \bar{g}(\theta_t)$. In a similar manner, after conditioning on θ_t, we can seamlessly apply Lemmas 3 and 5, as is done in inequality (2.16) of the proof below.

Proof of Theorem 2. The TD algorithm updates the parameters as: $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$. Thus, for

each $t \in \mathbb{N}_0$, we have,

$$\|\theta^* - \theta_{t+1}\|_2^2 \ = \|\theta^* - \theta_t\|_2^2 - 2\alpha_t g_t(\theta_t)^\top(\theta^* - \theta_t) + \alpha_t^2 \|g_t(\theta_t)\|_2^2.$$

Under the hypotheses of (a), (b) and (c), we have that $\alpha_t \leq (1-\gamma)/8$. Taking expectations and applying Lemma 3 and Lemma 5 implies,

$$\begin{aligned}
\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] &= \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha_t \mathbb{E}\left[g_t(\theta_t)^\top(\theta^* - \theta_t)\right] + \alpha_t^2 \mathbb{E}\left[\|g_t(\theta_t)\|_2^2\right] \\
&= \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha_t \mathbb{E}\left[\mathbb{E}\left[g_t(\theta_t)^\top(\theta^* - \theta_t) \mid \theta_t\right]\right] + \alpha_t^2 \mathbb{E}\left[\mathbb{E}\left[\|g_t(\theta_t)\|_2^2 \mid \theta_t\right]\right] \\
&\leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - \left(2\alpha_t(1-\gamma) - 8\alpha_t^2\right) \mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] + 2\alpha_t^2\sigma^2 \qquad (2.16) \\
&\leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - \alpha_t(1-\gamma)\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] + 2\alpha_t^2\sigma^2. \qquad (2.17)
\end{aligned}$$

The inequality (2.16) follows from Lemmas 3 and 5. The application of these lemmas uses that the random tuple $O_t = (s_t, r_t, s_t')$ influencing $g_t(\cdot)$ is independent of the iterate, θ_t.

Part (a). Consider a constant step-size of $\alpha_T = \cdots = \alpha_0 = 1/\sqrt{T}$. Starting with Equation (2.17) and summing over t gives

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|V_{\theta^*} - V_{\theta_t}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{\alpha_0(1-\gamma)} + \frac{2\alpha_0 T\sigma^2}{(1-\gamma)} = \frac{\sqrt{T}\|\theta^* - \theta_0\|_2^2}{(1-\gamma)} + \frac{2\sqrt{T}\sigma^2}{(1-\gamma)}.$$

We find

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2\right] \leq \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \|V_{\theta^*} - V_{\theta_t}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + 2\sigma^2}{\sqrt{T}(1-\gamma)}.$$

Part (b). Consider a constant step-size of $\alpha_0 \leq \omega(1-\gamma)/8$. Applying Lemma 1 to Equation (2.17) implies

$$\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] \leq (1 - \alpha_0(1-\gamma)\omega)\, \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + 2\alpha_0^2\sigma^2. \qquad (2.18)$$

Iterating this inequality establishes that for any $T \in \mathbb{N}_0$,

$$\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] \leq (1 - \alpha_0(1-\gamma)\omega)^T \, \mathbb{E}\left[\|\theta^* - \theta_0\|_2^2\right] + 2\alpha_0^2\sigma^2 \sum_{t=0}^{\infty} (1 - \alpha_0(1-\gamma)\omega)^t \, .$$

The result follows by solving the geometric series and using that $(1 - \alpha_0(1-\gamma)\omega) \leq e^{-\alpha_0(1-\gamma)\omega}$ along with Lemma 1.

Part (c). Note that by the definitions of ν, λ and β, we have

$$\nu = \max\{2\beta^2\sigma^2 \, , \, \lambda\|\theta^* - \theta_0\|_2^2\}.$$

We then have $\|\theta^* - \theta_0\|_2^2 \leq \frac{\nu}{\lambda}$ by the definition of ν. Proceeding by induction, suppose $\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] \leq \frac{\nu}{\lambda+t}$. Then,

$$\begin{aligned}
\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] &\leq (1 - \alpha_t(1-\gamma)\omega)\, \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + 2\alpha_t^2\sigma^2 \\
&\leq \left(1 - \frac{(1-\gamma)\omega\beta}{\hat{t}}\right)\frac{\nu}{\hat{t}} + \frac{2\beta^2\sigma^2}{\hat{t}^2} \qquad [\text{where } \hat{t} \equiv \lambda + t] \\
&= \left(\frac{\hat{t} - (1-\gamma)\omega\beta}{\hat{t}^2}\right)\nu + \frac{2\beta^2\sigma^2}{\hat{t}^2} \\
&= \left(\frac{\hat{t} - 1}{\hat{t}^2}\right)\nu + \frac{2\beta^2\sigma^2 - ((1-\gamma)\omega\beta - 1)\,\nu}{\hat{t}^2} \\
&= \left(\frac{\hat{t} - 1}{\hat{t}^2}\right)\nu + \frac{2\beta^2\sigma^2 - \nu}{\hat{t}^2} \qquad [\text{using } \beta = \frac{2}{(1-\gamma)\omega}] \\
&\leq \frac{\nu}{\hat{t} + 1},
\end{aligned}$$

where the final inequality uses that $2\beta^2\sigma^2 - \nu \leq 0$, which holds by the definition of ν and the fact that $\hat{t}^2 \geq (\hat{t}-1)(\hat{t}+1)$. The final result follows by invoking the inequality $\|V_{\theta^*} - V_{\theta_T}\|_D^2 \leq \|\theta^* - \theta_T\|_2^2$ as shown in Lemma 1. □

2.9 Analysis for the Markov chain observation model: Projected TD algorithm

In Section 2.8, we developed a method for analyzing TD under an i.i.d. sampling model in which tuples are drawn independently from the stationary distribution of the underlying MDP. But a more realistic setting is one in which the observed tuples used by TD are gathered from a single trajectory of the Markov chain. In particular, if for a given sample path the Markov chain visits states $(s_0, s_1, \ldots s_t, \ldots)$, then these are processed into tuples $O_t = (s_t, r_t = \mathcal{R}(s_t, s_{t+1}), s_{t+1})$ that are fed into the TD algorithm. Mathematical analysis is difficult since the tuples used by the algorithm can be highly correlated with each other. We outline the main challenges below.

Challenges in the Markov chain noise model. In the i.i.d. observation setting, our analysis relied heavily on a Martingale property of the noise sequence. This no longer holds in the Markov chain model due to strong dependencies between the noisy observations. To understand this, recall the expression of the TD update,

$$g_t(\theta) = \Big(r_t + \gamma\phi(s_{t+1})^\top\theta - \phi(s_t)^\top\theta\Big)\phi(s_t). \tag{2.19}$$

To make the statistical dependencies more transparent, we can overload notation to write this as $g(\theta, O_t) \equiv g_t(\theta)$, where $O_t = (s_t, r_t, s_{t+1})$. Assuming the sequence of states is stationary, we have defined the function $\bar{g} : \mathbb{R}^d \to \mathbb{R}^d$ by $\bar{g}(\theta) = \mathbb{E}[g(\theta, O_t)]$, where, since θ is non-random, this expectation integrates over the marginal distribution of the tuple O_t. However, $\mathbb{E}[g(\theta_t, O_t) \mid \theta_t = \theta] \neq \bar{g}(\theta)$ because θ_t is a function of past tuples $\{O_1, \ldots, O_{t-1}\}$, potentially introducing strong dependencies between θ_t and O_t. Similarly, in general $\mathbb{E}[g(\theta_t, O_t) - \bar{g}(\theta_t)] \neq 0$, indicating bias in the algorithm's semi-gradient evaluation. A related challenge arises in trying to control the norm of the semi-gradient step, $\mathbb{E}[\|g_t(\theta_t)\|_2^2]$. Lemma 5 does not yield a bound due to coupling between the iterate θ_t and the observation O_t.

Our analysis uses an information-theoretic technique to control for this coupling and explicitly account for the semi-gradient bias. This technique may be of broader use in analyzing reinforce-

ment learning and stochastic approximation algorithms. However, our analysis also requires some strong regularity conditions, as outlined below.

Projected TD algorithm. Our technique for controlling the semi-gradient bias relies critically on a condition that, when step-sizes are small, the iterates $(\theta_t)_{t\in\mathbb{N}_0}$ do not change too rapidly. This is the case as long as norms of the semi-gradient steps do not explode. For tractability, we modify the TD algorithm itself by adding a projection step that ensures semi-gradient norms are uniformly bounded across time. In particular, starting with an initial guess of θ_0 such that $\|\theta_0\|_2 \leq R$, we consider the Projected TD algorithm, which iterates

$$\theta_{t+1} = \Pi_{2,R}\left(\theta_t + \alpha_t g_t(\theta_t)\right) \;\; \forall \; t \in \mathbb{N}_0, \tag{2.20}$$

where

$$\Pi_{2,R}(\theta) = \underset{\theta' : \|\theta'\|_2 \leq R}{\arg\min} \; \|\theta - \theta'\|_2$$

is the projection operator onto a norm ball of radius $R < \infty$. The subscript 2 on the operator indicates that the projection is with respect the unweighted Euclidean norm. This should not be confused with the projection operator Π_D used earlier, which projects onto the subspace of approximate value functions with respect to a weighted norm. One may wonder whether this projection step is practical. We note that, from a computational perspective, it only involves rescaling of the iterates, as $\Pi_{2,R}(\theta) = R\theta/\|\theta\|$ if $\|\theta\|_2 > R$ and is simply θ otherwise. In addition, Subsection 2.9.2 suggests that by using aprori bounds on the value function, it should be possible to estimate a projection radius containing the TD fixed point. However, at this stage, we view this mainly as a tool that enables clean finite time analysis, rather than a practical algorithmic proposal.

It is worth mentioning that projection steps have a long history in the stochastic approximation literature, and many of the standard analyses for stochastic gradient descent rely on projection steps to control the norm of the gradient [18, 56, 38, 53].

Structural assumptions on the Markov reward process. To control the statistical bias in the semi-gradient updates–which is the main challenge under the Markov observation model–we assume that the Markov chain mixes at a uniform geometric rate, as stated below.

Assumption 1. *There are constants $m > 0$ and $\rho \in (0, 1)$ such that*

$$\sup_{s \in \mathcal{S}} \; d_{TV}\left(\mathbb{P}(s_t \in \cdot | s_0 = s), \pi\right) \leq m\rho^t \quad \forall\, t \in \mathbb{N}_0,$$

where $d_{TV}(P, Q)$ denotes the total-variation distance between probability measures P and Q. In addition, the initial distribution of s_0 is the steady-state distribution π, so $(s_0, s_1, \ldots)$ is a stationary sequence.

This uniform mixing assumption always holds for irreducible and aperiodic finite-state Markov chains [57]. See [58] or [59] for a discussion of uniform ergodicity and relaxations of this concept in general state space Markov chains. We emphasize that the assumption that the chain begins in steady-state is not essential: given the uniform mixing assumption, we can always apply our analysis after the Markov chain has approximately reached its steady-state. However, adding this assumption allows us to simplify many mathematical expressions. Another useful quantity for our analysis is the mixing time which we define as

$$\tau^{\text{mix}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \epsilon\}. \tag{2.21}$$

For interpreting the bounds, note that from Assumption 1,

$$\tau^{\text{mix}}(\epsilon) \sim \frac{\log(m/\epsilon)}{\log(1/\rho)} \qquad \text{as } \epsilon \to 0.$$

We can therefore evaluate the mixing time at very small thresholds like $\epsilon = 1/T$ while only contributing a logarithmic factor to the bounds.

A bound on the norm of the semi-gradient: Before proceeding, we also state a bound on the Euclidean norm of the semi-gradient under TD(0) that follows from the uniform bound on rewards, along with feature normalization[10] and boundedness of the iterates through the projection step. Under projected TD(0) with projection radius R, this lemma implies that $\|g_t(\theta_t)\|_2 \leq (r_{\max} + 2R)$. This semi-gradient bound plays an important role in our convergence bounds.

Lemma 6. *For all $\theta \in \mathbb{R}^d$, $\|g_t(\theta)\|_2 \leq r_{\max} + 2\|\theta\|_2$ with probability 1.*

Proof. Using the expression of $g_t(\theta)$ in Equation (2.19), we have

$$\|g_t(\theta)\|_2 \;\leq\; |r_t + (\gamma\phi(s_t') - \phi(s_t))^\top\theta|\,\|\phi(s_t)\|_2 \;\leq\; r_{\max} + \|\gamma\phi(s_t') - \phi(s_t)\|_2\|\theta\|_2 \;\leq\; r_{\max} + 2\|\theta\|_2.$$

□

2.9.1 Finite time bounds

Following Section 2.8, we state several finite time bounds on the performance of the Projected TD algorithm. As before, in the spirit of robust stochastic approximation [53], the bound in part (a) gives a comparatively slow convergence rate of $\tilde{O}(1/\sqrt{T})$, but where the bound and step-size sequence are independent of the conditioning of the feature covariance matrix Σ. The bound in part (c) gives a faster convergence rate in terms of the number of samples T, but the bound and as well as the step-size sequence depend on the minimum eigenvalue ω of Σ. Part (b) confirms that for sufficiently small step-sizes, the value functions converge at an exponential rate to within some radius of the TD fixed-point, V_{θ^*}.

It is also instructive to compare the bounds for the Markov model vis-a-vis the i.i.d. model. One can see that in the case of part (b) for the Markov chain setting, a $O\left(G^2\tau^{\text{mix}}(\alpha_0)\right)$ term controls the limiting error due to semi-gradient noise. This scaling by the mixing time is intuitive, reflecting that roughly every cycle of $\tau^{\text{mix}}(\cdot)$ observations provides as much information as a single independent sample from the stationary distribution. We can also imagine specializing the results to the case of

[10]Recall that we assumed $\|\phi(s)\|_2 \leq 1$ for all $s \in \mathcal{S}$ and $|\mathcal{R}(s, s')| \leq r_{\max}$ for all $s, s' \in \mathcal{S}$

Projected TD under the i.i.d. model, thereby eliminating all terms depending on the mixing time. We would attain bounds that mirror those in Theorem 2, except that the semi-gradient noise term σ^2 there would be replaced by G^2. This is a consequence using G as a uniform upper bound on the semi-gradient norm in the proof, which is possible because of the projection step. Astute readers may notice the stepsize choices in parts (b) and (c) differ from those in parts (b) and (c) of Theorem 2. For each result, we have aimed for stepsize choices that lead to the simplest proofs of strong finite time bounds. In Theorem 3, the projection step allowed us to give a simple proof without requiring as small a stepsize as in Theorem 2. This choice may reflect our analysis technique more than any fundamental differences between the problems settings.

Theorem 3. *Suppose the Projected TD algorithm is applied with parameter $R \geq \|\theta^*\|_2$ under the Markov chain observation model with Assumption 1. Set $G = (r_{\max} + 2R)$. Then the following claims hold.*

(a) With a constant step-size sequence $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left(9 + 12\tau^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\gamma)}.$$

(b) With a constant step-size sequence $\alpha_0 = \cdots = \alpha_T < 1/(2\omega(1-\gamma))$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_T}\right\|_D^2\right] \leq \left(e^{-2\alpha_0(1-\gamma)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \alpha_0\left(\frac{G^2\left(9 + 12\tau^{\text{mix}}(\alpha_0)\right)}{2(1-\gamma)\omega}\right).$$

(c) With a decaying step-size sequence $\alpha_t = 1/(\omega(t+1)(1-\gamma))$ for all $t \in \mathbb{N}_0$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{G^2\left(9 + 24\tau^{\text{mix}}(\alpha_T)\right)}{T(1-\gamma)^2\omega}\left(1 + \log T\right),$$

There are two noteworthy points here. First, the proof of part (c) also implies an $\tilde{O}(1/T)$ convergence rate for the value function V_{θ_T} itself; however the bound degrades by a factor of ω. We refer the readers to Equation (A.4) in Appendix A.1.2 for the complete result. Second, it is

likely possible to eliminate the $\log T$ term in the numerator of part (c) to get a $O(1/T)$ convergence rate. One approach is to use a different weighting of the iterates when averaging, as in [56]. For brevity and simplicity, we do not pursue this direction.

2.9.2 Choice of the projection radius

We briefly comment on the choice of the projection radius, R. Note that Theorem 3 assumes that $\|\theta^*\|_2 \leq R$, so the TD limit point lies within the projected ball. How do we choose such an R when θ^* is unknown? It turns out we can use Lemma 2, which relates the value function at the limit of convergence V_{θ^*} to the true value function, to give a conservative upper bound. This is shown in the following lemma.

Lemma 7. *$\|\theta^*\|_\Sigma \leq \frac{2r_{\max}}{(1-\gamma)^{3/2}}$ and hence $\|\theta^*\|_2 \leq \frac{2r_{\max}}{\sqrt{\omega}(1-\gamma)^{3/2}}$.*

Proof. See Appendix A.3 for a detailed proof. □

It is important to remark here that this bound is *problem dependent* as it depends on the minimum eigenvalue ω of the steady-state feature covariance matrix Σ. We believe that estimating ω online would make the projection step practical to implement. We also remark that while we have assumed for that feature vectors are bounded as $\|\phi(s)\|_2 \leq 1$, this is not required for the conclusion in Lemma 7. The required projection radius automatically reflects any scaling of the feature vectors through the minimum eigenvalue ω.

2.9.3 Analysis

We now present the key analysis used to establish Theorem 3. Throughout, we assume the conditions of the theorem hold: we consider the Markov chain observation model with Assumption 1 and study the Projected TD algorithm applied with parameter $R \geq \|\theta^*\|_2$ and some step-size sequence $(\alpha_0, \cdots, \alpha_T)$.

We fix some notation throughout the scope of this subsection. Define the set $\Theta_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$, so $\theta_t \in \Theta_R$ for each t because of the algorithm's projection step. Set $G = (r_{\max} + 2R)$,

so $\|g_t(\theta)\|_2 \leq G$ for all $\theta \in \Theta_R$ by Lemma 6. Finally, we set

$$\zeta_t(\theta) \equiv (g_t(\theta) - \bar{g}(\theta))^\top (\theta - \theta^*) \quad \forall\ \theta \in \Theta_R,$$

which can be thought of as the error in the evaluation of semi-gradient-update under parameter θ at time t.

Referring back to the analysis of the i.i.d. observation model, one can see that an error decomposition given in Equation (2.17) is the crucial component of the proof. The main objective in this section is to establish two key lemmas that yield a similar decomposition in the Markov chain observation model. The result can be stated cleanly in the case of a constant step-size. If $\alpha_0 = \cdots = \alpha_T = \alpha$, we show

$$\begin{aligned}\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] &\leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha(1-\gamma)\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] + 2\mathbb{E}[\alpha\zeta_t(\theta_t)] + \alpha^2 G^2 \\ &\leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha(1-\gamma)\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_t}\|_D^2\right] + 2\alpha^2\left(5 + 6\tau^{\text{mix}}(\alpha)\right)G^2.\end{aligned} \tag{2.22}$$

The first inequality follows from Lemma 8. The second follows from Lemma 11, which in the case of a constant step-size α shows $\mathbb{E}[\alpha\zeta_t(\theta_t)] \leq G^2(4 + 6\tau^{\text{mix}}(\alpha))\alpha^2$. Notice that bias in the semi-gradient enters into the analysis as if by scaling the magnitude of the noise in semi-gradient evaluations by a factor of the mixing time. From this decomposition, parts (a) and (b) of Theorem 3 follow by essentially copying the proof of Theorem 2. Similar, but messier, inequalities hold for any decaying step-size sequence, which allows us to establish part (c).

Error decomposition under Projected TD

The next lemma establishes a recursion for the error under projected TD(0) that hold for each sample path.

Lemma 8. *With probability 1, for every $t \in \mathbb{N}_0$,*

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2.$$

Proof. From the projected TD(0) recursion in Equation (2.20), for any $t \in \mathbb{N}_0$,

$$\begin{aligned}
\|\theta^* - \theta_{t+1}\|_2^2 &= \|\theta^* - \Pi_{2,R}(\theta_t + \alpha_t g_t(\theta_t))\|_2^2 \\
&= \|\Pi_{2,R}(\theta^*) - \Pi_{2,R}(\theta_t + \alpha_t g_t(\theta_t))\|_2^2 \\
&\leq \|\theta^* - \theta_t - \alpha_t g_t(\theta_t)\|_2^2 \\
&= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t g_t(\theta_t)^\top(\theta^* - \theta_t) + \alpha_t^2\|g_t(\theta_t)\|_2^2 \\
&\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t g_t(\theta_t)^\top(\theta^* - \theta_t) + \alpha_t^2 G^2. \\
&= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t \bar{g}(\theta_t)^\top(\theta^* - \theta_t) + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2. \\
&\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2.
\end{aligned}$$

The first inequality used that orthogonal projection operators onto a convex set are non-expansive[11], the second used Lemma 6 together with the fact that $\|\theta_t\|_2 \leq R$ due to projection, and the third used Lemma 3. □

By taking expectation of both sides, this inequality could be used to produce bounds in the same manner as in the previous section, except that in general $\mathbb{E}[\zeta_t(\theta_t)] \neq 0$ due to bias in the semi-gradient evaluations.

Information–theoretic techniques for controlling the semi-gradient bias

The uniform mixing condition in Assumption 1 can be used in conjunction with some information theoretic inequalities to control the magnitude of the semi-gradient bias. This section presents a general lemma, which is the key to this analysis. We start by reviewing some important properties

[11] Let $\mathcal{P}_C(x) = \arg\min_{x' \in C} \|x' - x\|$ denote the projection operator onto a closed, non-empty, convex set $C \subset \mathbb{R}^d$. Then $\|\mathcal{P}_C(x) - \mathcal{P}_C(y)\| \leq \|x - y\|$ for all vectors x and y.

of information-measures.

Information theory background. The total-variation distance between two probability measures is a special case of the more general f-divergence defined as

$$d_f(P||Q) = \int f\left(\frac{dP}{dQ}\right) dQ,$$

where f is a convex function such that $f(1) = 0$. By choosing $f(x) = |x-1|/2$, one recovers the total-variation distance. A choice of $f(x) = x\log(x)$ yields the Kullback-Leibler divergence. This yields a generalization of the mutual information between two random variables X and Y. The f-information between X and Y is the f-divergence between their joint distribution and the product of their marginals:

$$I_f(X,Y) = d_f(\mathbb{P}(X=\cdot, Y=\cdot)\,,\, \mathbb{P}(X=\cdot)\otimes\mathbb{P}(Y=\cdot)).$$

This measure satisfies several nice properties. By definition it is symmetric, so $I_f(X,Y) = I_f(Y,X)$. It can be expressed in terms of the expected divergence between conditional distributions:

$$I_f(X,Y) = \sum_x \mathbb{P}(X=x) d_f(\mathbb{P}(Y=\cdot|X=x), \mathbb{P}(Y=\cdot)). \tag{2.23}$$

Finally, it satisfies the following data-processing inequality. If $X \to Y \to Z$ forms a Markov chain, then

$$I_f(X,Z) \le I_f(X,Y).$$

Here, we use the notation $X \to Y \to Z$, which is standard in information theory and the study of graphical models, to indicate that the random variables Z and X are independent conditioned on Y. Note that by symmetry we also have $I_f(X,Z) \le I_f(Y,Z)$. To use these results in conjunction with Assumption 1, we can specialize to total-variation distance (d_{TV}) and total-variation mutual information (I_{TV}) using $f(x) = |x-1|/2$. The total-variation is especially useful for our purposes

because of the following variational representation.

$$d_{\mathrm{TV}}(P, Q) = \sup_{v:\|v\|_\infty \le \frac{1}{2}} \left| \int v dP - \int v dQ \right|. \tag{2.24}$$

In particular, if P and Q are close in total-variation distance, then the expected value of any bounded function under P will be close to that under Q.

Information theoretic control of coupling. With this background in place, we are ready to establish a general lemma, which is central to our analysis. We use $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ to denote the supremum norm of a function $f : \mathcal{X} \to \mathbb{R}$.

Lemma 9 (Control of coupling)**.** *Consider two random variables X and Y such that*

$$X \to s_t \to s_{t+\tau} \to Y$$

for some fixed $t \in \{0, 1, 2, \ldots\}$ and $\tau > 0$. Assume the Markov chain mixes uniformly, as stated in Assumption 1. Let X' and Y' denote independent copies drawn from the marginal distributions of X and Y, so $\mathbb{P}(X' = \cdot, Y' = \cdot) = \mathbb{P}(X = \cdot) \otimes \mathbb{P}(Y = \cdot)$. Then, for any bounded function v,

$$|\mathbb{E}\left[v(X, Y)\right] - \mathbb{E}\left[v(X', Y')\right]| \le 2\|v\|_\infty (m\rho^\tau).$$

Proof. Let $P = \mathbb{P}(X \in \cdot, Y \in \cdot)$ denote the joint distribution of X and Y and $Q = \mathbb{P}(X \in \cdot) \otimes \mathbb{P}(Y \in \cdot)$ denote the product of the marginal distributions. Let $h = \frac{v}{2\|v\|_\infty}$, which is the function v rescaled to take values in $[-1/2, 1/2]$. Then, by Equation (2.24)

$$\mathbb{E}[h(X, Y)] - \mathbb{E}[h(X', Y')] = \int h dP - \int h dQ \le d_{\mathrm{TV}}(P, Q) = I_{\mathrm{TV}}(X, Y),$$

where the last equality uses the definition of the total-variation mutual information, I_{TV}. Then,

$$\begin{aligned} I_{\mathrm{TV}}(X,Y) \leq I_{\mathrm{TV}}(s_t, s_{t+\tau}) &= \sum_{s\in\mathcal{S}} \mathbb{P}(s_t = s) d_{\mathrm{TV}}\left(\mathbb{P}(s_{t+\tau} = \cdot \mid s_t = s)\right), \mathbb{P}(s_{t+\tau} = \cdot)) \\ &\leq \sup_{s\in\mathcal{S}} d_{\mathrm{TV}}\left(\mathbb{P}(s_{t+\tau} = \cdot \mid s_t = s), \pi\right) \\ &\leq m\rho^{\tau}, \end{aligned}$$

where the steps follow, respectively, from the data-processing inequality, the property in Equation (2.23), the stationarity of the Markov chain, and the the uniform mixing condition in Assumption 1. Combining these steps gives

$$|\mathbb{E}\left[v(X,Y)\right] - \mathbb{E}\left[v(X',Y')\right]| \leq 2\|v\|_{\infty} I_{\mathrm{TV}}(X,Y) \leq 2\|v\|_{\infty} m\rho^{\tau}.$$

□

Bounding the semi-gradient bias

We are now ready to bound the expected semi-gradient error $\mathbb{E}[\zeta_t(\theta_t)]$. First, we establish some basic regularity properties of the function $\zeta_t(\cdot)$.

Lemma 10 (Semi-gradient error is bounded and Lipschitz)**.** *With probability 1,*

$$|\zeta_t(\theta)| \leq 2G^2 \qquad \forall\, \theta \in \Theta_R$$

and

$$|\zeta_t(\theta) - \zeta_t(\theta')| \leq 6G\|(\theta - \theta')\|_2 \qquad \forall\, \theta, \theta' \in \Theta_R.$$

Proof. The result follows from a straightforward application of the bounds $\|g_t(\theta)\|_2 \leq G$ and $\|\theta\|_2 \leq R \leq G/2$, which hold for each $\theta \in \Theta_R$. A full derivation is given in Appendix A.1.3. □

We now use Lemmas 9 and 10 to establish a bound on the expected semi-gradient error.

Lemma 11 (Bound on semi-gradient bias)**.** *Consider a non-increasing step-size sequence,* $\alpha_0 \geq \alpha_1 \ldots \geq \alpha_T$*. Fix any* $t < T$*, and set* $t^* \equiv \max\{0, t - \tau^{\text{mix}}(\alpha_T)\}$*. Then,*

$$\mathbb{E}\left[\zeta_t(\theta_t)\right] \leq G^2 \left(4 + 6\tau^{\text{mix}}(\alpha_T)\right) \alpha_{t^*}.$$

The following bound also holds:

$$\mathbb{E}\left[\zeta_t(\theta_t)\right] \leq 6G^2 \sum_{i=0}^{t-1} \alpha_i.$$

Proof. We break the proof down into three steps.

Step 1: Relate $\zeta_t(\theta_t)$ and $\zeta_t(\theta_{t-\tau})$.

Note that for any $i \in \mathbb{N}_0$,

$$\|\theta_{i+1} - \theta_i\|_2 = \left\|\Pi_{2,R}\left(\theta_i + \alpha_i g_i(\theta_i)\right) - \Pi_{2,R}(\theta_i)\right\|_2 \leq \|\theta_i + \alpha_i g_i(\theta_i) - \theta_i\|_2 = \alpha_i \|g_i(\theta_i)\|_2 \leq \alpha_i G.$$

Therefore,

$$\|\theta_t - \theta_{t-\tau}\|_2 \leq \sum_{i=t-\tau}^{t-1} \|\theta_{i+1} - \theta_i\|_2 \leq G \sum_{i=t-\tau}^{t-1} \alpha_i.$$

Applying Lemma 10, we conclude

$$\zeta_t(\theta_t) \leq \zeta_t(\theta_{t-\tau}) + 6G^2 \sum_{i=t-\tau}^{t-1} \alpha_i \qquad \text{for all } \tau \in \{0, \cdots, t\}. \tag{2.25}$$

Step 2: Bound $\mathbb{E}[\zeta_t(\theta_{t-\tau})]$ using Lemma 9.

Recall that the semi-gradient $g_t(\theta)$ depends implicitly on the observed tuple $O_t = (s_t, \mathcal{R}(s_t, s_{t+1}), s_{t+1})$. Let us overload notation to make this statistical dependency more transparent. Put

$$g(\theta, O_t) := g_t(\theta) = \left(r_t + \gamma\phi(s_{t+1})^\top\theta - \phi(s_t)^\top\theta\right)\phi(s_t) \qquad \theta \in \Theta_R$$

and

$$\zeta(\theta, O_t) := \zeta_t(\theta) = (g(\theta, O_t) - \bar{g}(\theta))^\top (\theta - \theta^*) \qquad \theta \in \Theta_R.$$

We have defined $\bar{g} : \Theta_R \to \mathbb{R}^d$ as $\bar{g}(\theta) = \mathbb{E}[g(\theta, O_t)]$ for all $\theta \in \Theta_R$, where this expectation integrates over the marginal distribution of O_t. Then, by definition, for any fixed (non-random) $\theta \in \Theta_R$,

$$\mathbb{E}[\zeta(\theta, O_t)] = (\mathbb{E}[g(\theta, O_t)] - \bar{g}(\theta))^\top (\theta - \theta^*) = 0.$$

Since $\theta_0 \in \Theta_R$ is non-random, it follows immediately that

$$\mathbb{E}[\zeta(\theta_0, O_t)] = 0. \tag{2.26}$$

We use Lemma 9 to bound $\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)]$. First, consider random variables $\theta'_{t-\tau}$ and O'_t drawn independently from the marginal distributions of $\theta_{t-\tau}$ and O_t, so $\mathbb{P}(\theta'_{t-\tau} = \cdot, O'_t = \cdot) = \mathbb{P}(\theta_{t-\tau} = \cdot) \otimes \mathbb{P}(O_t = \cdot)$. Then $\mathbb{E}[\zeta(\theta'_{t-\tau}, O'_t)] = \mathbb{E}[\mathbb{E}[\zeta(\theta'_{t-\tau}, O'_t) \mid \theta'_{t-\tau}]] = 0$. Since $|\zeta(\theta, O_t)| \leq 2G^2$ for all $\theta \in \Theta_R$ by Lemma 10 and $\theta_{t-\tau} \to s_{t-\tau} \to s_t \to O_t$ forms a Markov chain, applying Lemma 9 gives

$$\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] \leq 2(2G^2)(m\rho^\tau) = 4G^2 m\rho^\tau. \tag{2.27}$$

Step 3: Combine terms.

The second claim follows immediately from Equation (2.25) together with Equation (2.26). We focus on establishing the first claim. Taking the expectation of Equation (2.25) implies

$$\mathbb{E}[\zeta_t(\theta_t)] \leq \mathbb{E}[\zeta_t(\theta_{t-\tau})] + 6G^2\tau\alpha_{t-\tau} \qquad \forall \tau \in \{0, \cdots, t\}.$$

For $t \leq \tau^{\text{mix}}(\alpha_T)$, choosing $\tau = t$ gives

$$\mathbb{E}[\zeta_t(\theta_t)] \leq \underbrace{\mathbb{E}[\zeta_t(\theta_0)]}_{=0} + 6G^2 t\alpha_0 \leq 6G^2\tau^{\text{mix}}(\alpha_T)\alpha_0.$$

For $t > \tau^{\text{mix}}(\alpha_T)$, choosing $\tau = \tau_0 \equiv \tau^{\text{mix}}(\alpha_T)$ gives

$$\mathbb{E}[\zeta_t(\theta_t)] \leq 4G^2 m\rho^{\tau_0} + 6G^2\tau_0\alpha_{t-\tau_0} \leq 4G^2\alpha_T + 6G^2\tau_0\alpha_{t-\tau} \leq G^2\left(4 + 6\tau_0\right)\alpha_{t-\tau_0}.$$

where the second inequality used that $m\rho^{\tau_0} \leq \alpha_T$ by the definition of the mixing time $\tau_0 \equiv \tau^{\text{mix}}(\alpha_T)$ and the second inequality uses that step-sizes are non-increasing. □

Completing the proof of Theorem 3

Combining Lemmas 8 and 10 gives the error decomposition in Equation 2.22 for the case of a constant step-size. As noted at the beginning of this subsection, from this decomposition, parts (a) and (b) of Theorem 3 can be established by essentially copying the proof of Theorem 2. For completeness, this is included in Appendix A.1. For part (c), we closely follow analysis of SGD with decaying step-sizes presented in [56]. However, some headache is introduced because Lemma 11 includes terms of the form $\alpha_{t-\tau^{\text{mix}}(\alpha_T)}$ instead of the typical α_t terms present in analyses of SGD. A complete proof of part (c) is given in Appendix A.1 as well.

2.10 Extension to TD with eligibility traces

This section extends our analysis to provide finite time guarantees for temporal difference learning *with eligibility traces*. We study a class of algorithms, denoted by TD(λ) and parameterized by $\lambda \in [0, 1]$, that contains as a special case the TD(0) algorithm studied in previous sections[12]. For $\lambda > 0$, the algorithm maintains an eligibility trace vector, which is a geometric weighted average of the negative semi-gradients at all previously visited states, and makes parameter updates in the direction of the eligibility vector rather than the negative semi-gradient. Eligibility traces sometimes provide substantial performance improvements in practice [19]. Unfortunately, they also introduce subtle dependency issues that complicate theoretical analysis; to our knowledge, this section provides the *first* non-asymptotic analysis of TD(λ).

[12] TD(0) corresponds to $\lambda = 0$.

Our analysis focuses on the Markov chain observation model studied in the previous section and we mirror the technical assumptions used there. In particular, we assume that the Markov chain is stationary and mixes at a uniform geometric rate (Assumption 1). As before, for tractability, we study a projected variant of TD(λ).

2.10.1 Projected TD(λ) algorithm

TD(λ) makes a simple, but a highly consequential, modification to TD(0). The pseudo-code for this algorithm is presented below in Algorithm 2. As with TD(0), it observes a tuple $O_t = (s_t, r_t = \mathcal{R}(s_t, s_{t+1}), s_{t+1})$ at each time-step t and computes the TD error $\delta_t(\theta_t) = r_t + \gamma V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t)$. However, while TD(0) makes an update $\theta_{t+1} = \theta_t + \alpha_t \delta_t(\theta_t)\phi(s_t)$ in the direction of the feature vector at the current state, TD(λ) makes the update $\theta_{t+1} = \theta_t + \alpha_t \delta_t(\theta_t) z_{0:t}$. The vector $z_{0:t} = \sum_{k=0}^{t} (\gamma\lambda)^k \phi(s_{t-k})$ is called the eligibility trace which is updated incrementally as shown below in Algorithm 2. As the name suggests, the components of $z_{0:t}$ roughly capture the extent to which each feature is eligible for receiving credit or blame for an observed TD error [19, 60].

Algorithm 2: Projected TD(λ) with linear function approximation

Input : radius R, initial guess $\{\theta_0 : \|\theta_0\|_2 \leq R\}$, and step-size sequence $\{\alpha_t\}_{t\in\mathbb{N}}$
Initialize: $\bar{\theta}_0 \leftarrow \theta_0$, $z_{-1} = 0$, $\lambda \in [0, 1]$.
for $t = 0, 1, \ldots$ **do**
  Observe tuple: $O_t = (s_t, r_t, s_{t+1})$
  Get TD error: $\delta_t(\theta_t) = r_t + \gamma V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t)$ `/* sample Bellman error */`
  Update eligibility trace: $z_{0:t} = (\gamma\lambda) z_{0:t-1} + \phi(s_t)$ `/* Geometric weighting */`
  Compute update direction: $x_t(\theta_t, z_{0:t}) = \delta_t(\theta_t) z_{0:t}$
  Take a projected update step: $\theta_{t+1} = \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))$ `/* `α_t`: step-size */`
  Update averaged iterate: $\bar{\theta}_{t+1} \leftarrow \left(\frac{t}{t+1}\right)\bar{\theta}_t + \left(\frac{1}{t+1}\right)\theta_{t+1}$ `/* `$\bar{\theta}_{t+1} = \frac{1}{t+1}\sum_{\ell=1}^{t+1}\theta_\ell$ `*/`
end

Some new notation in Algorithm 2 should be highlighted. We use $x_t(\theta, z_{0:t}) = \delta_t(\theta) z_{0:t}$ to denote the update to the parameter vector θ at time t. This plays a role analogous to the negative semi-gradient $g_t(\theta)$ in TD(0).

2.10.2 Limiting behavior of TD(λ)

We now review results on the asymptotic convergence of TD(λ) due to [9]. This provides the foundation of our finite time analysis and also offers insight into how the algorithm differs from TD(0).

Before giving any results, let us note that just as the true value function $V_\mu(\cdot)$ is the unique solution to Bellman's fixed point equation $V_\mu = T_\mu V_\mu$, it is also the unique solution to a k-step Bellman equation $V_\mu = T_\mu^k V_\mu$. This can be written equivalently as

$$V_\mu(s) = \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t \mathcal{R}(s_t) + \gamma^{k+1} V(s_{k+1}) \mid s_0 = s\right] \qquad \forall s \in S,$$

where the expectation is over states sampled when policy μ is applied to the MDP. The asymptotic properties of TD(λ) are closely tied to a geometrically weighted version of the k-step Bellman equations described above. Define the averaged Bellman operator

$$(T_\mu^{(\lambda)} V)(s) = (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t \mathcal{R}(s_t) + \gamma^{k+1} V(s_{k+1}) \mid s_0 = s\right]. \tag{2.28}$$

One interesting interpretation of this equation is as a k-step Bellman equation, but where the horizon k itself is a random geometrically distributed random variable.

Tsitsiklis and Van Roy [9] showed that under appropriate technical conditions, the approximate value function $V_{\theta_t} = \Phi\theta_t$ estimated by TD(λ) converges almost surely to the unique solution, θ^* of the projected fixed point equation

$$\Phi\theta = \Pi_D T_\mu^{(\lambda)} \Phi\theta.$$

TD(λ) is then interpreted as a stochastic approximation scheme for solving this fixed point equation. The existence and uniqueness of such a fixed point is implied by the following lemma, which shows that $\Pi_D T^{(\lambda)}(\cdot)$ is a contraction operator with respect to the steady-state weighted norm $\|\cdot\|_D$. Throughout this section, we let θ^* denote be the unique fixed point of the projected TD(λ) operator as shown above.

Lemma 12. ***[Tsitsiklis and Van Roy [9]]*** *$\Pi_D T_\mu^{(\lambda)}(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus*

$$\kappa = \frac{\gamma(1-\lambda)}{1-\gamma\lambda} \leq \gamma < 1.$$

As with TD(0), the limiting value function under TD(λ) comes with some competitive guarantees. A short argument using Lemma 12 shows

$$\left\|V_{\theta^*} - V_\mu\right\|_D \leq \frac{1}{\sqrt{1-\kappa^2}}\left\|\Pi_D V_\mu - V_\mu\right\|_D. \tag{2.29}$$

See for example Chapter 6 of [48] for a proof. It is important to note the distinction between the convergence guarantee results for TD(λ) and TD(0) in terms of the contraction factors. The contraction factor κ is always less than γ, the contraction factor under TD(0). In addition, as $\lambda \to 1$, $\kappa \to 0$ implying that the limit point of TD(λ) for large enough λ will be arbitrarily close to $\Pi_D V_\mu$, which minimizes the mean-square error in value predictions among all value functions representable by the features. This calculation suggests a choice of $\lambda = 1$ will offer the best performance. However, the *rate of convergence* also depends on λ, and may degrade as λ grows. Disentangling such issues requires also a careful study of the statistical efficiency of TD(λ), which we undertake in the following subsection.

2.10.3 Finite time bounds for Projected TD(λ)

Following Section 2.9, we establish three finite time bounds on the performance of the Projected TD(λ) algorithm. The first bound in part (a) does not depend on any special regularity of the problem instance but gives a comparatively slow convergence rate of $\tilde{O}(1/\sqrt{T})$. It applies with the robust (problem independent) and aggressive step-size of $1/\sqrt{T}$. Part (b) shows an exponential rate of convergence to within some radius of the TD(λ) fixed-point under a sufficiently small step-size. Part (c) attains an improved dependence on T of $\tilde{O}(1/T)$, but the step-size sequence requires knowledge of the minimum eigenvalue ω of Σ.

Compared to the results for TD(0), our bounds depend on a slightly different definition of the mixing time that takes into account the geometric weighting in the eligibility trace term. Define

$$\tau_\lambda^{\text{mix}}(\epsilon) = \max\{\tau^{\text{MC}}(\epsilon), \tau^{\text{Algo}}(\epsilon)\}, \tag{2.30}$$

where we denote $\tau^{\text{MC}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \epsilon\}$ and $\tau^{\text{Algo}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid (\gamma\lambda)^t \leq \epsilon\}$. As we show next, this definition of mixing time enables compact bounds for convergence rates of TD(λ).

Theorem 4. *Suppose the Projected TD(λ) algorithm is applied with parameter $R \geq \|\theta^*\|_2$ under the Markov chain observation model with Assumption 1. Set $B = \frac{(r_{\max}+2R)}{(1-\gamma\lambda)}$. Then the following claims hold.*

(a) With a constant step-size $\alpha_t = \alpha_0 = 1/\sqrt{T}$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2\left(13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\kappa)}.$$

(b) With a constant step-size $\alpha_t = \alpha_0 < 1/(2\omega(1-\kappa))$ and $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_T}\right\|_D^2\right] \leq \left(e^{-2\alpha_0(1-\kappa)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \alpha_0\left(\frac{B^2\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right)}{2(1-\kappa)\omega}\right).$$

(c) With a decaying step-size $\alpha_t = 1/(\omega(t+1)(1-\kappa))$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{B^2\left(13 + 52\tau_\lambda^{\text{mix}}(\alpha_T)\right)}{T(1-\kappa)^2\omega}\left(1 + \log T\right).$$

As was the case for TD(0), the proof of part (c) also implies an $\tilde{O}(1/T)$ convergence rate for the value function V_{θ_T} itself; however the bound degrades by a factor of ω. We refer the readers to Equation (A.4) in Appendix A.2.2 for the complete result. Again, a different weighting of the iterates as shown in [56] might enable us to eliminate the $\log T$ term in the numerator of part (c) to

give a $O(1/T)$ convergence rate. For brevity, we do not pursue this direction.

We now compare the bounds for TD(λ) with that of TD(0) ignoring the constant terms. It should be emphasized that these are only *upper bounds*, so differences could be due to looseness of the analysis rather than true differences in statistical performance. First, let us look at the results for the robust step-size $\alpha_t = 1/\sqrt{T}$ in part (a) of Theorems 3 and 4. Approximately, for the TD(λ) case, we have the term $\frac{B^2}{\sqrt{T}(1-\kappa)}$ vis-a-vis the term $\frac{G^2}{\sqrt{T}(1-\gamma)}$ for the TD(0) case. A simple argument below clarifies the relationship between these two.

$$\frac{B^2}{\sqrt{T}(1-\kappa)} = \frac{(r_{\max}+2R)^2}{\sqrt{T}(1-\kappa)(1-\gamma\lambda)^2} = \frac{G^2}{\sqrt{T}(1-\kappa)(1-\gamma\lambda)^2} \geq \frac{G^2}{\sqrt{T}(1-\kappa)(1-\gamma\lambda)} = \frac{G^2}{\sqrt{T}(1-\gamma)}.$$

As we will see later, B is an upper bound to the norm of $x_t(\theta_t, z_{0:t})$, the update direction for TD(λ). Correspondingly, from Section 2.9, we know that G is the upper bound on semi-gradient norm, $g_t(\theta_t)$ for TD(0). Intuitively, for TD(λ), the bound B is larger (due to the presence of the eligibility trace term) and more so as $\lambda \to 1$. This calculation reveals that our bounds give a slower rate of convergence for TD(λ) than for TD(0). This means more data is required for our bound to guarantee TD(λ) is close to its limit point. In this context, however, the trade-off we remarked on in Section 2.10.2 is noteworthy as the fixed point for TD(λ) comes with a better error guarantee.

Interestingly, for decaying step-sizes $\alpha_t = 1/(\omega(t+1)(1-\kappa))$, the bounds are qualitatively the same. This follows as the terms that dominate part (c) of Theorems 3 and 4 are equal:

$$\frac{B^2}{T(1-\kappa)^2} = \frac{(r_{\max}+2R)^2}{T(1-\kappa)^2(1-\gamma\lambda)^2} = \frac{G^2}{T(1-\kappa)^2(1-\gamma\lambda)^2} = \frac{G^2}{T(1-\gamma)^2}.$$

It is unclear whether the difference between the two stepsize regimes is an artifact of our analysis technique.

2.11 Extension: Q-learning for high dimensional Optimal Stopping

So far, this chapter has dealt with the problem of approximating the value function of a fixed policy in a computationally and statistically efficient manner. The Q-learning algorithm is one natural extension of temporal-difference learning to control problems, where the goal is to learn an effective policy from data. Although it is widely applied in reinforcement learning, in general Q-learning is unstable and its iterates may oscillate forever. An important exception to this was discovered by [20], who showed that Q-learning converges asymptotically for optimal stopping problems. In this section, we show how the techniques developed in Sections 2.8 and 2.9 can be applied *in an identical manner* to give finite time bounds for Q-learning with linear function approximation applied to optimal-stopping problems with high dimensional state spaces. To avoid repetition, we only state key properties satisfied by Q-learning in this setting which establish exactly the same convergence bounds as shown in Theorems 2 and 3.

2.11.1 Problem formulation

The optimal stopping problem is that of determining the time to terminate a process to maximize cumulative expected rewards accrued. Problems of this nature arise naturally in many settings, most notably in the pricing of financial derivatives [21, 22, 23]. We first give a brief formulation for a class of optimal stopping problems. A more detailed exposition can be found in [20], or Chapter 5 of the thesis work of [61].

Consider a discrete-time Markov chain $\{s_t\}_{t\geq 0}$ with finite state space $\mathcal{S}$ and unique stationary distribution distribution π. At each time t, the decision-maker observes the state s_t and decides whether to stop or continue. Let $\gamma \in [0, 1)$ denote the discount factor and let $u(\cdot)$ and $U(\cdot)$ denote the reward functions associated with *continuation* and *termination* decisions respectively. Let the stopping time τ denote the (random) time at which the decision-maker stops. The expected total

discounted reward from initial state s associated with the stopping time τ is

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t u(s_t) + \gamma^\tau U(s_\tau) \,\middle|\, s_0 = s\right], \tag{2.31}$$

where $U(s_\tau)$ is defined to be zero for $\tau = \infty$. We seek an optimal stopping policy, which determines when to stop as a function of the observed states so as to maximize (2.31).

For any Markov decision process, the optimal state-action value function $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the expected value to go from choosing an action $a \in \mathcal{A}$ in a state $s \in \mathcal{S}$ and following the optimal policy in subsequent states. In optimal stopping problems, there are only two possible actions at every time step: whether to *terminate* or to *continue*. The value of stopping in state s is just $U(s)$, which allows us to simplify notation by only representing the continuation value.

For the remainder of this section, we let $Q^* : \mathcal{S} \to \mathbb{R}$ denote the optimal continuation-value function. It can be shown that Q^* is the unique solution to the Bellman equation $Q^* = FQ^*$, where the Bellman operator is given by

$$FQ(s) = u(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s) \max\{U(s'), Q(s')\}.$$

Given the optimal continuation values $Q^*(\cdot)$, the optimal stopping time is simply

$$\tau^* = \min\{t \mid U(s_t) \geq Q^*(s_t)\}. \tag{2.32}$$

2.11.2 Q-Learning for high dimensional Optimal Stopping

In principle, one could generate the optimal stopping time using Equation (2.32) by applying exact dynamic programming algorithms to compute the optimal Q-function. However, such methods are only implementable for small state spaces. To scale to high dimensional state spaces, we consider a feature-based approximation of the optimal continuation value function, Q^*. We focus

on linear function approximation, where $Q^*(s)$ is approximated as

$$Q^*(s) \approx Q_\theta(s) = \phi(s)^\top \theta,$$

where $\phi(s) \in \mathbb{R}^d$ is a fixed feature vector for state s and $\theta \in \mathbb{R}^d$ is a parameter vector that is shared across states. As shown in Section 2.3, for a finite state space, $\mathcal{S} = \{s_1, \ldots, s_n\}$, $Q_\theta \in \mathbb{R}^n$ can be expressed compactly as $Q_\theta = \Phi\theta$, where $\Phi \in \mathbb{R}^{n\times d}$ and $\theta \in \mathbb{R}^d$. We also assume that the d feature vectors $\{\phi_k\}_{k=1}^d$, forming the columns of Φ are linearly independent.

We consider the Q-learning approximation scheme in Algorithm 3. The algorithm starts with an initial parameter estimate of θ_0 and observes a data tuple $O_t = (s_t, u(s_t), s_t')$. This is used to compute the target $y_t = u(s_t) + \gamma \max\{U(s_t'), Q_{\theta_t}(s_t')\}$, which is a sampled version of the $F(\cdot)$ operator applied to the current Q–function. The next iterate, θ_{t+1}, is computed by taking a semi-gradient step with respect to a loss function measuring the distance between y_t and predicted value-to-go. An important feature of this method is that problem data is generated by the exploratory policy that chooses to continue at all time-steps.

Algorithm 3: Q-Learning for Optimal Stopping problems.

Input : initial guess θ_0, step-size sequence $\{\alpha_t\}_{t\in\mathbb{N}}$ and radius R.
Initialize: $\bar{\theta}_0 \leftarrow \theta_0$.
for $t = 0, 1, \ldots$ **do**
  Observe tuple: $O_t = (s_t, u(s_t), s_t')$
  Target: $y_t = u(s_t) + \gamma \max\{U(s_t'), Q_{\theta_t}(s_t')\}$ `/* sample Bellman operator */`
  Define loss function: $\frac{1}{2}(y_t - Q_\theta(s_t))^2$ `/* sample Bellman error */`
  Compute negative semi-gradient: $g_t(\theta_t) = -\frac{\partial}{\partial\theta}\frac{1}{2}(y_t - Q_\theta(s_t))^2\big|_{\theta=\theta_t}$
  Take a semi-gradient step: $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$ `/* step-size */`
  Update averaged iterate: $\bar{\theta}_{t+1} \leftarrow \left(\frac{t}{t+1}\right)\bar{\theta}_t + \left(\frac{1}{t+1}\right)\theta_t$ `/*` $\bar{\theta}_{t+1} = \frac{1}{t+1}\sum_{\ell=1}^t \theta_\ell$ `*/`
end

2.11.3 Asymptotic guarantees

Similar to the asymptotic results for TD algorithms, [20] show that the variant of Q-learning detailed above in Algorithm 3 converges to the unique solution, θ^*, of the projected Bellman equa-

tion,

$$\Phi\theta = \Pi_D F \Phi\theta.$$

This results crucially relies on the fact that the projected Bellman operator $\Pi_D F(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus γ. The analogous result for our study of TD(0) was stated in Lemma 2. [20] also give error bounds for the limit of convergence with respect to Q^*, the optimal Q-function. In particular, it can be shown that

$$\|\Phi\theta^* - Q^*\|_D \leq \frac{1}{\sqrt{1-\gamma^2}} \|\Pi_D Q^* - Q^*\|_D$$

where the left hand side measures the error between the estimated and the optimal Q-function which is upper bounded by the *representational power* of the linear approximation architecture, as given on the right hand side. In particular, if Q^* can be represented as a linear combination of the feature vectors then there is no approximation error and the algorithm converges to the optimal Q-function. Finally, one can ask whether the stopping times suggested by this approximate continuation value function, $\Phi\theta^*$, are effective. Let $\tilde{\mu}$ be the policy that stops at the first time t when

$$U(s_t) \geq (\Phi\theta^*)(s_t).$$

Then, for an initial state s_0 drawn from the stationary distribution π,

$$\mathbb{E}\left[V^*(s_0)\right] - \mathbb{E}\left[V_{\tilde{\mu}}(s_0)\right] \leq \frac{2}{(1-\gamma)\sqrt{1-\gamma^2}} \|\Pi_D Q^* - Q^*\|_D,$$

where V^* and $V_{\tilde{\mu}}$ denote the value functions corresponding, respectively, to the optimal stopping policy the approximate stopping policy μ. Again, this error guarantee depends on the choice of feature representation.

2.11.4 Finite time analysis

In this section, we show how our results in Sections 2.8 and 2.9 for TD(0) and its projected counterpart can be extended, without any modification, to give convergence bounds for the Q-function approximation algorithm described above. To this effect, we highlight that the key lemmas which enable our analysis in Sections 2.8 and 2.9 also hold in this setting. The contraction property of the $F(\cdot)$ operator will be crucial to our arguments here. Convergence rates for an i.i.d. noise model, mirroring those established for TD(0) in Theorem 2, can be shown for Algorithm 3. Results for the Markov chain sampling model, mirroring those established for TD(0) in Theorem 3, can be shown for a projected variant of Algorithm 3.

First, we give mathematical expressions for the negative semi-gradient. As a general function of θ and tuple $O_t = (s_t, u(s_t), s_t')$, the negative semi-gradient can be written as

$$g_t(\theta) = \Big(u(s_t) + \gamma \max\left\{U(s_t'), \phi(s_t')^\top \theta\right\} - \phi(s_t)^\top \theta\Big)\phi(s_t). \tag{2.33}$$

The negative expected semi-gradient, when the tuple $(s_t, u(s_t), s_t')$ follows its steady-state behavior, can be written as

$$\bar{g}(\theta) = \sum_{s,s' \in \mathcal{S}} \pi(s)\mathcal{P}(s'|s) \left(u(s) + \gamma \max\left\{U(s'), \phi(s')^\top \theta\right\} - \phi(s)^\top \theta\right) \phi(s).$$

Additionally, using $\sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s) \left(u(s) + \gamma \max\left\{U(s'), \phi(s')^\top \theta\right\}\right) = \left(F\Phi\theta\right)(s)$, it is easy to show

$$\bar{g}(\theta) = \Phi^\top D(F\Phi\theta - \Phi\theta).$$

Note the close similarity of this expression with its counterparts for TD learning (see Section 2.4 and Appendix A.2); the only difference is that the appropriate Bellman operator(s) for TD learning, $T_\mu(\cdot)$, has been replaced with the appropriate Bellman operator $F(\cdot)$ for this optimal stopping problem.

Analysis with i.i.d. noise

In this section, we show how to analyze the Q-learning algorithm under an i.i.d. observation model, where the random tuples observed by the algorithm are sampled i.i.d. from the stationary distribution of the Markov process. All our ideas follow the presentation in Section 2.8, a careful understanding of which reveals that Lemmas 3 and 5 form the backbone of our results. Recall that Lemma 3 establishes how, at any iterate θ, TD updates point in the descent direction of $\|\theta^* - \theta\|_2^2$. Lemma 5 bounds the expected norm of the stochastic semi-gradient, thus giving a control over system noise.

In Lemmas 13 and 14, given below, we state exactly the same results for the Q-function approximation algorithm under the i.i.d. sampling model. With these two key lemmas, convergence bounds shown in Theorem 2 follows by repeating the analysis in Section 2.8. Recall that Q_{θ^*} denotes the unique fixed point of $\Pi_D F(\cdot)$, i.e. $Q_{\theta^*} = \Pi_D F Q_{\theta^*}$.

Lemma 13. ***[Tsitsiklis and Van Roy [20]]*** *For any $\theta \in \mathbb{R}^d$,*

$$\bar{g}(\theta)^\top (\theta^* - \theta) \geq (1-\gamma)\|Q_{\theta^*} - Q_\theta\|_D^2.$$

Proof. This property is a consequence of the fact that $\Pi_D F(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus γ. It was established by [20] in the process of proving their Lemma 8. For completeness, we provide a standalone proof in Appendix A.3. □

Lemma 14. *For any fixed $\theta \in \mathbb{R}^d$, $\mathbb{E}\left[\|g_t(\theta)\|_2^2\right] \leq 2\sigma^2 + 8\|Q_\theta - Q_{\theta^*}\|_D^2$ where $\sigma^2 = \mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right]$.*

Proof. See Appendix A.3 for a detailed proof. □

Analysis under the Markov chain model

Analogous to Section 2.9, we analyze a projected variant of Algorithm 3 under the Markov chain sampling model. Let $\Theta_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$. Starting with an initial guess of $\theta_0 \in \Theta_R$, the algorithm updates to the next iterate by taking a semi-gradient step followed by

projection onto Θ_R, so iterates satisfy the stochastic recursion $\theta_{t+1} = \Pi_{2,R}(\theta_t + \alpha_t g_t(\theta_t))$. We make the similar structural assumptions to those in Section 2.9. In particular, assume the feature vectors and the continuation, termination rewards to be uniformly bounded, with $\|\phi(s)\|_2 \leq 1$ and $\max\{|u(s)|, |U(s)|\} \leq r_{\max}$ for all $s \in \mathcal{S}$. We assume $r_{\max} \leq R$, which can always be ensured by rescaling rewards or the projection radius. We first state a uniform bound on the semi-gradient norm.

Lemma 15. *Define $G = (r_{\max} + 2R)$. With probability 1, $\|g_t(\theta)\|_2 \leq G$ for all $\theta \in \Theta_R$.*

Proof. See Appendix A.3 for a detailed proof. □

If we assume the Markov process $(s_0, s_1, \ldots)$ satisfies Assumption 1, then Lemma 15 paves the way to show exactly the same convergence bounds as given in Theorem 3. For this, we refer the readers to Section 2.9 and Appendix A.1, where we show all the key lemmas and a detailed proof of Theorem 3. One can mirror the same proof, using Lemmas 13 and 15 in place of Lemmas 3 and 11, which apply to TD(0). In particular, note that we can use Lemma 15 along with some basic algebraic inequalities to show the semi-gradient bias, $\zeta_t(\theta)$, to be Lipschitz and bounded. This, along with the information-theoretic arguments of Lemma 9 enables the exact same upper bound on the semi-gradient bias as shown in Lemma 11. Combining these with standard proof techniques for SGD [56, 53] shows the convergence bounds for Q-learning.

2.12 Conclusions

In this chapter we provide a simple finite time analysis of a foundational and widely used algorithm known as temporal difference learning. Although asymptotic convergence guarantees for the TD method were previously known, characterizing its data efficiency stands as an important open problem. Our work makes a substantial advance in this direction by providing a number of explicit finite time bounds for TD, including in the much more complicated case where data is generated from a single trajectory of a Markov chain. Our analysis inherits the simplicity of and elegance enjoyed by SGD analysis and can gracefully extend to different variants of TD, for exam-

ple TD learning with eligibility traces (TD(λ)) and Q-function approximation for optimal stopping problems. Owing to the close connection with SGD, we believe that optimization researchers can further build on our techniques to develop principled improvements to TD.

There are a number of research directions one can take to extend our work. First, we use a projection step for analysis under the Markov chain model, a choice we borrowed from the optimization literature to simplify our analysis. It will be interesting to find alternative ways to add regularity to the TD algorithm and establish similar convergence results; we think analysis without the projection step is possible if one can show that the iterates remain bounded under additional regularity conditions. Second, the $\tilde{O}(1/T)$ convergence rate we showed used step-sizes which crucially depends on the minimum eigenvalue ω of the feature covariance matrix, which would need to be estimated from samples. While such results are common in optimization for strongly convex functions, very recently [17] showed TD(0) with iterate averaging and *universal constant step-sizes* can attain an $\tilde{O}(1/T)$ convergence rate in the i.i.d. sampling model. Extending our analysis for problem independent, robust step-size choices is a research direction worth pursuing.

Chapter 3: Global Optimality Guarantees For Policy Gradient Methods

3.1 Introduction

Many recent successes in reinforcement learning are driven by a class of algorithms called policy gradient methods. These methods search over a parameterized class of polices by performing stochastic gradient descent on a cost function capturing the cumulative expected cost incurred. Specifically, for discounted or episodic problems, they treat the scalar cost function $\ell(\pi) = \int J_\pi(s) d\rho(s)$, which averages the total cost-to-go function J_π over a random initial state distribution ρ. Policy gradient methods aim to optimize over a smooth, and often stochastic, class of parameterized policies $\{\pi_\theta\}_{\theta \in \Theta}$ by performing stochastic gradient descent on $\ell(\cdot)$,

$$\theta_{k+1} = \theta_k - \alpha_k \left(\nabla_\theta \ell(\pi_{\theta_k}) + \text{noise} \right).$$

This approach has several attractive features leading to its widespread adoption including in conjunction with deep neural network based approaches [62, 63, 64]. It can be easily implemented using simulation based Monte Carlo approximations to the true gradient in an end-to-end fashion, directly optimizing the true decision objective rather than searching for approximate models of the underlying problem or other dynamic programming methods which approximate value functions that minimize Bellman error. Therefore, it can be useful in problems where there is an inductive bias about the form of policy that might be effective, rather than the form of an approximate model or value function. Gradient descent with finite stepsizes often makes small updates to the policy in each iteration making this scheme more stable than say approximate policy iteration which is prone to chattering.

Unfortunately, while policy gradient methods can be applied to a very broad class of problems,

it is not clear whether they adequately address even simple control problems solvable by classical methods. The challenge is that total cost $\ell(\cdot)$ is a non-convex function of the chosen policy. Typical of results concerning black-box optimization of non-convex functions, policy gradient methods are widely understood to converge asymptotically to a stationary point or a local minimum. Existing optimization theory guarantees this under technical conditions [65, 66, 67] and it is widely repeated in textbooks and surveys [68, 69, 1]. But the reinforcement learning literature seems to provide almost no guarantees into the *quality* of the points to which policy gradient methods converge. Worse yet, Example 1 clearly shows how policy gradient methods could get stuck in a bad local minima for very simple examples where the policy class is rich enough to contain the optimal policy.

Example 1 (Failure of policy gradient with a constrained policy class that contains the optimal policy)**.** *Consider the MDP depicted in Figure 3.1 (a). There are two states, left* (S_L) *and right* (S_R)*, and two possible actions, L and R, which move the agent to the desired state in the next period. Staying in the state L incurs a cost,* $g(S_L, L) = 1$ *per period, whereas staying in the right state is cost-less,* $g(S_R, R) = 0$*. Moving between states incurs a per-period cost of* 2*. As long as the discount factor exceeds* $1/2$*, it is optimal policy to play action R in either state*[1]*. For* $\gamma > 1/2$*, it is therefore reasonable to search in a constrained policy class,* $\{\pi_\theta : \theta \in [0, 1]\}$ *which plays the action R with probability* $\theta \in [0, 1]$ *regardless of the current state, as it contains the optimal policy. Unfortunately, as shown in Figure 3.1 (b), the total discounted cost incurred* $\ell(\pi_\theta)$ *is a non-convex function of* θ*. When initialized with small value of* θ*, cost is locally increasing as a function of* θ*, and so a gradient method updates the policy toward a bad local minimum at* $\theta = 0$*. Once there, any local policy search approach gets stuck as there are no descent directions to improve the total cost. It is worth noting here that in general, policy gradient methods face many additional challenges, for instance due to unsophisticated exploration or policy parameterization. This example however, clearly highlights the challenges presented by non-convexity of the objective* ℓ *which is fundamental to gradient based policy search methods.*

[1] Although it is always optimal to play R in S_R, one can easily check that for $\gamma < 1/2$, it is optimal to play L in S_L.

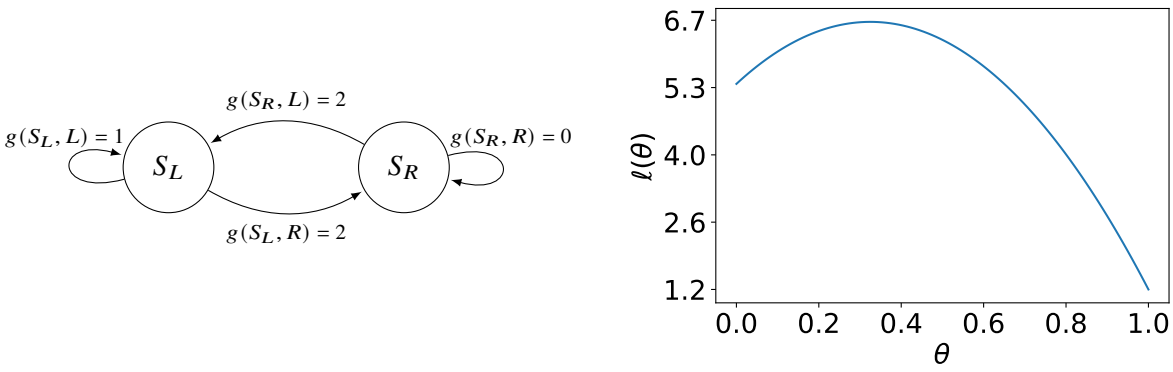


Figure 3.1: Policy gradient fails with the constrained policy class for a simple deterministic MDP. (a) Two state, two action MDP where the optimal policy, π^* plays right in both states. (b) For the constrained policy class, $\pi_\theta(R|S_L) = \pi_\theta(R|S_R) = \theta \in [0, 1]$, policy gradient objective is non-convex with two local minima. For (b), we took $\gamma = 0.8$ and $\rho = [0.6, 0.4]$. We remark that this example is not cherry picked. For any $\gamma > 1/2$ and different initial distributions ρ which put non-zero weight on both states, $\ell(\theta)$ has two local minima at $\theta = 0$ and $\theta = 1$.

In marked contrast to the example above, important recent work of [70] showed that for the deterministic linear quadratic control problem, policy gradient method with the class of linear policies converges to the global optimum, despite non-convexity of the objective. Here, the authors provided an intricate analysis, leveraging a variety of closed form expressions available for linear-quadratic problems. Separate from the Reinforcement learning literature, [71] propose a stochastic approximation method for setting base-stock levels in inventory control. In this example too, the objective is non-convex but surprisingly, the authors manage to establish convergence to the globally optimal solution using a complex analysis quite different from that of [70]. How do we reconcile these success stories with the simple counterexample given in Example 1?

3.1.1 Our Contribution

Our work aims to construct a simple and more general understanding of the global convergence properties of policy gradient methods. As a consequence of our general framework, we can show that for several classic dynamic programming problems, policy gradient methods performed with respect to natural structured policy classes faces no suboptimal local minima. More precisely,

despite its non-convexity, any stationary point[2] of the policy gradient cost function is a global optimum. The examples we treat include:

1. Finite state and action MDPs with the class of all stochastic policies over the simplex.

2. Linear quadratic (LQ) control problems with the class of linear policies.

3. Optimal stopping problem with the class of threshold policies.

4. Finite horizon inventory control problem with the class of non-stationary base-stock policies.

These canonical control problems provide an important benchmark and sanity check for policy gradient methods. The central question we are interested in is *under what conditions do policy gradient methods converge to globally optimal solutions*? What is so special about the examples listed above and why do policy gradient methods fail for Example 1?

Interestingly, these examples share some important structural properties. Consider a linear quadratic control problem. Starting with a linear policy and performing a policy iteration step yields another linear policy. That is, the policy class is *closed* under policy improvement. In addition, although the cost-to-go function is non-convex, the policy iteration update involves just solving a quadratic minimization problem. These same properties apply to each of the first three examples. The policy class is closed under policy improvement and the policy iteration problem is solvable by first-order methods – it is either a convex optimization problem or can be easily shown to have no suboptimal stationary points. A strikingly simple proof shows that these two properties, together with mild regularity conditions, imply that any stationary point of the policy gradient loss function is globally optimal. We remark that our results also apply to the case of finite horizon problems with non-stationary policy classes (in particular, the finite horizon inventory control example) under weaker conditions; in this case we only require that the policy class contains an optimal policy.

[2]For unconstrained problems, stationary points of a function f satisfy $\nabla f(x) = 0$. More generally, for constrained optimization over some set $\mathcal{X}$, any stationary point x satisfies the first order necessary conditions for optimality, $\nabla f(x)^\top (x' - x) \geq 0 \ \forall x' \in \mathcal{X}$.

We also generalize these results to the case where the policy class is closed under approximate policy improvement — meaning that the policy iteration problem can be solved in the given policy class upto a small error. We interpret this closure assumption as a requirement that the policy class is sufficiently rich. Crucially, however, the (approximate) closure condition is much weaker than requiring the policy class to contain (approximately) all possible policies. This is useful, for example in problems where simple structured policy classes may be naturally aligned with the problem objective. However, note that the closure property is stronger than only requiring the policy class to contain (near) optimal policies. Indeed Example 1 shows that is necessary. In that case, even though the policy class contains the optimal policy, it is not closed under policy improvement and hence is susceptible to bad local minima.

Beyond studying the quality of stationary points, we also study stronger properties of the policy gradient objective that lead to fast converge rates. In particular, we show conditions under which the total cost function, $\ell(\cdot)$, satisfies a Polyak-lojasiewicz condition [72, 73, 74] (also popularly known as gradient dominance) which guarantees fast convergence rates for many first-order optimization algorithms for non-convex objectives. Essentially, for this case we require a weighted policy iteration objective to be gradient dominated.

Scope of this work. There are many reasons why practitioners may find simple policy gradient methods, like the classic REINFORCE algorithm [75], offer poor performance in practice. In an effort to clarify the scope of our contribution, and its place in the literature, let us briefly review some of these challenges.

1. *Non-convexity of the loss function:* Policy gradient methods apply (stochastic) gradient descent on a non-convex loss function. Such methods are usually expected to converge toward a stationary point of the objective function. Unfortunately, a general non-convex function could have many stationary points that are far from optimal.

2. *Unnatural policy parameterization:* It is possible for parameters that are far apart in Euclidean distance to describe nearly identical polices. Precisely, this happens when the Jacobian matrix

of the policy $\pi_\theta(\cdot \mid s)$ vanishes or becomes ill conditioned. Researchers have addressed this challenge through natural gradient algorithms [76, 77], which perform steepest descent in a different metric. The issue can also be alleviated with regularized policy gradient algorithms [62, 64].

3. *Insufficient exploration:* Although policy gradients are often applied with stochastic policies, convergence with this kind of naive random exploration can require a number of iterations that scales exponentially with the number of states in the MDP. [78] provide a striking example. Combining efficient exploration methods with policy gradients algorithms is challenging, but is an active area of research [see e.g. 79, 80].

4. *Large variance of stochastic gradients:* The variance of estimated policy gradients generally increases with the problem's effective time horizon, usually expressed in terms of a discount factor or the average length of an episode. Considerable research is aimed at alleviating this problem through the use of actor-critic methods [66, 81, 65] and appropriate baselines [63, 82].

We emphasize that this paper is primarily focused on the first challenge of studying the optimization landscape rather than analyzing particular policy gradient algorithms and their practical implementations. Such an investigation is relevant to many strategies for searching locally over the policy space, including policy gradient methods, natural gradient methods [77] , finite difference schemes [83], random search [84], and evolutionary strategies [85]. For most parts of this paper, we will imagine applying policy gradient algorithms in simulation, where an appropriate restart distribution ρ provides sufficient exploration and consider an idealized update with access to exact gradient evaluations, $\theta_{k+1} = \theta_k - \alpha_k \nabla \ell(\theta_k)$.

3.2 Further Related Literature

Beyond reinforcement learning, this work connects to a large body of work on first-order methods in non-convex optimization. Under broad conditions, these methods are guaranteed to converge asymptotically to stationary points of the objective function under a variety of noise models [86,

87]. The ubiquity of non-convex optimization problems in machine learning and especially deep learning has sparked a slew of recent work [88, 89, 90, 91] giving rates of convergence and ensuring convergence to approximate local minima rather than saddle points. A complementary line of research studies the optimization landscape of specific problems to essentially ensure that local minima are global, [92, 93, 94, 95, 96]. Taken together, these results show interesting non-convex optimization problems can be efficiently solved using gradient descent. Our work contributes to the second line of research, offering insight into the optimization landscape of $\ell(\cdot)$ for classic dynamic programming problems.

Related work along this direction includes the aforementioned work by [71] and [70]. For tabular MDPs with softmax policy parameterization, [97] gives a simple argument that the gradient of the policy gradient cost function is never exactly equal to zero. Work on conservative policy iteration [78] laid some intellectual groundwork for studying policy gradient methods. An under-appreciated paper by [98] extends the analysis of conservative policy iteration to study the stationary points of policy gradient methods. Relative to that work, our results regarding the quality of stationary points in Section 5 are more general as they deal with (1) problems with infinite action spaces and structured cost functions and (2) problems where the parameterized policy class is not convex (see remark 2).

Concurrently with this work, [99] provide a detailed study of the rate of convergence of policy gradient methods. Their work primarily focuses on natural gradient methods in problems with finite action spaces, both for tabular environments and larger state spaces where a (sufficiently accurate) function approximation architecture is employed. By contrast, our work gives a unified treatment of several foundational dynamic programming problems – reaching beyond tabular settings. This unified analysis seems to offer insight into when and why policy gradient methods can succeed despite non-convexity.

3.3 Problem formulation

Consider a Markov decision process (MDP), which is a six-tuple $(\mathcal{S}, \mathcal{A}, g, P, \gamma, \rho)$, consisting of a state space $\mathcal{S}$, action space $\mathcal{A}$, cost function g, transition kernel P, discount factor $\gamma \in (0, 1)$ and initial distribution ρ. We assume the state space $\mathcal{S}$ is at most countably infinite, in which case we can index the states as $\mathcal{S} = \{s_1, \cdots, s_n\}$ where n is possibly infinite. For each state $s \in \mathcal{S}$, let $\mathcal{A}_s \subset \mathbb{R}^k$ denote the set of feasible actions. We take $\mathcal{A} = \cup_s \mathcal{A}_s$. The transition kernel P specifies the probability $P(s'|s, a)$ of transitioning to a state s' upon choosing action a in state s. The cost function $g(s, a)$ denotes the instantaneous expected cost incurred when selecting action a in state s. We assume that per-period costs are uniformly bounded, meaning $\sup_{s\in\mathcal{S}, a\in\mathcal{A}_s} |g(s, a)| < \infty$.

Remark 1. *The assumptions that state spaces are at most countably infinite and per-period costs are bounded are standard in textbook treatments of dynamic programming [100, 101]. A rigorous study of dynamic programming with uncountable state spaces and infinite actions spaces is possible, but it requires restrictions that ensure various functions are measurable [102] or through the theory of universally measurable policies [103]. To avoid excessive technicality while being rigorous, our general results are stated for countably infinite state spaces. In specific problems like inventory control, there are no measurability issues and so we safety substitute infinite summations for integrals.*

Cost-to-go-functions and Bellman operators. A stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ is a function that prescribes a feasible action, $\pi(s) \in \mathcal{A}_s$ for each state $s \in \mathcal{S}$. Let Π denote the set of all stationary policies. Let $\mathcal{J} = \{J : \mathcal{S} \to \mathbb{R} : \|J\|_\infty < \infty\}$ denote the set of bounded functions on the state space. The cost-to go function $J_\pi \in \mathcal{J}$ associated with policy π is defined by $J_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t g(s_t, \pi(s_t))|s_0 = s\right]$. Define the Bellman operator $T_\pi : \mathcal{J} \to \mathcal{J}$ under the policy π as

$$(T_\pi J)(s) := g(s, \pi(s)) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s, \pi(s))J(s').$$

The cost-to-go function under policy π is the unique solution to the Bellman equation $J_\pi = T_\pi J_\pi$. The Bellman optimality operator is denoted by $T : \mathcal{J} \to \mathcal{J}$ and defined by $(TJ)(s) = \min_{\pi\in\Pi}(T_\pi J)(s)$. For simplicity of exposition, we assume throughout that this minimum is exists. The unique fixed point of T, denoted by J^*, is called the optimal cost-to-go function and satisfies $J^*(s) = \min_\pi J_\pi(s)$ for all $s \in \mathcal{S}$. There is at least one optimal policy, π^*, that attains this minimum for every $s \in \mathcal{S}$. It is well known that for uniformly bounded per-period costs, T and T_π are monotone and are contraction operators with respect to the maximum norm $\|\cdot\|_\infty$. Additional background is given in Appendix B.1.

The state-action cost-to-go function corresponding to a policy $\pi \in \Pi$,

$$Q_\pi(s,a) = g(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s' \mid s,a) J_\pi(s'), \tag{3.1}$$

measures the cumulative expected cost of taking action a in state s and applying π thereafter. The state-action value function provides an alternative notation for Bellman operators. Define $Q^*(\cdot,\cdot) = Q_{\pi^*}(\cdot,\cdot)$ for some optimal policy π^*. This is the same as (3.1) except the optimal cost-to-go function J^* appears on the right hand side. Notice that for any polices $\pi, \pi' \in \Pi$, we have the relations

$$Q_\pi(s,\pi(s)) = J_\pi(s), \qquad Q_\pi(s,\pi'(s)) = (T_{\pi'}J_\pi)(s), \qquad \min_{a\in\mathcal{A}_s} Q_\pi(s,a) = (TJ_\pi)(s). \tag{3.2}$$

Vector notation. We use vector and matrix notation that is standard in the study of Markov decision processes, including for problems with infinite state spaces. The readers can refer to [100, Sections 5.5, 6.1–6.2, Appendix C] for background. For each stationary policy $\pi \in \Pi$, we use the compact notation $g_\pi \in \mathcal{J}$ for the function $g_\pi(s) = g(s,\pi(s))$ for all $s \in \mathcal{S}$. We define P_π to be a transition operator associated with a policy π. It operates on elements $J \in \mathcal{J}$ on the right as

$$(P_\pi J)(s) = \sum_{s'\in\mathcal{S}} P(s'|s,\pi(s))J(s') = \mathbb{E}_\pi[J(s_1) \mid s_0 = s] \tag{3.3}$$

and operates on probability distributions ρ (viewed as row vectors) on the left as

$$(\rho P_\pi)(s) = \sum_{s' \in \mathcal{S}} \rho(s') P(s|s', \pi(s')) = \mathbb{P}_\pi(s_1 = s | s_0 \sim \rho) \tag{3.4}$$

We let P_π^t denote the t-step transition operator; the t-step counterparts to (3.3) and (3.4) can be naturally written as a composition. When the state space is finite, $P_\pi = (P(s'|s, \pi(s)))_{s,s' \in \mathcal{S}} \in \mathbb{R}^{n \times n}$ denotes the usual transition matrix under π. In this notation, one can write the Bellman operator concisely as $T_\pi J = g_\pi + \gamma P_\pi J$ and the cost to-go function as

$$J_\pi = g_\pi + \gamma P_\pi J_\pi = \cdots = \sum_{t=0}^{\infty} \gamma^t P_\pi^t g_\pi = (I - \gamma P_\pi)^{-1} g_\pi.$$

where $(I - \gamma P_\pi)^{-1}$ is invertible since γP_π has operator norm less than one.

Performance difference. A helpful "variational form" the Bellman equation [101] is

$$\begin{aligned} J_\pi - J = (T_\pi J - J) + (T_\pi J_\pi - T_\pi J) &= (T_\pi J - J) + \gamma P_\pi (J_\pi - J) \\ &= \cdots = (I - \gamma P_\pi)^{-1} (T_\pi J - J), \end{aligned} \tag{3.5}$$

for any $J \in \mathcal{J}$. A closely related expression is called the performance difference lemma in the reinforcement literature, after [78].

State distributions. We define the discounted state-occupancy measure under any policy π and initial state distribution ρ as:

$$\eta_\pi = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \rho P_\pi^t = (1 - \gamma) \rho (I - \gamma P_\pi)^{-1}. \tag{3.6}$$

When the state space is finite, η_π and ρ are both row vectors. Recognize that ρP_π^t is the distribution of the state at time t, so $\eta_\pi(s)$ evaluates the discounted fraction of time the system spends at state s under policy π.

Scalar loss function. Although classical dynamic programming methods seek a policy that minimizes the expected cost incurred *from every* state, for policy gradient methods it is more natural to study a scalar loss function

$$\ell(\pi) = (1-\gamma)\rho J_\pi = (1-\gamma)\sum_{s\in\mathcal{S}} J_\pi(s)\,\rho(s),$$

in which the states are weighted by their initial probabilities under ρ and we have normalized costs by $(1-\gamma)$ for convenience.

Exploratory initial distributions. We assume throughout that ρ is supported on the entire state space, i.e. $\rho(s) > 0$ for all $s \in \mathcal{S}$. Our results critically rely on the use of an exploratory initial distribution. This is not an artifact of the proof technique and it is well known that, in the absence of strong assumptions on the transition kernel[3], policy gradient methods have poor convergence properties if applied without an exploratory initial distribution. See Appendix B.3 for a full discussion. While this aspect of policy gradient methods is not always highlighted in the literature, many applied papers assume access to a diverse set of starting states using either explicit restarts [104, 105] or some form of continual learning that aims to increase the support of the training distribution [106, 107]. Since $\rho(s) > 0\ \forall\, s \in \mathcal{S}$, note that $\pi \in \arg\min_{\bar\pi} \ell(\bar\pi)$ if and only if $\pi \in \arg\min_{\bar\pi} J_{\bar\pi}(s)\ \ \forall\, s \in \mathcal{S}$.

Parameterized policies. Policy gradient methods search over a parameterized class of policies, $\Pi_\Theta = \{\pi_\theta(\cdot) : \theta \in \Theta\} \subset \Pi$ which have corresponding cost-to-go functions $\mathcal{J}_\Theta = \{J_{\pi_\theta} : \theta \in \Theta\}$. To indicate that we are referring to a policy in the restricted policy class, rather than an arbitrary stationary policy $\pi \in \Pi$, we typically either write π_θ or specify that $\pi \in \Pi_\Theta$. We assume that $\Theta \subset \mathbb{R}^d$ is convex and $\mathcal{A}_s \subset \mathbb{R}^k$ is convex for each $s \in \mathcal{S}$. In some cases, like inventory or linear quadratic control problems, the set of actions is naturally taken to be convex. In others, like MDPs

[3]Our assumption that $\rho(s) > 0$ was is used in our analysis to ensure that the state occupancy measure is lower bounded, meaning $\inf_\pi \eta_\pi(s) > 0$. Some authors assume the environment automatically drives the agent into each state.

with a finite set of base actions, the action set is convexified by taking $\mathcal{A} = \Delta^{k-1}$ to the probability simplex over k elements. See Example 3 in Section 3.5.

We overload notation, writing $\ell(\theta) = \ell(\pi_\theta)$. Policy gradient methods aim to minimize this loss function using gradient descent or a related first-order method. We later state results on smoothness conditions which ensure differentiability of $\ell(\cdot)$.

Norms. For our results, we often consider the weighted 1-norm, $\|J\|_{1,w} = \sum_s |J(s)|\, w(s)$ and the weighted maximum norm $\|J\|_{\infty,w} = \sup_{s\in\mathcal{S}} |J(s)|w(s)$ for some $w : \mathcal{S} \to \mathbb{R}_+$. When $w(s) = 1$ for all $s \in \mathcal{S}$, we simplify notation and write $\|J\|_1$ and $\|J\|_\infty$ respectively.

3.4 Convergence to stationary points in smooth optimization

Given that the policy gradient objective is almost always non-convex, optimization algorithms generally will not converge to a global minimum. Instead, classical theory suggests many algorithms converge to stationary points of the objective. This motivates our approach of studying the landscape of the policy gradient objective — and in particular the quality of its (approximate) stationary points — rather than studying convergence of specific algorithms.

We briefly review the theory of convergence to stationary points. A much more complete treatment can be found in nonlinear optimization textbooks [see e.g 108]. As made formal by the following definition, a stationary point satisfies the first-order necessary conditions for optimality. We say that a function has no-suboptimal stationary points if all such points are global minima. That is, the first-order necessary conditions are also sufficient. Note that, in unconstrained problems, where $\mathcal{X} = \mathbb{R}^d$, a stationary point x satisfies $\nabla f(x) = 0$.

Definition 1. *Consider the optimization problem* $\min_{x\in\mathcal{X}} f(x)$ *where* $\mathcal{X} \subset \mathbb{R}^d$ *is a closed convex set and* f *is continuously differentiable on an open set containing* $\mathcal{X}$*. A point* $x \in \mathcal{X}$ *is a stationary point*[4] *if* $\langle \nabla f(x), x' - x\rangle \geq 0$ *for all* $x' \in \mathcal{X}$ *. We say* $f(\cdot)$ *has no suboptimal stationary points if any stationary point* x *satisfies* $f(x) = \inf_{x'\in\mathcal{X}} f(x')$*.*

[4]These points are sometimes called first order stationary points, to distinguish them from points that also satisfy second order necessary conditions for optimality. Throughout this paper, we refer only to first order stationary points.

Under appropriate smoothness and regularity conditions, many optimization algorithms are guaranteed to converge to first-order stationary points. To make this concrete, we include an illustrative result that applies to the projected gradient descent algorithm. This result can be generalized in numerous ways. One strengthening of this result provides finite time bound on the rate of convergence to a stationary point, which we will review in Section 3.8. Another generalization, which we do not consider here, would treat stochastic noise in gradient evaluations. A rich literature on stochastic approximation shows that, under regularity conditions and appropriately decaying step-sizes, most noisy iterative algorithms converge to the same limit as their deterministic counterparts. See [32] for a very general treatment and [86] for a very readable introduction. A rapidly growing literature also studies convergence of stochastic first order methods in non-convex optimization [see e.g. 109, 110, 111, 112, 113, 114, 115, 116, 117].

The result below covers two cases. The first, which is standard in optimization literature [118], assumes ∇f is Lipschitz, or, for twice differentiable functions, that its Hessian is uniformly bounded. The second relaxes this condition, only requiring regularity properties on the sub-level set of the initial iterate. This is possible because projected gradient descent with sufficiently small step-sizes is guaranteed to reduce cost in each iteration, so all iterates lie in certain sub-level sets. The restriction that f has bounded sub-level sets is satisfied if the feasible region $\mathcal{X}$ is itself a bounded set or if the function is coercive, meaning $f(x) \to \infty$ as $\|x\| \to \infty$. In problems where this is not naturally satisfied, it can sometimes be enforced by adding a small penalty function (e.g. an entropy regularizer) to the objective. Recall that a point x_∞ is said to be a limit point of a sequence $\{x_k\}$ if some sub-sequence converges to x_∞.

Lemma 16. *Consider the optimization problem* $\min_{x\in\mathcal{X}} f(x)$ *where* $\mathcal{X} \subset \mathbb{R}^d$ *is a closed convex set. Assume* f *is bounded below and its* α*–sub-level sets* $\{x \in \mathcal{X} : f(x) \leq \alpha\}$ *are bounded for each* $\alpha \in \mathbb{R}$*. Consider the sequence* $x_{k+1} = \text{Proj}_{\mathcal{X}}\left(x_k - \alpha \nabla f(x_k)\right)$

1. *(Amir Beck, [119, 118]) Assume* f *is differentiable on an open set containing* $\mathcal{X}$ *and its gradient* ∇f *is Lipschitz continuous on* $\mathcal{X}$ *with Lipschitz constant* L*. If* $\alpha \in (0, 1/L]$*, the sequence* $\{x_k\}$ *has at least one limit point and any limit point* x_∞ *is a stationary point of* $f(\cdot)$

on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.

2. *Suppose $f(\cdot)$ is continuously twice differentiable on an open set containing the sub-level set $\{x \in \mathcal{X} : f(x) \leq f(x_0)\}$ and its gradient ∇f is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant L. Then, for a sufficiently small $\alpha > 0$, the sequence $\{x_k\}$ has at least one limit point and any limit point x_∞ is a stationary point of $f(\cdot)$ on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.*

Proof. Proof for the second part closely follows proof for the first part as shown in [119, 118] with slight modifications. For completeness, we provide a sketch in Appendix B.2. □

3.5 Closed policy classes and the optimality of stationary points

We hope for results that suggest local policy search may succeed when the policy class is well suited to the decision objective. In the next subsection, we look at the example of linear quadratic control and identify structural properties that ensure policy gradient methods avoid bad local minima despite non-convexity of the objective. With this as motivation, we proceed to show these structural properties ensure the policy iteration loss function has no suboptimal stationary points more broadly. Beyond linear quadratic control, we instantiate this theory for finite state and action MDPs as well as for stopping problems with threshold policies.

3.5.1 Motivation from linear quadratic control

Leveraging many of the closed form expressions available for linear quadratic (LQ) control, recent work of [70] showed that policy gradient methods with linear policies converge to the globally optimal policy under some technical conditions. This is true even though the total expected cost incurred is a nasty non-convex function. The key to this result is that the loss function $\ell(\cdot)$ has no suboptimal stationary points (and in fact a stronger gradient dominance property holds). Given the failure of policy gradient methods in Example 1, there must be some special problem structure driving this, but what?

We identify two key properties for LQ control. First, as highlighted in Equation (3.8) below,

the class of linear policies is *closed under policy improvement steps*. Second, the policy iteration problem as shown below in (3.10) could be solved to optimality by a gradient method, since it is convex quadratic and hence has no suboptimal stationary points. Both properties relate to the nice structure of the policy iteration objective in (3.10) which considers the impact of changing the policy for only one time step. By contrast, the infinite horizon cost function $\ell(\cdot)$ is non-convex and difficult to analyze. Our approach will be to show that, despite non-convexity, $\ell(\cdot)$ has nice optimization structure as an immediate consequence of the two properties we've identified, leading to a very simple understanding of policy gradient methods.

Like [70], we simplify the presentation by studying deterministic linear quadratic control[5]. It is easy to extend our ideas to noisy dynamics of the form $s_{t+1} = As_t + B_a + \zeta_t$ for i.i.d noise ζ_t with zero mean and finite second moment.

Example 2 (Linear Quadratic Control)**.** *For symmetric positive definite matrices R and C, we face the optimal control problem:*

$$\begin{aligned} \text{Minimize} \quad & \sum_{t=0}^{\infty} \gamma^t \left(a_t^\top R a_t + s_t^\top C s_t \right) \\ \text{Subject to} \quad & s_{t+1} = As_t + Ba_t, \quad s_0 \sim \rho \end{aligned}$$

where $s_t \in \mathbb{R}^n$ is a continuous state variable and $a_t \in \mathbb{R}^k$ is the action chosen at time t. We assume per-step costs to be finite, $\|R\|_2, \|C\|_2 < \infty$, and that the second moment of the initial distribution, $\mathbb{E}_{s\sim\rho}\left[ss^\top\right]$, is finite and positive definite.

A linear policy $\pi_\theta(s) = \theta s$ is known to be optimal for some $\theta \in \mathbb{R}^{k\times n}$, see for example [101, 120, 121]. We consider the search for optimal θ via a gradient method. Unfortunately, the loss function $\ell(\theta) = \mathbb{E}_{s\sim\rho}\left[J_{\pi_\theta}(s)\right]$ is non-convex (see Appendix B in [70] for a simple example), making it unclear if gradient descent on $\ell(\theta)$ would reach the global minimum.

For LQ control, if a linear policy π_θ is applied from a state s_0 then from the linear dynamics

[5]This choice has several benefits. First, it gives an easy expression $s_t = (A + B\theta)^t s_0$ for the state evolution helping readers see the source of non-convexity in $J_{\pi_\theta}(s)$. Second, in the noisy case, the cost-to-go functions have an additional constant term, $J_{\pi_\theta}(s) = s^\top K_\theta s + \mathbb{E}\left[\zeta^\top K_\theta \zeta\right]$, as compared to the noiseless case, simplifying some expressions.

we have $s_t = (A + B\theta)s_{t-1} = \cdots = (A + B\theta)^t s_0$. *From this, we can write the cost-to-go function as*

$$J_{\pi_\theta}(s_0) = \sum_{t=0}^{\infty} \gamma^t \left(s_t^\top \theta^\top R\theta s_t + s_t^\top C s_t\right) = s_0^\top \underbrace{\left[\sum_{t=0}^{\infty} \gamma^t \left((A + B\theta)^t\right)^\top \left(\theta^\top R\theta + C\right)(A + B\theta)^t\right]}_{:=K_\theta} s_0$$

A linear policy π_θ, *is said to be stable if all eigenvalues of the matrix* $\sqrt{\gamma}(A+B\theta)$ *lies strictly within the unit circle. Let* $\Theta_S \subset \mathbb{R}^{k\times n}$ *denote the set of all parameters defining stable linear policies.*

$$\Theta_S := \left\{\theta \in \mathbb{R}^{k\times n} \mid \max_{x:\|x\|_2\leq 1} \left\|\sqrt{\gamma}(A + B\theta)x\right\|_2 < 1\right\}$$

For a stable linear policy, K_θ *is finite and positive definite, ensuring that the cost-to-go is finite. We assume the system* (A, B) *is controllable so there exists at least one stable policy. Note that when* $\gamma = 1$ *our definition reduces to the more standard definition of a stable linear policy in undiscounted problems. See [101] for further discussion of stability in discounted linear quadratic control.*

Even though the total cost function $\ell(\theta)$ *is non-convex, classical dynamic programming theory applies to the policy iteration algorithm for LQ control. This dates back at least to the works of [122] and [123], who showed that even in the undiscounted case, beginning with a stable linear policy, policy iteration produces a sequence of stable linear policies with strictly improving cost-to-go and converges to an optimal policy. Essentially, the complexity in analyzing* $\ell(\theta)$ *is due to its multi-period nature where any changes in the control,* θ, *have a compounding influence on states visited far out into the future. On the other hand, policy iteration converges to an optimal policy by solving a sequence of much simpler single period decision problems. Beginning with a stable linear policy* $\pi_\theta(s) = \theta s$, *applying the Bellman operator involves solving a quadratic optimization*

problem, making it easy to plan over single period:

$$
\begin{aligned}
(TJ_{\pi_\theta})(s) &= \min_{a \in \mathbb{R}^k} \left[a^\top R a + s^\top C s + \gamma J_{\pi_\theta}(As + Ba) \right] \\
&= \min_{a \in \mathbb{R}^k} \left[\underbrace{a^\top R a + s^\top C s + \gamma (As + Ba)^\top K_\theta (As + Ba)}_{= Q_{\pi_\theta}(s,a)} \right]. \qquad (3.7)
\end{aligned}
$$

It is easy to see that for any state s, a minimizing action is $a^* = -\gamma(R + \gamma B^\top K_\theta B)^{-1} B^\top K_\theta A s$. Therefore, the linear feedback policy $\pi_{\overline{\theta}}$ with $\overline{\theta} = -\gamma(R + \gamma B^\top K_\theta B)^{-1} B^\top K_\theta A$ is a policy iteration update. In terms of Bellman operators, this can be expressed as

$$
T_{\pi_{\overline{\theta}}} J_{\pi_\theta} = T J_{\pi_\theta} \qquad (3.8)
$$

where the Bellman operator corresponding to a linear policy is given by

$$
(T_{\pi_{\overline{\theta}}} J)(s) = (\overline{\theta} s)^\top R (\overline{\theta} s) + s^\top C s + \gamma J(As + B\overline{\theta} s). \qquad (3.9)
$$

Thus, one can perform policy iteration for LQ control by searching over the restricted class of linear policies. In other words, for LQ control, the class of linear polices is closed under policy iteration which is the first key property we identified in the beginning of this section.

We can equivalently write (3.8) *in terms of Q-functions as: $Q_{\pi_\theta}(s, \overline{\theta} s) = \min_a Q_{\pi_\theta}(s, a)$ for all s. For a given s, this optimization problem depends on $\overline{\theta}$ only though $\overline{\theta} s$ and hence there is an entire subspace of optimal solutions. The solution becomes unique either by requiring optimality at a set of states that span $\mathbb{R}^n$ or by solving the weighted policy iteration problem*

$$
\min_{\overline{\theta}} \ \mathbb{E}_{s \sim \eta} \left[Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) \right], \qquad (3.10)
$$

where $\mathbb{E}_{s \sim \eta}[s s^\top]$ has full rank. Since $Q_{\pi_\theta}(s, a)$ is a convex quadratic function[6] *of a, $Q_{\pi_\theta}(s, \overline{\theta} s)$ is*

[6]Cost matrices R, C are assumed to be positive definite and it is easy to show that $K_\theta > 0$ for all $\theta \in \Theta_S$.

a quadratic function of $\overline{\theta}$ (viewing $\overline{\theta}$ is a stacked vector) and this property is preserved by taking the expectation over s. This shows the weighted policy iteration problem in (3.10) *is strongly convex and hence can be solved efficiently by a gradient method which is the second key property we identified.*

Before giving a simple analysis of the stationary points of the policy gradient loss functions, we need some basic preliminaries that are specialized to linear quadratic control. First, notice that the LQ control example does not fit within our general problem formulation, which assumed single period costs to be uniformly bounded. For linear quadratic control, costs tend to infinity as the norm of the state grows and unstable policies will have infinite cost-to-go from some initial states. However, by restricting to quadratic cost-to-go functions and stable linear policies, important properties of Bellman operators still hold. The following lemma states the three properties we use. These are standard results[7] and the proofs is omitted. Define the set of strictly convex quadratic cost-to-go functions, as

$$\mathcal{J}_q = \{J : s \in \mathbb{R}^n \mapsto s^\top K s \,\Big|\, K \in \mathbb{R}^{n \times n}\,,\, K \succ 0\,\}.$$

Lemma 17 (Bellman operators for LQ control)**.** *Consider the linear quadratic control problem formulated in Example 2. For $J, \bar{J} \in \mathcal{J}_q$ and a stable linear policy π_θ, the following properties hold:*

1. *(Closure on the set of quadratic cost functions) $T_{\pi_\theta} J \in \mathcal{J}_q$ and $TJ \in \mathcal{J}_q$.*
2. *(Monotonicity) If $J \preceq \bar{J}$ then $T_{\pi_\theta} J \preceq T_{\pi_\theta} \bar{J}$ and $TJ \preceq T\bar{J}$.*
3. *(Bellman equation) $J_{\pi_\theta} = T_{\pi_\theta} J_{\pi_\theta}$ and $J_{\pi_\theta} = \lim_{k \to \infty} T_{\pi_\theta}^k J$. Moreover, $J = TJ$ if and only if $J = J^*$.*

[7] Many references state such results in terms of the cost matrices instead of the functions $J \in \mathcal{J}_q$. For example, the uniqueness of solutions to the Bellman optimality equation within $\mathcal{J}_q$ is identical to the more common statement that the algebraic Riccatti equation has a unique positive definite solution.

The next lemma establishes smoothness properties of $\ell(\cdot)$ on sub-level sets. If initialized with some stable linear policy θ_0, first order algorithms that are descent methods — meaning they reduce the cost function in each iteration — are assured to stay within the sub-level set, $C_{\ell(\theta_0)}$, on which these properties are satisfied. Known results in optimization, like Lemma 16, then imply that gradient descent on $\ell(\cdot)$ with sufficiently small stepsizes will converge to a first-order stationary points. Some additional details are provided in Appendix B.4, but the properties in this lemma are known in the control theory literature [124, 125].

Lemma 18. *Consider the linear quadratic control problem formulated in Example 2. The set Θ_S is open and ℓ is twice continuously differentiable on Θ_S. For any $\alpha \in \mathbb{R}$, the sublevel set $C_\alpha := \left\{\theta \in \mathbb{R}^{n\times k} : \ell(\theta) \leq \alpha\right\}$ is a compact subset of Θ_S and $\sup_{\theta\in C_\alpha} \|\nabla^2\ell(\theta)\| < \infty$.*

With this background in place, a simple proof shows that for LQ control, the policy gradient loss function has no suboptimal stationary points despite being non-convex. Starting from a suboptimal stable linear policy π_θ, the policy iteration step produces a linear policy $\pi_{\overline{\theta}}$ with reduced cost, i.e. $\|K_{\overline{\theta}}\|_2 < \|K_\theta\|_2 < \infty$ which is stable (following Lemma 18). Using the convexity of the policy iteration objective along with standard dynamic programming arguments, we show that a soft-policy iteration update forms a descent direction for loss $\ell(\cdot)$, implying that θ is not a stationary point. The argument here is strongly reminiscent of the standard analysis of policy iteration.

Lemma 19. *For the linear quadratic control problem formulated in Example 2, any stable linear policy θ satisfies $\nabla\ell(\theta) = 0$ if and only if $J_{\pi_\theta} = J^*$.*

Proof. Consider a stable linear policy π_θ and take $\pi_{\overline{\theta}}$ to be a policy iteration update to π_θ as shown in (3.8). Set $\theta^\alpha = (1-\alpha)\theta + \alpha\overline{\theta}$ for $\alpha \in [0,1]$, which implies that $\pi_{\theta^\alpha}(s) = (1-\alpha)\theta s + \alpha\overline{\theta}s$ (both

π_θ and $\pi_{\overline{\theta}}$ are linear policies). For every $s \in \mathbb{R}^n$,

$$\begin{aligned} T_{\pi_{\theta^\alpha}} J_{\pi_\theta}(s) = Q_{\pi_\theta}(s, \pi_{\theta^\alpha}(s)) = Q_{\pi_\theta}(s, (1-\alpha)\theta s + \alpha\overline{\theta} s) &\leq (1-\alpha) Q_{\pi_\theta}(s, \theta s) + \alpha Q_{\pi_\theta}(s, \overline{\theta} s) \\ &= (1-\alpha) T_{\pi_\theta} J_{\pi_\theta}(s) + \alpha T_{\pi_{\overline{\theta}}} J_{\pi_\theta}(s) \\ &= (1-\alpha) J_{\pi_\theta}(s) + \alpha T J_{\pi_\theta}(s) \\ &= J_{\pi_\theta}(s) - \alpha \left(J_{\pi_\theta}(s) - T J_{\pi_\theta}(s) \right) \end{aligned}$$

where the inequality uses that $a \mapsto Q_{\pi_\theta}(s, a)$ is convex as shown in the discussion of Example (2). Repeatedly applying the Bellman operator and using the monotonicity property, $J_{\pi_\theta} \leq T J_{\pi_\theta}$, shown in Lemma 17 gives:

$$J_{\pi_\theta} \geq T_{\pi_{\theta^\alpha}} J_{\pi_\theta} \geq T^2_{\pi_{\theta^\alpha}} J_{\pi_\theta} \geq \cdots \geq J_{\pi_{\theta^\alpha}}.$$

Notice that π_{θ^α} is stable following Lemma 18, since its cost-to-go is always lower than that of π_θ. From this, we have

$$\frac{J_{\pi_{\theta^\alpha}} - J_{\pi_\theta}}{\alpha} \leq \frac{T_{\pi_{\theta^\alpha}} J_{\pi_\theta} - J_{\pi_\theta}}{\alpha} \leq \left[T J_{\pi_\theta} - J_{\pi_\theta} \right].$$

Multiplying each side by $1 - \gamma$, taking the expectation over s drawn from the initial distribution ρ, and then taking $\alpha \to 0$ gives

$$\frac{d}{d\alpha} \ell(\theta^\alpha) \bigg|_{\alpha=0} \leq (1-\gamma) \mathbb{E}_{s\sim\rho} \left[T J_{\pi_\theta}(s) - J_{\pi_\theta}(s) \right].$$

Consider the error in Bellman's equation $E(s) \triangleq T J_{\pi_\theta}(s) - J_{\pi_\theta}(s)$. We know $E(s) \leq 0$ for all s. Since θ is suboptimal, $E(s) < 0$ for some state s. We use the assumption that $\Sigma := \mathbb{E}_{s\sim\rho}[ss^\top] \succ 0$ to show that $\mathbb{E}_{s\sim\rho}[E(s)] < 0$. To see this, note that as $J_{\pi_\theta}, T J_{\pi_\theta} \in \mathcal{J}_q$, $E(s)$ is the difference in quadratic functions and can be written as $E(s) = s^\top K s$ for some symmetric $K \preceq 0$ with $K \neq 0$. Taking the spectral decomposition, we can write $K = \sum_{i=1}^n \lambda_i q_i q_i^\top$, where $(q_1, \cdots, q_n)$ is an orthonormal basis of eigenvectors, $\lambda_i \leq 0$ and $\min_i \lambda_i < 0$. Therefore,

$\mathbb{E}_{s\sim\rho}[E(s)] = \sum_{i=1}^n \lambda_i \mathbb{E}_{s\sim\rho}[(q_i^\top s)^2] = \sum_{i=1}^n \lambda_i q_i^\top \Sigma q_i < 0$ as desired. □

Remark 2. *This argument of constructing an improved policy using soft policy iteration updates is strongly reminiscent of the idea underlying Conservative Policy Iteration (CPI) algorithm of [78] for finite MDPs. A similar argument is given in [98] to study the stationary points of the policy gradient loss function in problems with finite action spaces and convex policy classes. The argument here is more general as it applies to problems with infinite action spaces and structured cost functions.*

While we find this proof to be intuitive, it relies not just on the closure property in (3.10) *but on convexity of the policy class and on convexity of the policy iteration cost function* (3.10)*, which will not hold in all of our examples*[8]*. One contribution of this paper is to find a clean generalization of this intuitive but limited proof strategy, relaxing convexity conditions into Condition 2 in the next subsection.*

3.5.2 General results

Let us generalize (3.10) by introducing the weighted policy iteration, or "Bellman" objective,

$$\mathcal{B}(\pi' \mid \eta, J_\pi) = \mathbb{E}_{s\sim\eta}\left[(T_{\pi'}J_\pi)(s)\right] = \mathbb{E}_{s\sim\eta}\left[Q_\pi(s,\pi'(s))\right], \tag{3.11}$$

and overload notation to write $\mathcal{B}(\theta \mid \eta, J_\pi) = \mathcal{B}(\pi_\theta \mid \eta, J_\pi)$. Recall that $(T_{\pi'}J_\pi)(s) \equiv Q_\pi(s,\pi'(s))$, which is the reason for the second equality in (3.11).

In order to speak meaningfully about the convergence of policy gradient methods, some regularity conditions of the loss function $\ell(\theta)$ are needed. The following states smoothness conditions, related to partial differentiability of $\mathcal{B}(\cdot)$, that ensure $\ell(\cdot)$ is differentiable and its gradients satisfy a convenient formula used in practical implementations [126, 65, 1]. Subsection 3.5.3 will show these conditions arise quite naturally and will discus basic smoothness conditions on the problem under which it holds. We call this condition 0, because it is not related to global convergence of

[8]For example, the class of threshold policies, used in the optimal stopping problem in Example 4, is not convex. If $\pi_\theta(s) = \mathbf{1}(s \le \theta)$ for $\theta \in \mathbb{R}$ is a threshold policy, then $\frac{1}{2}\pi_\theta + \frac{1}{2}\pi_{\theta'}$ is not a threshold policy when $\theta \neq \theta'$.

policy gradient so much as whether the algorithms themselves are even well defined.

Condition 0 (Differentiability)**.** *For each $\theta \in \Theta$, the functions $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta}|\eta_{\pi_\theta}, J_{\pi_\theta})$ and $\overline{\theta} \mapsto \mathcal{B}(\theta|\eta_{\pi_{\overline{\theta}}}, J_{\pi_\theta})$ are continuously differentiable on an open set containing θ.*

Clearly, when $\eta(s) > 0 \; \forall \, s \in \mathcal{S}$, minimizing (3.11) over all policies π' is equivalent to classical policy iteration. Policy gradient methods are more closely related to the following weighted policy iteration scheme:

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \mathcal{B}(\theta \mid \eta_{\pi_{\theta_t}}, J_{\pi_{\theta_t}}),$$

which preforms policy iteration style updates over the parameterized policy class, Π_Θ. In general, this scheme may chatter endlessly. But, it is assured to converge to an optimal policy when the policy class is closed under policy improvement. As explained in the introduction, this condition is stronger than the requirement that Π_Θ contains an optimal policy. However, this condition is necessary, since Example 1 shows policy gradient methods can get stuck in bad local minima even in extremely simple examples in which Π_Θ contains an optimal policy but is not closed under policy improvement. Note that Condition 1 is much weaker than requiring the policy class is rich enough to contain nearly all policies, accommodating examples in which a certain class of policies is naturally aligned with the decision task, for example linear policies in LQ control or threshold policies in optimal stopping.

Condition 1 (Closure under policy improvement)**.** *For each $\pi \in \Pi_\Theta$, there exists $\pi^+ \in \Pi_\Theta$ such that $T_{\pi^+} J_\pi = T J_\pi$. Equivalently, $\mathcal{B}(\pi^+|\eta, J_\pi) = \min_{\pi' \in \Pi} \mathcal{B}(\pi'|\eta, J_\pi)$ for each probability distribution η over $\mathcal{S}$.*

As a first order method, policy gradients require additional local optimization structure to succeed. The following condition ensures that the weighted policy iteration problem is amenable to first-order optimization. It is worth emphasizing that the total cost function $\ell(\theta)$ is complicated and non-convex in all the examples we consider. This is due to the multi-period nature of the decision

problem, in which changes to the policy can have a compounding effect[9] over time. On the other hand, the weighted policy iteration objective $\bar{\theta} \mapsto \mathcal{B}(\bar{\theta}|\eta_{\pi_\theta}, J_{\pi_\theta})$ is typically much simpler as it considers only a single period decision problem and often has a nice structure.

Condition 2 (Stationary points of the policy iteration objective)**.** *For each $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ has no sub-optimal stationary points.*

Recall, we already noted that $Q_\pi(s, \pi_\theta(s))$ is a convex quadratic function of θ for the LQ control problem, which by definition implies that $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ is convex quadratic. Similarly, for finite MDPs, we show it to be linear in Section 3.5.5. We also verify Condition 2 for the optimal stopping problem in Section 3.5.5.

The next theorem offers a broad generalization of the result for LQ control shown in Lemma 19. After developing some supporting results in the next subsection, we prove Theorem 5 in Section 3.5.4.

Theorem 5. *Suppose Conditions 0, 1, and 2 hold. Then, ℓ is continuously differentiable and $\theta \in \Theta$ is a stationary point of $\ell(\cdot)$ if and only if $J_{\pi_\theta} = J^*$.*

3.5.3 A sharp connection between policy gradient and weighted policy iteration

The key to our approach is a sharp relationship between between policy gradient methods and the weighted policy iteration scheme

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \mathcal{B}(\theta \mid \eta_{\pi_{\theta_t}}, J_{\pi_{\theta_t}}), \tag{3.12}$$

which performs policy iteration style updates over Π_Θ. In light of the policy gradient theorem below, when $\Theta = \mathbb{R}^d$ is unconstrained, gradient descent for $\ell(\theta)$ with a constant stepsize α can be shown to be equivalent to gradient updates with the weighted policy iteration objective.

$$\theta_{t+1} = \theta_t - \alpha \nabla_{\bar{\theta}} \mathcal{B}(\bar{\theta} \mid \eta_{\pi_{\theta_t}}, J_{\pi_{\theta_t}})\Big|_{\bar{\theta}=\theta_t}.$$

[9] In terms of the distribution of states and actions visited over a trajectory under the updated policy vs the old policy.

Policy gradients and weighted policy iteration differ in two ways. First, policy gradient methods are incremental, making a parameter update based on a gradient of (3.11) rather than solving it exactly. Second, the state-relevance weights η are updated over time to reflect the frequency of states visited by the current policy. This idea is central to our main results presented subsequently. Under suitable regularity conditions, allowing the exchange of differentiation and integration, gradients of $\mathcal{B}$ can be rewritten as

$$\nabla_\theta \mathcal{B}(\theta|\eta_\pi, J_\pi) = \nabla_\theta \mathbb{E}_{s\sim\eta_\pi}\left[Q_\pi(s, \pi_\theta(s))\right] = \mathbb{E}_{s\sim\eta_\pi}\left[\nabla_\theta Q_\pi(s, \pi_\theta(s))\right].$$

It is worth mentioning that, although policy gradient theorems are widely used in reinforcement learning [66, 126, 127], our very general presentation of this result and the short proof does not seem to be widely understood in the literature.

Lemma 20 (Policy gradient theorem)**.** *Under Condition 0, $\ell(\theta)$ is continuously differentiable and*

$$\nabla \ell(\theta) = \nabla_{\overline{\theta}} \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})\Big|_{\overline{\theta}=\theta}.$$

Proof. Denote $J_\theta \equiv J_{\pi_\theta}$, $P_\theta \equiv P_{\pi_\theta}$ and $\eta_\theta \equiv \eta_{\pi_\theta}$. The variational Bellman equation in (3.5) states

$$J_{\overline{\theta}} - J_\theta = \sum_{t=0}^{\infty} \gamma^t P_{\overline{\theta}}^t (T_{\overline{\theta}} J_\theta - J_\theta).$$

Recall the definition of the discounted state occupancy measure, $\eta_{\overline{\theta}} = \rho \sum_{t=0}^{\infty} \gamma^t P_{\overline{\theta}}^t$, in (3.6) and the

definition of the policy iteration objective in (3.11). Then,

$$\begin{aligned}
\ell(\overline{\theta}) - \ell(\theta) = (1-\gamma)\rho\,(J_{\overline{\theta}} - J_\theta) &= (1-\gamma)\rho \sum_{t=0}^{\infty} \gamma^t P_{\overline{\theta}}^t\,(T_{\overline{\theta}} J_\theta - J_\theta) \\
&= (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \rho P_{\overline{\theta}}^t\,(T_{\overline{\theta}} J_\theta - J_\theta) \\
&= \sum_{s\in\mathcal{S}} \eta_{\overline{\theta}}(s)(T_{\overline{\theta}} J_\theta - J_\theta)(s) \\
&= \mathcal{B}(\overline{\theta}|\eta_{\pi_{\overline{\theta}}},\, J_{\pi_\theta}) - \mathcal{B}(\theta|\eta_{\pi_{\overline{\theta}}},\, J_{\pi_\theta})
\end{aligned}$$

Expanding the total derivative in terms of partial derivatives gives,

$$\nabla \ell(\overline{\theta}) = \nabla_{\overline{\theta}}\, \mathcal{B}(\overline{\theta}|\eta_{\overline{\theta}},\, J_\theta)\Big|_{\overline{\theta}=\theta} - \nabla_{\overline{\theta}}\, \mathcal{B}(\theta|\eta_{\overline{\theta}},\, J_\theta)\Big|_{\overline{\theta}=\theta} = \nabla_{\overline{\theta}}\, \mathcal{B}(\overline{\theta}|\eta_\theta,\, J_\theta)\Big|_{\overline{\theta}=\theta}.$$

□

Although Condition 0 may initially appear unnatural, the proof shows that it arises naturally in calculating the derivative of $\ell(\theta)$. Let us briefly remark on what is required for the condition to hold. The differentiability of $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta}|\eta_{\pi_\theta}, J_{\pi_\theta})$ holds if $\nabla_{\overline{\theta}} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))$ exists at every $s \in \mathcal{S}$ and $\mathbb{E}_{s\sim\eta_{\pi_\theta}} \left|\frac{\partial}{\partial \overline{\theta}_i} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))\right| < \infty$. The second condition, that $\overline{\theta} \mapsto \mathcal{B}(\theta|\eta_{\pi_{\overline{\theta}}}, J_{\pi_\theta})$ is related to the existence of derivative of the state occupancy measure [128, 129].

There is a substantial literature in applied probability that studies technical sufficient conditions which imply differentiability and the validity of particular estimators. See [130, 131, 132] for broad introductions. We do not try to advance that literature, instead focusing on the convergence of policy gradient methods when they are well defined.

3.5.4 Proof of Theorem 5

We first give a key lemma, which can be viewed as a Bellman-type equation that holds when the single period objective $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})$ has no bad stationary points.

Lemma 21. *Suppose Condition 2 is satisfied. If θ is a stationary point of $\ell : \Theta \to \mathbb{R}$, then*

$$\mathbb{E}\left[J_{\pi_\theta}(S)\right] = \min_{\pi \in \Pi_\Theta} \mathbb{E}\left[(T_\pi J_{\pi_\theta})(S)\right],$$

where the expectation is over S drawn from η_{π_θ}.

Proof. If θ is a stationary point of $\ell : \Theta \to \mathbb{R}$, then by Lemma 20, it is a stationary point of the function $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})$. Since Condition 2 holds, this means

$$\mathcal{B}(\theta \mid \eta_{\pi_\theta}, J_{\pi_\theta}) = \min_{\overline{\theta} \in \Theta} \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta}).$$

Recalling the definition of $\mathcal{B}(\theta \mid \eta, J_\pi)$ in (3.11) lets us rewrite both sides of this equation. To simplify the expressions, take S to be a random state drawn from η_{π_θ}. Then,

$$\mathbb{E}\left[J_{\pi_\theta}(S)\right] = \mathbb{E}\left[(T_{\pi_\theta} J_{\pi_\theta})(S)\right] = \mathcal{B}(\theta \mid \eta_{\pi_\theta}, J_{\pi_\theta}) = \min_{\bar{\theta} \in \Theta} \mathcal{B}(\bar{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta}) = \min_{\bar{\theta} \in \Theta} \mathbb{E}\left[(T_{\pi_{\bar{\theta}}} J_{\pi_\theta})(S)\right].$$

□

We now state an "average" form of Bellman's equation in Lemma 22, which holds due to our assumption that the initial distribution ρ places positive probability on every state, ensuring that $\eta_\pi(s) \geq (1-\gamma)\rho(s) > 0$ for all $s \in \mathcal{S}$ and $\pi \in \Pi_\Theta$. Using this, the average Bellman equation reduces to a standard result in dynamic programming which argues that satisfying the Bellman's equation is necessary and sufficient for optimality, i.e. $J_\pi = J^* \iff J_\pi = TJ_\pi$. For completeness, we give a proof in Appendix B.1.

Lemma 22 (On average Bellman equation)**.** *For any $\pi \in \Pi$ and $S \sim \eta_\pi$,*

$$J_\pi = J^* \iff \mathbb{E}[J_\pi(S)] = \mathbb{E}[TJ_\pi(S)]$$

The proof of Theorem 5 now follows as an immediate consequence of the closure assumption as stated in Condition 1.

Completing the proof of Theorem 5. Suppose θ is a stationary point of $\ell(\cdot)$. We have

$$\mathbb{E}\left[J_{\pi_\theta}(S)\right] = \min_{\pi \in \Pi_\Theta} \mathbb{E}\left[T_\pi J_{\pi_\theta}(S)\right] = \mathbb{E}\left[T J_{\pi_\theta}(S)\right]$$

where the first equality uses Condition 2 to invoke Lemma 21 and the second equality uses Condition 1. Finally, Lemma 22 shows that satisfying the average Bellman equation implies optimality. □

3.5.5 Examples beyond LQ control

Having looked at the LQ control example in detail, we describe two other problem settings to which our results will apply. We show how Conditions 1 and 2 continue to hold for these problems as well.

Example 3 (Finite state action MDPs)**.** *Consider a problem with finite number of states, $\mathcal{S} = \{1, \cdots, n\}$. For notational simplicity, assume the set of feasible actions $\mathcal{A}_s$ is the same for every state s and denote this by $\mathcal{A}$. We also assume there is a finite set of k deterministic actions to choose from and take $\mathcal{A} = \Delta^{k-1}$ to be the set of all probability distributions over these actions. That is, any action $a \in \mathcal{A}$ is a probability vector where each component a_i denotes the probability of taking the i^{th} action. Cost and transition functions can be naturally extended to functions on the probability simplex by defining:*

$$g(s,a) = \sum_{i=1}^{k} g(s,e_i)\, a_i \qquad\qquad P(s'|s,a) = \sum_{i=1}^{k} P(s'|s,e_i)\, a_i. \tag{3.13}$$

where e_i is the i-th standard basis vector, representing one of the k possible deterministic actions.

For this tabular setting, a natural parametrization considers the policy $\pi_\theta(s) = \theta_s \in \Delta^{k-1}$ which associates each state with a probability distribution over actions. Rather than track the policy parameter $\theta = (\theta_s : s = 1, \cdots, n) \in \mathbb{R}^{n \times k}$ we work directly with a stochastic policy $\pi \in \mathbb{R}^{n \times k}$, viewed as a matrix whose rows are probability vectors. In this case, the set of all stationary randomized policies can be written as $\Pi = \{\pi \in \mathbb{R}_+^{n \times k} : \sum_{i=1}^{k} \pi_{s,i} = 1 \ \forall s \in \{1, \cdots, n\}\}$.

Since Π contains all stationary policies, it is clearly closed under policy improvement. It is also worth noting that for any $\pi \in \Pi, s \in \mathcal{S}$ and $a \in \Delta^{k-1}$, the Q-function is linear in a, as we can write: $Q_\pi(s,a) = \sum_{i=1}^k Q_\pi(s,e_i)a_i = \langle Q_\pi(s,\cdot), a\rangle$. Therefore, the weighted policy iteration objective,

$$\mathcal{B}(\pi'|\eta_\pi, J_\pi) = \mathbb{E}_{s\sim\eta_\pi}\left[Q_\pi\left(s,\pi'(s)\right)\right]$$

is convex (linear) in π' and can be optimized efficiently using projected gradient descent, for example.

Remark 3 (Regularized softmax policies)**.** *For tabular MDPs, it is quite common to use a softmax policy parameterized by $\theta \in \mathbb{R}^{n\times k}$ where for any state s, the policy $\pi_\theta(s) \in \Delta^{k-1}$ is a probability distribution with components $\pi_\theta(s) \equiv (\pi_\theta(s,1), \cdots, \pi_\theta(s,k))$ such that*

$$\pi_\theta(s,i) = \frac{e^{\theta_{s,i}}}{\sum_{j=1}^k e^{\theta_{s,j}}} \quad i = 1, \cdots, k$$

To avoid over parameterization, we assume $\theta_{s,1} = 1$ is fixed implying that each θ defines a unique policy. It is important to note that our result about stationary points in Theorem 5 does not apply in a meaningful way to the class of softmax policies as it is not closed. For non-degenerate values of θ, any policy π_θ is suboptimal as all stationary points lie at corners of the probability simplex, i.e. $\pi_\theta(i|s) = 1$ for some $i \in \{1,2,\ldots,k\}$ for every state s, which corresponds to some components of θ becoming infinite. Convergence to stationary points can only occur in the limit as some components of θ tend to infinity, sending the probability of certain actions to zero. This kind of convergence is not treated in standard optimization results like Lemma 16.

One way to make our results meaningful for softmax policies is by adding a small regularizer to the cost function that penalizes near-deterministic actions[10]*. To sketch this idea, consider defining $g(s,a) = \sum_{i=1}^k g(s,e_i)a_i + R(a)$ where $R(a) \to \infty$ if $a_i \to 0$ for any i. This is a feature, for example, of the relative entropy function $R(a) = D_{\mathrm{KL}}(U||a)$ where U is the uniform distribution ($U_i = 1/k$*

[10]An alternative is to regularize $\ell(\theta)$ directly. We do not consider this as as it does not lie within the scope of our formulation.

for each i). For such a regularizer, $R(\pi_\theta(s)) \to \infty$ as $\|\theta_s\| \to \infty$, implying $\ell(\theta)$ is coercive. Continuous and coercive functions are known to attain a global minimum, so $\arg\min_\theta \ell(\theta)$ *is non-empty and $\ell(\cdot)$ has an interior minimizer. Our result in Theorem 5 can now be used to show $\ell(\theta)$ has no suboptimal stationary points, and invoking Lemma 16, gradient descent converges to the global optimum.*

We now turn to an example with a structured policy class.

Example 4 (Optimal Stopping)**.** *The optimal stopping problem is most naturally formulated as a reward maximization problem*[11]*. In each round the agent observes a state variable x_t in a finite set $\mathcal{X}$, which evolves according to an uncontrolled Markov chain with time-homogenous transition probabilities $\mathbb{P}(x_{t+1} = x' | x_t = x) = p(x'|x)$. Conditioned on x_t, the agent receives an offer y_t drawn i.i.d from a probability distribution that has a density function $q_{x_t}(\cdot)$ supported over $\mathcal{Y}$. We assume q_{x_t} has a continuous derivative and the offer set $\mathcal{Y} = [y_{\min}, y_{\max}]$ is an interval in $\mathbb{R}$ with $y_{\min} > 0$. If the offer is accepted in round t, the process terminates and the decision maker accrues a reward of $\gamma^t y_t$ while rejecting the offer in any round is cost-less. The agent's objective is to maximize the expected revenue.*

The problem can be formalized as a Markov decision process with the state-space $\mathcal{S} = \mathcal{S}_C \cup \{T\}$, consisting of a set of continuation states $\mathcal{S}_C = (\mathcal{X} \times \mathcal{Y})$ and a terminal state T that is cost-less, $g(T, a) = 0$ and absorbing, $P(T|T, a) = 1$. To simplify notation, we assume ρ is an initial distribution over continuation states $\mathcal{S}_C$ and there is zero probability of trivial problem instances that start in the terminal state. We assume the initial distribution factors as $\rho(x, y) = \nu(x)q_x(y)$ where $\nu(x) > 0 \; \forall x \in \mathcal{X}$. The action action $a = 0$ corresponds to accepting the offer and terminating while action $a = 1$ continues the game by transitioning to a new state with probabilities given by

$$\mathbb{P}\left[s_{t+1} = (x', dy') \mid s_t = (x, y), a = 1\right] = p(x' \mid x) q_{x'}(y').$$

[11] We can imagine costs to be the negative reward in order to be consistent with our formulation.

We consider the class of threshold policies where the vector $\theta \in \Theta := [y_{\min}, y_{\max}]^{|\mathcal{X}|}$ *specifies one stopping threshold per context. The policy* $\pi_\theta(x, y) = \mathbb{1}(y < \theta_x)$ *rejects all offers below the threshold.*

It is easy to verify that the class of threshold policies is closed under policy improvement, i.e. Condition 1. For any $\pi \in \Pi_\Theta$*, consider the policy iteration update for any state* $s = (x, y) \in \mathcal{S}_C$*:*

$$\begin{aligned}\pi^+(x, y) &= \arg\max_{a\in\{0,1\}} Q_\pi((x, y), a) \\ &= \arg\max_{a\in\{0,1\}} \left[ay + (1 - a)\gamma \sum_{x'\in\mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J_\pi((x', y'))q_{x'}(y')dy'\right].\end{aligned}$$

Clearly, $\pi^+(x, y) = 1$ *if and only if* y *exceeds the continuation value which is defined as,* $c_\pi(x) := \gamma \sum_{x'\in\mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J_\pi(x', y')q_x(y')dy'$*. Thus,* π^+ *is itself a threshold policy.*

We also verify Condition 2 to show that the weighted policy iteration objective has no suboptimal stationary points and can be therefore solved to optimality by a gradient method. For any $\pi \in \Pi_\Theta$*, the fundamental theorem of calculus implies*

$$\begin{aligned}\frac{\partial}{\partial\theta_x}\mathcal{B}(\theta|\eta_\pi, J_\pi) &= \eta'_\pi(x)\frac{\partial}{\partial\theta_x}\int_{y\in\mathcal{Y}} Q_\pi\left((x, y), \pi_\theta(x, y)\right) q_x(y)dy \\ &= \eta'_\pi(x)\frac{\partial}{\partial\theta_x}\int_{y\in\mathcal{Y}} \left[\mathbb{1}(y \geq \theta_x)y + \mathbb{1}(y < \theta_x)c_\pi(x)\right] q_x(y)dy \\ &= \left(c_\pi(x) - \theta_x\right)\eta'_\pi(x)q_x(\theta_x). \qquad (3.14)\end{aligned}$$

where η'_π *denotes the marginal distribution over* $\mathcal{X} \cup \{T\}$ *under* η_π*. As we formulate this as a maximization problem, any stationary point of* $\mathcal{B}(\theta|\eta_\pi, J_\pi)$ *satisfies,*

$$\max_{\theta'_x\in\mathcal{Y}} \frac{\partial}{\partial\theta_x}\mathcal{B}(\theta|\eta_\pi, J_\pi)\cdot(\theta'_x - \theta_x) \leq 0 \implies \max_{\theta'_x\in\mathcal{Y}} \left(c_\pi(x) - \theta_x\right)\cdot(\theta'_x - \theta_x) \leq 0 \qquad (3.15)$$

Note that $c_\pi(x) < y_{\max}$ *for any* $\pi \in \Pi_\Theta$ *and* $x \in \mathcal{X}$*,. This follows by assumption that* $q_x(\cdot)$ *is supported over* $\mathcal{Y}$ *and therefore* $\gamma \int_{\mathcal{Y}} J_\pi(x, y)q_x(y)dy < y_{\max}$*. Using this with* (3.15) *implies that*

if θ is a stationary point then

$$\theta_x = \begin{cases} c_\pi(x) & \text{if } c_\pi(x) \in (y_{\min}, y_{\max}), \\ y_{\min} & \text{if } y_{\min} \geq c_\pi(x) \end{cases} \tag{3.16}$$

To show that $\theta' \mapsto \mathcal{B}(\theta'|\eta_\pi, J_\pi)$ has no suboptimal stationary points, consider the policy iteration update,

$$\begin{aligned} \theta = \arg\max_{\theta' \in \Theta} \mathcal{B}(\theta'|\eta_\pi, J_\pi) &= \arg\max_{\theta' \in \Theta} \mathbb{E}_{\eta_\pi(x,y)} \left[Q_\pi((x,y), \pi_{\theta'}(x,y)) \right] \\ &= \arg\max_{\theta' \in \Theta} \mathbb{E}_{\eta_\pi(x,y)} \left[\mathbb{1}(y \geq \theta'_x) y + \mathbb{1}(y < \theta'_x) c_\pi(x) \right] \end{aligned}$$

For any state $(x, y) \in \mathcal{S}$, the improved policy should accept only if the offer y is greater than the continuation value $c_\pi(x)$. Hence,

$$\theta_x = \begin{cases} c_\pi(x) & \text{if } c_\pi(x) > y_{\min}, \\ y_{\min} & \text{if } c_\pi(x) \leq y_{\min} \end{cases} \tag{3.17}$$

From (3.16) *and* (3.17) *its clear that any stationary point solves the weighted policy iteration objective. We verify the additional smoothness properties in Condition 0 in Appendix B.5.*

3.6 Beyond closed policy classes: the case of non-stationary policy classes

For finite horizon problems with non-stationary policy classes, we can show that policy gradient methods face no spurious local minima under a much weaker condition. Rather than require the policy class is closed under improvement, it is sufficient that the policy class contains the optimal policy[12]. For this reason, our theory will cover as special cases a broad variety of finite horizon dynamic programming problems for which structured policy classes are known to be optimal. Interestingly, this result relies critically on the use of a non-stationary policy class. In particular,

[12]Closure of the policy class implies that it contains the optimal policy.

Example 1 shows that policy gradient performed with respect to stationary policy classes can get stuck in bad local minima even if the policy class contains an optimal policy.

As motivation, consider the finite-horizon inventory control in Example 5 for which [71] previously showed through a somewhat intricate analysis that a stochastic approximation algorithm converges to the optimal policy, despite non-convexity of the objective.

Example 5 (Finite horizon inventory control)**.** *We consider a multi-period inventory control problem (also popularly known as the newsvendor problem) with backlogged demands where at time t, we denote s_t to be the state of the seller's inventory, $a_t \geq 0$ to be the quantity of inventory ordered (only non-negative orders are allowed) and w_t to be the random demand (assumed to be i.i.d for simplicity). We assume the demand distribution to have a density which is supported over a bounded set $[0, w_{\max}]$. For a problem with horizon H, the seller's objective is to minimize total expected cost*

$$\mathbb{E}\left[\sum_{t=0}^{H-1}(ca_t + b\max\{s_t + a_t - w_t, 0\} + p\max\{-s_t + a_t - w_t, 0\})\right] \tag{3.18}$$

where $c, b, p > 0$ denote the per unit costs of ordering, holding and backlogging items, respectively. The inventory level evolves as: $s_{t+1} = s_t + a_t - w_t \;\forall\, t = \{0, \ldots, H-1\}$. Negative inventory levels correspond to backlogged demand that is filled when additional inventory becomes available.

It is well known that a base-stock policy is optimal for this setting [101]. Therefore, we consider the class of base-stock-policies parameterized as $\Pi_\Theta = \left\{\theta = (\theta_0, \ldots, \theta_{H-1}) \in \mathbb{R}^H : \theta_t > 0\right\}$ which orders inventory $\pi_\theta(s_t) = \max\{0, \theta_t - s_t\}$ at time t. That is, it orders enough inventory to reach a target level θ_t, whenever feasible. We also assume that $p > c$. Otherwise, the optimal policy, θ^* never orders inventory.

We can state our formal result without introducing new notation for the finite horizon setting, by a well known trick that treats finite-horizon time-inhomogenous MDPs as a special case of infinite horizon MDPs (see e.g. [133]). Essentially, one can imagine that the state space factorizes

into $H+1$ components as $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H \cup \mathcal{S}_{H+1}$, thought of as stages or time periods of the decision problem. For any policy, a state $s \in \mathcal{S}_i$ transitions to a state in $\mathcal{S}_{i+1}$ until stage $H+1$ is reached and the interaction effectively ends. We also assume the policy class factors into separate components. This structure allows us to change the policy in stage h without influencing the policy at other stages and essentially encodes time-inhomogenous policies.

Condition 3. *Suppose the state space factors as $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H \cup \mathcal{S}_{H+1}$, where for a state $s \in \mathcal{S}_h$ with $h \leq H$, $\sum_{s' \in \mathcal{S}_{h+1}} P(s'|s,a) = 1$ for all $a \in \mathcal{A}_s$. The final subset $\mathcal{S}_{H+1} = \{\tau\}$ contains a single cost-less absorbing state, with $P(\tau|\tau,a) = 1$ and $g(\tau,a) = 0$ for any action a. The parameter space is the product set $\Theta = \Theta_1 \times \cdots \times \Theta_H$, where a policy parameter $\theta = (\theta_1, \ldots, \theta_H) \in \Theta$ is the concatenation of H sub-vectors.*

Remark 4. *As a remark on notation, we assume the state is always endowed with a time index for simplicity. We write s_h or $s \in \mathcal{S}_h$ and therefore in the notation $\pi_\theta(s)$ for $s \in \mathcal{S}_h$ or $\pi_\theta(s_h)$, the policy depends only on θ_h. We assume the initial distribution to be supported over all time periods implying that we can start from any subproblem. For each time period, $h \in \{1, \ldots, H\}$, we assume ρ to have a density supported over $\mathcal{S}_h$.*

For the inventory control proble in Example 5, we assumed the demand distribution to be bounded for the finite horizon inventory problem. Therefore, $\mathcal{S}_h = [-M, M]$ and $\Theta_h = [0, M]$ for some finite M[13].

We now state the main result of this subsection, which applies under conditions much weaker than those for Theorem 5. As opposed to Condition 1, we only require Π_Θ to contain the optimal policy. We also need the following which is weaker than Condition 2, since it only treats the weighted policy iteration objective corresponding to the optimal cost-to-go function, J^*.

Condition 4. *The problem $\min_{\theta \in \Theta} \mathcal{B}(\theta|\eta_\pi, J^*)$ has no suboptimal stationary points.*

Recall that Condition 2 considers stationary points of single-period problems, $\theta \to \mathcal{B}(\theta|\eta_\pi, J_\pi)$, induced by any suboptimal policy $\pi \in \Pi_\Theta$. For the inventory control problem in Example 5, a sim-

[13] We only need to search over the set of possible non-negative states to find the optimal threshold.

ple argument shown in Appendix B.6 verifies Condition 3. The result, essentially follows by using convexity of $Q^*(s, a)$ (see Chapter 3 in [101]).

Theorem 6. *Suppose Conditions 3 and 4 hold. If the parameterized policy class Π_Θ contains an optimal policy π^*, then any stationary point θ of $\ell : \Theta \to \mathbb{R}$ satisfies $J_{\pi_\theta} = J^*$.*

The full proof is given in Appendix B.6 and proceeds by backward induction. We first show how all stationary points must play according to an optimal policy at the final-period states $s \in \mathcal{S}_H$. From this, we argue that at any stationary point, the policy must also act optimally from any state $s \in \mathcal{S}_h$ for all $h < H$.

3.7 The exploratory initial distribution and concentrability coefficients

This section defines a constant, which we call the effective concentrability coefficient, that measures the efficacy of an exploratory initial distribution. This constant will impact the quality of the convergence rates in established in Section 3.8 and the approximation results established in Section 3.9.

To motivate this definition, let us first recall some of our results in Section 3.5 which show that any stationary point of the policy gradient loss function has a corresponding cost-to-go function which satisfies an average form of Bellman's equation. Crucially, as confirmed in Lemma 22, this implies such policies are optimal. Extending our analysis requires connecting the magnitude of errors in Bellman's equation to the optimality gap. As an example, if the Bellman optimality operator T is a contraction in a norm $\|\cdot\|$ then

$$\|J - J^*\| \leq \frac{1}{(1-\gamma)}\|J - TJ\| \quad \forall\, J \in \mathcal{J}. \tag{3.19}$$

See [101] or (B.3) in Appendix B.1. Here it is critical that T is a contraction in the same norm that is used to measure distance from the optimal cost-to-go function.

We define a constant, κ_ρ, which enables the same inequality as in (3.19) in the weighted norm $\|\cdot\|_{1,\rho}$, in which T is typically not contractive. Intuitively, κ_ρ captures how errors in the cost-to-go

functions manifest in Bellman errors when sampling from the exploratory initial distribution ρ. Recall that $\mathcal{J}_\Theta = \{J_{\pi_\theta} : \theta \in \Theta\}$ is the set of cost-to-go functions induced by the parameterized policy class. It is critical, at least for some of our results (Lemma 26 for example), that we measure κ_ρ only on this subclass of cost-to-go functions and not all functions $J : \mathcal{S} \to \mathbb{R}$.

Definition 2. *Define the effective concentrability coefficient κ_ρ of the class of cost-to-go functions $\mathcal{J}_\Theta$ to be the smallest scalar such that*

$$\|J - J^*\|_{1,\rho} \leq \frac{\kappa_\rho}{(1-\gamma)} \|J - TJ\|_{1,\rho} \qquad \forall J \in \mathcal{J}_\Theta. \tag{3.20}$$

If no such scalar exists then we say $\kappa_\rho = \infty$.

This definition is motivated by two important factors. First, the optimality gap under $\ell(\cdot)$ can be written as $\ell(\pi_\theta) - \min_{\pi \in \Pi} \ell(\pi) = (1-\gamma)\|J_{\pi_\theta} - J^*\|_{1,\rho}$, mirroring the left hand side of (3.20) modulo a constant factor. Second, due to the policy gradient formula in Lemma 20, our results naturally depend on errors Bellman equation weighted under the state occupancy measure η_π. See Lemma 22, for example. As $\eta_\pi(s) \geq (1-\gamma)\rho(s)$, it makes sense to therefore measure the Bellman errors in $\|\cdot\|_{1,\rho}$.

We call κ_ρ the *effective concentrability coefficient*, since it plays a role similar to the concentrability coefficients ([134, 135]) that are widely used in the analysis of approximate value and policy iteration algorithms [134, 135, 136, 137, 78, 98, 138]. See [139] for a detailed comparison on different notions of the concentrability coefficient. Note, instead of stating a more restrictive regularity assumption on the MDP or the initial distribution, our definition of κ_ρ in (3.20) is precisely the quantity we need in our analysis. We now give various bounds on κ_ρ below.

The first bound depends on the likelihood ratio between the state occupancy measure under the optimal policy and the initial distribution. This yields the simple bound $\kappa_\rho < 1/\min_s \rho(s)$ in any finite state problem, but it could also be finite in some infinite state problems.

Lemma 23. *Let π^* denote any optimal stationary policy. Then,*

$$\kappa_\rho \leq \sup_{s \in \mathcal{S}} \frac{\eta_{\pi^*}(s)}{\rho(s)}$$

This result is, essentially, a restatement of a key observation in [78] and can be derived using the variational form of the Bellman equation (see (B.5) in Appendix B.1), also known as the performance difference lemma [78]. For completeness, we give a short proof in Appendix B.7. Such *distributional mismatch* terms also appears in the works of [98, 99].

An alternative approach to bounding κ_ρ is to relate the weighted 1-norm to a different norm in which the Bellman operator is a contraction.

Lemma 24 (Concentrability via norm equivalence)**.** *If T is a contraction with modulus γ in a norm $\|\cdot\|$ that satisfies*

$$c\|J\| \leq \|J\|_{1,\rho} \leq C\|J\| \qquad \forall\, J \in \mathcal{J}, \tag{3.21}$$

then $\kappa_\rho \leq C/c$.

Proof. Using that T is contraction with modulus γ in $\|\cdot\|$ implies that $\|J - J^*\| \leq \frac{1}{(1-\gamma)}\|J - TJ^*\|$ (see (B.3) in Appendix B.1). Then,

$$\|J - J^*\|_{1,\rho} \leq C\|J - J^*\| \leq \frac{C}{(1-\gamma)}\|J - TJ\| \leq \frac{C}{c(1-\gamma)}\|J - TJ\|_{1,\rho}$$

□

Lemma 24 is potentially useful for many problems where the Bellman operator is a contraction with respect to a certain weighed norm, as it suggests ρ should be chosen in a manner that aligns with that norm's state weighting. Optimal stopping problem is one such special case in which a very natural choice of ρ is suggested by the contraction properties of T. In particular, if ρ is chosen to be the stationary distribution of the underlying Markov chain – assuming it is never interrupted by stopping – then $\kappa_\rho \leq 1$. In practical problems, one could easily sample initial states from ρ by simulating this Markov process.

Lemma 25 (Concentrability in optimal stopping)**.** *For the optimal stopping problem in Example 4 consider the policy π_C that never stops ($\pi_C(s) = 1$ for each $s \in \mathcal{S}_C$) and suppose the induced Markov process has stationary distribution $\mu = \mu P_{\pi_C}$. Then, for the choice $\rho = \mu$, $\kappa_\rho \leq 1$.*

Proof. The analysis in [140] shows T is a contraction in $\|\cdot\|_{2,\mu}$. Similarly, it is easy to show T is a contraction with modulus γ in $\|\cdot\|_{1,\mu}$. The result then follows immediately from Lemma 24. □

Note that the definition of κ_ρ in (3.20) allows for bounds that depend on regularity properties in the cost-to-go functions of interest. This is potentially useful in cases where generic bounds, for example in Lemma 23, are pessimistic.

Take the case of linear quadratic control, where the cost-to-go functions induced by the class of linear policies are quadratic. As a result, we need only the initial distribution to explore the basis of the state space sufficiently, rather than requiring it to almost perfectly mimic the steady state distribution of the (unknown) optimal policy. In this case, the following lemma bounds κ_ρ. One term depends on the condition number of the second moment matrix under initial distribution ρ while the other depends on $\|A + B\theta^*\|_2$ for the optimal linear policy θ^*. Recall, we assumed the system to be controllable. This implies the optimal policy is stable and satisfies $\|A + B\theta^*\|_2 < 1/\sqrt{\gamma}$, which ensures that the first term $\frac{(1-\gamma)}{1-\gamma\|A+B\theta^*\|_2^2}$ is finite, less than 1 when $\|A + B\theta^*\|_2 \leq 1$, and becomes large in problems where θ^* is only barely stable.

Lemma 26 (Concentrability in LQ control)**.** *Consider the linear quadratic control problem in Example 2. Suppose $\Sigma_\rho := \mathbb{E}_{s_0\sim\rho}[s_0 s_0^\top] > 0$ and let $\theta^* \in \mathbb{R}^{n\times k}$ denote the parameter of an optimal policy. Then,*

$$\kappa_\rho \leq \frac{(1-\gamma)}{1-\gamma\|A + B\theta^*\|_2^2} \cdot \frac{\lambda_{\max}\left(\Sigma_\rho\right)}{\lambda_{\min}\left(\Sigma_\rho\right)}.$$

3.8 Convergence rates for policy gradient methods

Our result in Theorem 5 guarantees that any stationary point of the policy gradient objective, $\ell(\cdot)$ is globally optimal assuming (i) the policy class is closed in policy improvement and (ii) the weighted policy iteration objective function, $\theta \mapsto \mathcal{B}(\theta \,|\, \eta_\pi, J_\pi)$ has no sub-optimal stationary

points for any $\pi \in \Pi_\Theta$ and distribution η supported over the entire state space $\mathcal{S}$. This however only implies an asymptotic result: optimizing the policy gradient objective with first order methods converges asymptotically to a stationary point which is also globally optimal. From a practitioners perspective however, we also care about finite time convergence rates which provide bounds on the optimality gap after say a finite number of policy gradient updates.

Our main insight in this section is to identify conditions which guarantee that the policy gradient objective is gradient dominated. The result follows if the weighted policy iteration objective, $\theta \mapsto \mathcal{B}(\theta \,|\, \eta_\pi, J_\pi)$ is gradient dominated and the policy class is closed (condition 1). We use the policy gradient theorem in Lemma 20 to show how these conditions translate to gradient dominance of $\ell(\cdot)$. This is useful as under suitable smoothness assumptions, it is well known that first order methods converge rapidly to the globally optimal solutions if the objective is gradient dominated [see e.g 141]. Such gradient dominance conditions also underly the proof of [70] for linear quadratic control.

We remark that assuming the weighted policy iteration objective to be gradient dominated is not entirely impractical. As noted before in Section 3.5, $\theta \mapsto \mathcal{B}(\theta \,|\, \eta_\pi, J_\pi)$, is linear for finite MDPs and quadratic for LQ control problem and therefore our results in this section immediately apply to these examples.

3.8.1 Background on Gradient Dominance

Throughout this section, we consider the optimization problem $\min_{x \in \mathcal{X}} f(x)$ where $\mathcal{X} \subseteq \mathbb{R}^d$. When $\mathcal{X} \subset \mathbb{R}^d$, we assume it to be a closed convex set and let f be differentiable on an open set containing $\mathcal{X}$. Recall that the defining feature of a convex function is that it lies above its tangents, that is for each $x \in \mathcal{X}$, $f(x') \geq f(x) + \langle \nabla f(x)\,,\, x' - x \rangle$ for every $x' \in \mathcal{X}$. In analysis of optimization algorithms, this property not only implies that $f(\cdot)$ has no suboptimal stationary points, it can be used to bound the optimality gap by a measure of distance from stationarity, as

$$\min_{x' \in \mathcal{X}} f(x') \geq f(x) + \min_{x' \in \mathcal{X}} \langle \nabla f(x)\, x' - x \rangle.$$

Assuming $\|x - x'\| \leq R$ for all $x, x' \in \mathcal{X}$, one can deduce that $\min_{x' \in \mathcal{X}} f(x') \geq f(x) + R\|\nabla f(x)\|$, indicating that the gradient norm bounds sub optimality. When $f(\cdot)$ is strongly convex, this conclusion can be strengthened, leading to faster convergence rates. In particular, if f is μ–strongly convex, then

$$\min_{x' \in \mathcal{X}} f(x') \geq f(x) + \min_{x' \in \mathcal{X}} \left[\langle \nabla f(x), x' - x \rangle + \frac{\mu}{2} \|x - x'\|_2^2 \right].$$

When $\mathcal{X} = \mathbb{R}^d$, this says $\min_{x' \in \mathcal{X}} f(x') \geq f(x) - \frac{\mu}{2} \|\nabla f(x)\|_2^2$, so the optimality gap is bounded by the *squared* norm of the gradient in this case.

Below, we introduce a notion of gradient dominance. This definition, essentially, assumes a critical implication of convexity or strong convexity rather than assuming these properties themselves. According to the definition below, a convex function is $(1, 0)$–gradient dominated. A μ–strongly convex function is $(1, \mu)$–gradient dominated.

Definition 3. *For $\mathcal{X} \subseteq \mathbb{R}^d$, we say f is (c, μ)–gradient dominated over $\mathcal{X}$ if there exists a constant $c > 0$ and $\mu \geq 0$ such that*

$$\min_{x' \in \mathcal{X}} f(x') \geq f(x) + \min_{x' \in \mathcal{X}} \left[c \langle \nabla f(x), x' - x \rangle + \frac{\mu}{2} \|x - x'\|_2^2 \right] \quad \forall x \in \mathcal{X}. \tag{3.22}$$

The function is said to be gradient dominated with degree 1 if $\mu = 0$ and gradient dominated of degree two if $\mu > 0$.

For $\mathcal{X} \subset \mathbb{R}^d$, the definition above may seem somewhat non-standard as authors typically consider unconstrained optimization (see [74, 70] for example), in which case the definition reduces to the well known Polyak-Lojasiewicz (PL) inequality [72],

$$\min_{x' \in \mathcal{X}} f(x) \geq f(x) - \frac{c^2}{2\mu} \|\nabla f(x)\|_2^2.$$

Under gradient dominance conditions, popular first order optimization algorithms are assured to converge to the global minimum and a simple analysis provides finite time rates of convergence as shown in Lemma 27 below. The first result focuses on projected gradient descent, strengthening

Lemma 16 by providing a convergence rate. This result is obtained by using a well known fact that projected gradient descent reaches an approximate stationary point rapidly (approximately at $O(1/\sqrt{T})$ rate), and then using the definition of gradient dominance to relate approximate stationarity to the optimality gap. The second case considers an unconstrained problem shows a geometric convergence rate. The analysis here is essentially identical to the typical analysis of gradient descent for strongly convex objectives. Recall that a differentiable function is said to be L-smooth if $\nabla f(x)$ is Lipschitz with constant L with respect to the Euclidean norm.

Lemma 27 (Convergence rates for gradient dominated smooth functions)**.** *Consider the problem,* $\min_{x\in\mathcal{X}} f(x)$ *where* $\mathcal{X} \subseteq \mathbb{R}^d$*. Assume* f *be* L*–smooth on* $\mathcal{X}$ *and a non-empty solution set. Denote* $f(x^*) = \min_{x'\in\mathcal{X}} f(x')$*. Consider the sequence* $x_{t+1} = \text{Proj}_{\mathcal{X}}\left(x_t - \alpha\nabla f(x_t)\right)$*.*

1. *Let* $\mathcal{X} \subset \mathbb{R}^d$ *such that* $\|x - x'\|_2 \leq R < \infty$ *for all* $x, x' \in \mathcal{X}$*. Assume* f *is* k*-lipschitz and* $(c, 0)$*–gradient-dominated and where* $\alpha \leq \min\{\frac{1}{k}, \frac{1}{L}\}$*. Then,*

$$\min_{t\leq T}\{f(x_t) - f(x^*)\} \leq \sqrt{\frac{2R^2c\left(f(x_0) - f(x^*)\right)}{\alpha T}}$$

2. *Karimi et. al. [74], Polyak [72] Assume* $\mathcal{X} = \mathbb{R}^d$ *and* $\alpha = 1/L$*. If* f *is* (c, μ)*–gradient-dominated for* $\mu > 0$*, then,*

$$f(x_t) - f(x^*) \leq \left(1 - \frac{\mu}{c^2L}\right)^t \left(f(x_0) - f(x^*)\right)$$

Proof. See Appendix B.2 for a detailed proof of Part (1). □

Remark 5. *Part (2) of Lemma 27 assumes* $\mathcal{X} = \mathbb{R}^d$*, in which case the sequence* x_t *is produced by gradient descent with a fixed stepsize. It is likely possible to show geometric rates for projected gradient descent on constrained subsets* $\mathcal{X} \neq \mathbb{R}^d$*. We do not consider this case for brevity.*

3.8.2 Gradient dominance of the policy gradient objective

We continue to assume that the policy class is closed under policy improvement, as stated in Condition 1. But now, instead of assuming the policy improvement objective has no sub-optimal stationary points, we impose a stronger property that it is gradient dominated. This condition holds whenever the single period objective solved by policy iteration is convex and therefore immediately applies to the LQ control problem as well as for tabular MDPs.

Condition 5 (Gradient dominance of the weighted policy iteration objective)**.** *For any $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ is (c, μ)–gradient-dominated over Θ.*

This gradient dominance condition ensures the single-period objective optimized by policy iteration problem could be solved efficiently by first order methods. Our main result in Theorem 7 shows that for closed policy classes, this gradient dominance condition is automatically inherited by the multi-period objective $\ell(\cdot)$, implying convergence rates for first-order methods applied to $\ell(\cdot)$. Notice that the constants degrade with the horizon and the concentrability coefficient κ_ρ stated in Definition 3.20. The Corollary 1 provides a more interpretable statement, which follows because any convex $\mathcal{B}(\theta|\eta_\pi, J_\pi)$ is $(1, 0)$–gradient-dominated any μ–strongly–convex $\mathcal{B}(\theta|\eta_\pi, J_\pi)$ is $(1, \mu)$–gradient-dominated.

Theorem 7. *If Conditions 0, 1, and 5 hold, then $\ell(\cdot)$ is $\left(\frac{1-\gamma}{\kappa_\rho} \cdot c\,,\ \frac{1-\gamma}{\kappa_\rho} \cdot \mu\right)$–gradient dominated.*

Corollary 1. *Suppose Conditions 0 and 1 hold. If, for every $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ is convex, then $\ell(\theta)$ is gradient dominated of degree one. If $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ is strongly convex then $\ell(\theta)$ is gradient dominated of degree two.*

The proof can be divided into two key steps. First, we use closure property of the policy class to upper bound the optimality gap of the policy gradient objective with that of the weighted policy iteration objective. Essentially, this result shows that the current policy is *nearly optimal* if the weighted policy iteration step offers little improvement over the current policy; assuming the policy iteration step can be performed exactly and there is sufficient exploration, $\rho(s) > 0\ \forall\, s \in \mathcal{S}$.

It is important to note how *optimality* here crucially depends on the use of an exploratory initial distribution under which $\kappa_\rho < \infty$. The second step of the proof translates gradient dominance of the policy iteration objective to that of $\ell(\cdot)$ by using the policy gradient theorem in Lemma 20.

Proof of Theorem 7. We first derive a consequence of the closure condition. Let S denote a random draw from η_{π_θ}. We have,

$$\begin{aligned}
\ell(\pi_\theta) - \min_\pi \ell(\pi) &= (1-\gamma) \sum_{s \in \mathcal{S}} \rho(s) \left(J_{\pi_\theta}(s) - J^*(s) \right) \\
&\overset{(a)}{=} (1-\gamma) \| J_{\pi_\theta} - J^* \|_{1,\rho} \\
&\overset{(b)}{\leq} \kappa_\rho \| J_{\pi_\theta} - T J_{\pi_\theta} \|_{1,\rho} \\
&\overset{(c)}{\leq} \frac{\kappa_\rho}{(1-\gamma)} \| J_{\pi_\theta} - T J_{\pi_\theta} \|_{1,\eta_{\pi_\theta}} \\
&= \frac{\kappa_\rho}{(1-\gamma)} \mathbb{E} \left[J_{\pi_\theta}(S) - T J_{\pi_\theta}(S) \right] \\
&\overset{(d)}{=} \frac{\kappa_\rho}{(1-\gamma)} \mathbb{E} \left[J_{\pi_\theta}(S) - \min_{\pi' \in \Pi_\Theta} T_{\pi'} J_{\pi_\theta}(S) \right] \\
&\overset{(e)}{=} \frac{\kappa_\rho}{(1-\gamma)} \left(\mathcal{B}(\theta \mid \eta_{\pi_\theta}, J_{\pi_\theta}) - \min_{\theta' \in \Theta} \mathcal{B}(\theta' \mid \eta_{\pi_\theta}, J_{\pi_\theta}) \right)
\end{aligned}$$

Here (a) uses that $J_{\pi_\theta} \geq T J_{\pi_\theta}$, (b) directly applies the definition of κ_ρ in Definition 3.20, (c) uses that $\eta_{\pi_\theta} \geq (1-\gamma)\rho$ (by definition, see (3.6)), (d) directly applies the policy closure condition in Condition 1, and (e) uses definition of the weighted policy iteration objective in (3.11).

As we assume $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ is (c, μ)–gradient dominated for each $\pi \in \Pi_\Theta$, we have that for all $\theta \in \Theta$,

$$\begin{aligned}
\mathcal{B}(\theta \mid \eta_{\pi_\theta}, J_{\pi_\theta}) - \min_{\theta' \in \Theta} \mathcal{B}(\theta' \mid \eta_{\pi_\theta}, J_{\pi_\theta}) &\leq - \min_{\nu \in \Theta} \left[c \left\langle \nabla_{\overline{\theta}} \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta}) \Big|_{\overline{\theta}=\theta}, \nu - \theta \right\rangle + \frac{\mu}{2} \| \nu - \theta \|_2^2 \right] \\
&\leq - \min_{\nu \in \Theta} \left[c \left\langle \nabla_\theta \ell(\theta), \nu - \theta \right\rangle + \frac{\mu}{2} \| \nu - \theta \|_2^2 \right],
\end{aligned}$$

which gives our desired result. The second inequality above uses the policy gradient theorem in Lemma 20. □

3.8.3 Gradient dominance and smoothness for examples 2-4

As noted earlier, the gradient dominance condition for LQ control and finite MDPs follows from (strong) convexity of the weighted policy iteration objective. However, the weighted policy iteration cost function is not convex for the optimal stopping problem as described in Example 4. Nevertheless, we show below in Lemma 28 that the total cost function, $\ell(\cdot)$ is in fact gradient dominated using a more direct argument.

Recall that for the optimal stopping problem we assume $\nu(x) > 0$ for all contexts $x \in \mathcal{X}$ and $q_x(\cdot)$ to be a density supported over the offer set $\mathcal{Y}$ such that the initial distribution factorizes as $\rho(x, y) = \nu(x)q_x(y)$. The gradient dominance constants depend on a measure of the degree of uniformity in the initial distribution $\rho(x, y)$. For a general start state distribution, Lemma 23 bounds κ_ρ but Lemma 25 shows that we get a tighter bound of $\kappa_\rho \leq 1$ is possible when ρ is chosen to be the stationary distribution.

Lemma 28 (Gradient dominance for optimal stopping)**.** *Denote κ_ρ to be the concentrability coefficient of the initial distribution $\rho(x, y) = \nu(x)q_x(y)$. Define $\psi := \min_{(x,y)\in\mathcal{S}_C} \nu(x)\, q_x(y)$ and $\beta := \max_{(x,y)\in\mathcal{S}_C} \nu(x)\, q_x(y)$. Then, $\ell(\cdot)$ is $(\frac{\beta}{\kappa_\rho \psi}, 0)$ gradient dominated.*

Proof. See Appendix B.5 for details. □

In addition to gradient dominance, the convergence rates described in Lemma 27 apply to smooth objectives. We refer the readers to Lemma E.3 in [99] which shows smoothness of $\ell(\cdot)$ for tabular MDPs. For the LQ control problem, recall how Lemma 18 shows that $\ell(\cdot)$ is smooth on sub-level sets. As the gradient update, $\theta' = \theta - \nabla\ell(\theta)$, stays in the sub-level set with an appropriately small step-size (see proof of Lemma 16 in Appendix B.2), a rate result simlar to Lemma 27 can therefore be shown for LQ control. We omit this for simplicity.

A smoothness results also holds for the optimal stopping objective as shown below. The proof follows by showing that $\ell(\theta)$ is twice continuously differentiable and leveraging that continuous functions are bounded on compact sets. See Appendix B.5 for details.

Lemma 29 (Cost function for optimal stopping)**.** *In the optimal stopping problem in Example 4,* $\max_{\theta\in\Theta} \|\nabla^2 \ell(\theta)\| < \infty$.

3.9 Policy classes closed under approximate policy improvement

So far, we have studied some classical dynamic programming problems that are ideally suited to policy iteration. The key property we used is that certain structured policy classes are closed under policy improvement, so that exact policy iteration can be performed when only considering that policy class. Although simple structured policy classes are common in some applications of stochastic approximation based policy search [e.g. 142, 143, 144], they are not widely used in the reinforcement learning literature. Instead, flexible policy classes like those parameterized by a deep neural network, a Kernel method [145], or using state aggregation [146, 147, 148] are preferred. We conclude this chapter by presenting some preliminary but interesting progress toward understanding why, for highly expressive policy classes, any local minimum of the policy gradient cost function might be near-optimal. We conjecture this theory can at least be clearly instantiated in special case of state aggregation given in Appendix B.8.

Given an expressive policy class Π_Θ,

$$\inf_{\pi\in\Pi_\Theta} \left\| T_\pi J_{\pi_\theta} - T J_{\pi_\theta} \right\|_{1,\eta_{\pi_\theta}} \tag{3.23}$$

measures the approximation error of the best approximate policy iteration update in the policy class to the current policy π_θ. If Π_Θ were closed under policy improvement steps, the approximation error would be zero since there would exist a $\pi \in \Pi_\Theta$ such that $T_\pi J_{\pi_\theta}(s) = T J_{\pi_\theta}(s)$ for every $s \in \mathcal{S}$. Equation (3.23) measures the deviation from this ideal case, in a norm that weights states by the discounted-state-occupancy distribution η_{π_θ} under the policy π_θ. Our formal result stated below in Theorem 8 bounds the optimality gap at a stationary point by the approximation error in (3.23). Our result in Theorem 8 is reminiscent of results in the study of approximate policy iteration methods, pioneered by [86, 134, 40, 136, 120], among others. The primary differences are that (1)

we directly consider an approximate policy class whereas that line of work considers the error in parametric approximations to the Q-function and (2) we make a specific link with the stationary points of a policy gradient method. The abstract framework of [78] is also closely related, though they do not study the stationary points of $\ell(\cdot)$. Recall the definition of the effective concentrability coefficient, κ_ρ, which relates errors in the Bellman equation to errors in the cost-to-go functions weighted under the initial distribution ρ.

Theorem 8. *Suppose Condition 2 holds. If θ is a stationary point of $\ell(\cdot)$, then*

$$\ell(\pi_\theta) - \min_{\pi \in \Pi} \ell(\pi) \leq \frac{\kappa_\rho}{(1-\gamma)} \min_{\pi \in \Pi_\Theta} \|T_\pi J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\eta_{\pi_\theta}}$$

Proof. Suppose θ is a stationary point of $\ell : \Theta \to \mathbb{R}$. Let S denote a random draw from η_{π_θ}. Since condition 2 holds, Lemma 21 implies

$$\mathbb{E}\left[\left(J_{\pi_\theta} - T J_{\pi_\theta}\right)(S)\right] \leq \left(\min_{\pi \in \Pi_\Theta} \mathbb{E}\left[T_\pi J_{\pi_\theta}(S)\right] - \mathbb{E}[T J_{\pi_\theta}(S)]\right) = \min_{\pi \in \Pi_\Theta} \|T_\pi J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\eta_{\pi_\theta}} := \epsilon.$$

where the final equality uses that $T_\pi J_{\pi_\theta} \geq T J_{\pi_\theta}$ for any $\pi \in \Pi_\Theta$. Then, we have

$$\begin{aligned}
\ell(\pi_\theta) - \min_\pi \ell(\pi) = (1-\gamma) \sum_{s \in \mathcal{S}} \rho(s) \left(J_{\pi_\theta}(s) - J^*(s)\right) &= (1-\gamma)\|J_{\pi_\theta} - J^*\|_{1,\rho} \\
&\leq \kappa_\rho \|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\rho} \\
&\leq \frac{\kappa_\rho}{(1-\gamma)} \|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\eta_{\pi_\theta}} \\
&= \frac{\kappa_\rho \cdot \epsilon}{(1-\gamma)},
\end{aligned}$$

where the first inequality follows from the definition of κ_ρ as given in (3.20) and the second inequality uses that $\eta_{\pi_\theta} \geq (1-\gamma)\rho$ (by definition, see (3.6)). □

3.10 Notation

Table 3.1: Table of Notation for our general problem formulation

γ	$\triangleq$	Discount factor
$\mathcal{S}$	$\triangleq$	State space
$\mathcal{A}_s \subset \mathbb{R}^k$	$\triangleq$	Convex set of feasible actions when in state s.
Π	$\triangleq$	Set of all stationary policies
$\mathcal{J}$	$\triangleq$	Set of bounded real-valued functions on $\mathcal{S}$.
$g(s,a)$	$\triangleq$	Single period expected cost of action a in state s
$P(s'\|s,a)$	$\triangleq$	Transition probability
g_π	$\triangleq$	Single period cost function under policy π
P_π	$\triangleq$	Markov transition matrix under policy π.
$J_\pi \in \mathcal{J}$	$\triangleq$	cost-to-go function under policy π
$Q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$	$\triangleq$	state-action cost-to-go function under policy π
$J^* \in \mathcal{J}$	$\triangleq$	optimal cost-to-go function
π^*	$\triangleq$	An optimal policy (satisfying $J_{\pi^*} = J^*$).
$Q^* = Q_{\pi^*}$	$\triangleq$	state-action cost-to go function associated with an optimal policy.
$T_\pi : \mathcal{J} \to \mathcal{J}$	$\triangleq$	Bellman operator associated with policy π.
$T : \mathcal{J} \to \mathcal{J}$	$\triangleq$	Bellman optimality operator.
ρ	$\triangleq$	initial distribution with $\rho(s) > 0\ \forall s \in \mathcal{S}$. A column vector.
$\eta_\pi = (1-\gamma)\rho(I - \gamma P_\pi)^{-1}$	$\triangleq$	The discounted state occupancy measure under policy π.
$\ell(\pi) = \rho J_\pi$	$\triangleq$	Expected discounted cost under a random initial state, policy π.
$\Theta \subset \mathbb{R}^d$	$\triangleq$	Convex set of policy parameters
$\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$	$\triangleq$	Parameterized policy class.
$\mathcal{J}_\Theta = \{J_\pi : \pi \in \pi_\Theta\}$	$\triangleq$	Set of cost-to-go functions under parameterized policies.
$\ell(\theta) = \ell(\pi_\theta)$	$\triangleq$	Overloaded notation for $\ell(\pi_\Theta)$.
$\mathcal{B}(\pi'\|\eta, J_\pi)$	$\triangleq$	Policy iteration objective defined in (3.11)
κ_ρ	$\triangleq$	Effective concentrability coefficient described in Section 3.7
$\\|J\\|_\infty$	$\triangleq$	Max-norm $\sup_s \|J(s)\|$
$\\|J\\|_{1,\eta}$	$\triangleq$	Weighted 1-norm $\sum_s \eta(s)\|J(s)\|$.
∇_θ	$\triangleq$	Gradient operator with respect to θ
α	$\triangleq$	Free step-size parameter in iterative algorithms

Chapter 4: On the Linear Convergence of Policy Gradient Methods

In this chapter, we revisit the finite time analysis of policy gradient methods in the simplest setting: finite state and action problems with a policy class consisting of all stochastic policies and with exact gradient evaluations. This setting was recently studied by by [99] and [149], who view these problems as instances of smooth nonlinear optimization problems and suggest small stepsizes to control for the error due to local linearization. Despite non-convexity of the objective, their proofs show policy gradient methods find an ϵ–optimal policy within either $O\left(\frac{1}{\epsilon}\right)$ or $O\left(\frac{1}{\epsilon^2}\right)$ iterations, depending on the precise algorithm used.

Instead of viewing the problem through the lens of nonlinear optimization, we take a policy iteration perspective. We highlight that many forms of policy gradient can work with extremely large stepsizes and attain a *linear* rate of convergence, meaning they require only $O(\log(1/\epsilon))$ iterations to reach an ϵ–optimal policy. Our results cover many first order methods applied to the policy gradient loss function, including projected gradient descent, Frank-Wolfe, mirror descent, and natural gradient descent. In an idealized setting where stepsizes are set by line search, a one paragraph proof applies to all algorithms. For natural gradient algorithms, a longer calculation studies a specific stepsize sequence.

Many caveats apply to these results. The literature claims to effectively approximate natural policy gradient updates with complex deep neural networks [62], but it is unclear whether understanding of other first order algorithms contributes to developing practical algorithms. Small stepsizes may be critical in practice for controlling certain approximation errors and for stabilizing algorithms. None of these issues are present in simple tabular RL problems, however, and given the flurry of interest in policy gradient convergence rates, we believe it is valuable for researchers to have a clear understanding in this idealized case.

On concurrent work. Concurrent to this work, two other sets of authors [150, 151] also show that policy gradient methods for finite MDPs converge linearly. Both these papers contain sophisticated analysis of particular policy gradient algorithms. In contrast, by making a more direct connection to policy iteration, we give simple proofs to show results that apply to a broad range of algorithms and may give readers a clear understanding of why one would expect policy gradient methods to converge geometrically for tabular MDPs.

4.1 Problem Formulation

The problem formulation follows that of Chapter 2 but is specialized to the settings of tabular MDPs. We choose to reproduce it here for convenience of readers. Consider a Markov decision process (MDP), which is a six-tuple $(\mathcal{S}, \mathcal{A}, g, P, \gamma, \rho)$, consisting of a state space $\mathcal{S}$, action space $\mathcal{A}$, cost function g, transition kernel P, discount factor $\gamma \in (0, 1)$ and initial distribution ρ. We assume the state space $\mathcal{S}$ to be finite and index the states as $\mathcal{S} = \{s_1, \cdots, s_n\}$. For each state $s \in \mathcal{S}$, we assume that there are is finite set of k deterministic actions to choose from and take the action space, $\mathcal{A} = \Delta^{k-1}$ to be the set of all probability distributions over the k deterministic actions. That is, any action $a \in \mathcal{A}$ is a probability vector where each component a_i denotes the probability of taking the i^{th} action. The transition kernel P specifies the probability $P(s'|s, a)$ of transitioning to a state s' upon choosing action a in state s. The cost function $g(s, a)$ denotes the instantaneous expected cost incurred when selecting action a in state s. Cost and transition functions can be naturally extended to functions on the probability simplex by defining:

$$g(s, a) = \sum_{i=1}^{k} g(s, e_i)\, a_i \qquad\qquad P(s'|s, a) = \sum_{i=1}^{k} P(s'|s, e_i)\, a_i. \tag{4.1}$$

where e_i is the i-th standard basis vector, representing one of the k possible deterministic actions. We assume that per-period costs are non-negative and uniformly bounded, meaning $0 \leq g(s, e_i) < \infty$ for all $s \in \mathcal{S}$ and $i \in \{1, \ldots, k\}$. The assumption that costs are non-negative is without loss of generality, as one can always add the same large constant to the cost of each state and action

without changing the decision problem.

Cost-to-go functions and Bellman operators. A stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ selects a distribution over the $k-1$ dimensional simplex, Δ^{k-1} for each state $s \in \mathcal{S}$. We use the notation $\pi(s,i)$ to denote the probability of selecting action i in state s under policy π. Let Π denote the set of all stationary policies over the simplex, $\Pi = \{\pi \in \mathbb{R}_+^{n\times k} : \sum_{i=1}^k \pi_{s,i} = 1 \ \forall s \in \mathcal{S}\}$ and $J_\pi : \mathcal{S} \to \mathbb{R}$ be the cost-to-go function for any policy $\pi \in \Pi$ defined as:

$$J_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t g(s, \pi(s)) \,\middle|\, s_0 = s \right].$$

Define the Bellman operator $T_\pi : \mathbb{R}^n \to \mathbb{R}^n$ under the policy π as

$$(T_\pi J)(s) := g(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) J(s').$$

The cost-to-go function under policy π is the unique solution to the Bellman equation, $J_\pi = T_\pi J_\pi$. Similarly, the Bellman optimality operator, denoted by $T : \mathbb{R}^n \to \mathbb{R}^n$ is defined as $(TJ)(s) = \min_{\pi\in\Pi}(T_\pi J)(s)$. The optimal cost-to-go function, J^* which satisfies $J^*(s) = \min_\pi J_\pi(s)$ for all $s \in \mathcal{S}$ is the unique fixed point of T and that there is at least one optimal policy, $\pi^* \in \Pi$ that attains this minimum for every $s \in \mathcal{S}$.

The state-action cost-to-go function corresponding to a policy $\pi \in \Pi$,

$$Q_\pi(s,a) = g(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) J_\pi(s'),$$

measures the cumulative expected cost of taking action a in state s and applying π thereafter. For any polices $\pi, \pi' \in \Pi$, we have the relations

$$Q_\pi(s, \pi(s)) = J_\pi(s), \qquad Q_\pi(s, \pi'(s)) = (T_{\pi'} J_\pi)(s), \qquad \min_{a\in\mathcal{A}} Q_\pi(s,a) = (TJ_\pi)(s).$$

Note that for any $\pi \in \Pi, s \in \mathcal{S}$ and $a \in \Delta^{k-1}$, the Q-function is linear in a, as we can write,

$Q_\pi(s,a) = \sum_{i=1}^k Q_\pi(s,e_i)a_i = \langle Q_\pi(s,\cdot), a\rangle$.

Loss function and initial distribution. Policy gradient methods seek to minimize the scalar loss function

$$\ell(\pi) = (1-\gamma)\sum_{s\in\mathcal{S}} J_\pi(s)\,\rho(s),$$

in which the states are weighted by their initial probabilities under ρ and we have normalized costs by $(1-\gamma)$ for convenience. We assume throughout that ρ is supported on $\mathcal{S}$, meaning that $\rho(s) > 0$ for all $s \in \mathcal{S}$ which implies that $\pi \in \arg\min_{\bar\pi} \ell(\bar\pi)$ if and only if $\pi \in \arg\min_{\bar\pi} J_{\bar\pi}(s) \;\; \forall\, s \in \mathcal{S}$.

State distributions. We define the discounted state-occupancy measure under any policy π and initial state distribution ρ as:

$$\eta_\pi = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \rho P_\pi^t = (1-\gamma)\rho(I-\gamma P_\pi)^{-1}.$$

where η_π and ρ are both row vectors and $P_\pi \in \mathbb{R}^{n\times n}$ denotes the Markov transition matrix under π, i.e. $P_\pi = (P(s'|s,\pi(s)))_{s,s'\in\mathcal{S}}$. Note that we have $\eta_\pi(s) \geq (1-\gamma)\rho(s) > 0$ as we assumed $\rho(s) > 0$ for all $s \in \mathcal{S}$.

4.2 Linear convergence of policy iteration

We briefly revisit the classic policy iteration algorithm as our analysis of policy gradient methods is intricately tied to it. At the same time, we review some basic properties related to Bellman operators (see [101] or [100] for proofs). Starting with an initial policy π, policy iteration first evaluates the corresponding cost-to-go function Q_π, and then finds the policy π^+ with

$$\pi^+(s) \in \arg\min_{a\in\mathcal{A}} Q_\pi(s,a) \qquad \forall\, s \in \mathcal{S}.$$

Expressed in terms of Bellman operators, π^+ is defined by the property $T_{\pi^+}J_\pi = TJ_\pi$. Analysis of policy iteration follows from a few basic properties of Bellman operators. First, they are monotone,

meaning the element-wise inequality $J \preceq J'$ implies $TJ \preceq TJ'$ and $T_\pi J \preceq T_\pi J'$. Next, they are contraction operators contraction operators with respect to the maximum norm. That is, $\|TJ - TJ'\|_\infty \leq \gamma \|J - J'\|_\infty$ and $\|T_\pi J - T_\pi J'\|_\infty \leq \gamma \|J - J'\|_\infty$ hold for for any $J, J' \in \mathbb{R}^n$. The unique fixed points of T and T_π are J^* and J_π, respectively.

Let us now apply these properties to analyze policy iteration. Observe that

$$J_\pi = T_\pi J_\pi \preceq TJ_\pi = T_{\pi^+} J_\pi. \tag{4.2}$$

Inductively applying T_{π^+} to each side and using the monotonicity property yields the following policy improvement property,

$$J_\pi \geq T_{\pi^+} J_\pi \geq T_{\pi^+}^2 J_\pi \geq \cdots \geq J_{\pi^+}. \tag{4.3}$$

Since $J_\pi \geq TJ_\pi \geq J_{\pi^+} \geq J^*$ we have,

$$\|J_{\pi^+} - J^*\|_\infty \leq \|TJ_\pi - J^*\|_\infty = \|TJ_\pi - TJ^*\|_\infty \leq \gamma \|J_\pi - J^*\|_\infty. \tag{4.4}$$

From this, we conclude that policy iteration converges at least linearly. Let $\{\pi^t\}_{t \geq 0}$ be the set of policies produced by policy iteration. Then iterating over (4.4) shows

$$\|J_{\pi^t} - J^*\|_\infty \leq \gamma \|J_{\pi^{t-1}} - J^*\|_\infty \leq \cdots \leq \gamma^t \|J_{\pi^0} - J^*\|_\infty.$$

In fact, sometimes policy iteration converges quadratically in the limit [100].

4.3 A sharp connection between policy gradient and policy iteration

We specialize the presentation of the policy gradient theorem in Chapter 2 to tabular problems. Recall that e_i denotes the i-th standard basis vector, representing one of the k possible deterministic

actions. Define the weighed policy iteration or "Bellman" objective:

$$\mathcal{B}(\bar{\pi}|\eta, J_\pi) = \sum_{s\in\mathcal{S}} \eta(s) Q_\pi(s, e_i)\bar{\pi}(s, i) = \langle Q_\pi \, , \pi' \rangle_{\eta\times 1}$$

where $\langle v, u\rangle_W = \sum_{s=1}^n \sum_{i=1}^k v(s,i)u(s,i)W(s,i)$ denotes the W-weighted inner product and $\eta_\pi \times 1$ denotes a weighting that places weight $\eta_\pi(s) \cdot 1$ on any state-action pair (s, i). For a probability distribution η with $\eta(s) > 0$ for each $s \in \mathcal{S}$, the policy iteration update to π is a minimizer $\pi^+ \in \arg\min_{\bar{\pi}\in\Pi} \mathcal{B}(\bar{\pi}|\eta, J_\pi)$.

The policy gradient theorem connects gradients of the infinite horizon cost function $\ell(\cdot)$ to gradients of the single period cost function underlying policy iteration. In particular, we have

$$\nabla\ell(\pi) = \nabla\mathcal{B}(\bar{\pi}|\eta_\pi, J_\pi)\Big|_{\bar{\pi}=\pi} = (\eta_\pi(s)Q_\pi(s, e_i))_{s\in\mathcal{S}, i\in[k]}$$

Equivalently, we can write a first order Taylor expansion of $\ell(\cdot)$ as

$$\begin{aligned}\ell(\bar{\pi}) &= \ell(\pi) + \langle \nabla\ell(\pi) \, , \, \bar{\pi} - \pi\rangle + O(\|\bar{\pi} - \pi\|^2) \\ &= \ell(\pi) + \langle Q_\pi \, , \, \bar{\pi} - \pi\rangle_{\eta_\pi\times 1} + O(\|\bar{\pi} - \pi\|^2).\end{aligned}$$

Essentially, we interpret policy gradient $\nabla\ell(\pi)$ as the gradient of a weighted policy iteration objective. For tabular MDPs, the policy iteration step is simple as it reduces to solving a linear optimization problem over the probability simplex, and the solution is to select the best action for each state. First order methods applied to $\mathcal{B}(\cdot|\eta_\pi, J_\pi)$ can essentially solve such a problem in a single iteration using a large stepsize and we show that for this reason several first order methods applied to $\ell(\cdot)$ will converge as rapidly as policy iteration.

4.4 Policy gradient methods for finite MDPs

We write all algorithms in terms of their evolution in the space of policies Π. Several of them could instead be viewed as operating in the space of parameters for some parameterized policy

class. We discuss this in Remark 6, but keep our formulation and results focused on the space of policies Π. Note that $\Pi = \Delta^{k-1} \times \cdots \times \Delta^{k-1}$ is the n-fold product of the probability simplex. This form of the policy class will cause certain policy gradient updates to decouple across states.

Frank-Wolfe. Starting with some policy $\pi \in \Pi$, an iteration of the Frank-Wolfe algorithm computes

$$\pi^+ = \arg\min_{\bar{\pi} \in \Pi} \langle \nabla \ell(\pi), \bar{\pi} \rangle = \arg\min_{\bar{\pi} \in \Pi} \langle Q_\pi, \bar{\pi} \rangle_{\eta_\pi \times 1} \tag{4.5}$$

and then updates the policy to $\pi' = (1 - \alpha)\pi + \alpha\pi^+$. We use the notation π' in (4.5) as it is exactly a policy iteration update to π so *Frank-Wolfe mimics a soft-policy iteration step*, akin to the *conservative policy iteration* update in [78]. Note, the minimization problem in (4.5) decouples across states to optimize a linear objective over the probability simplex, so

$$\pi^+(s) \in \arg\min_{d \in \Delta^{k-1}} d^\top Q_\pi(s, \cdot)$$

is a point-mass that places all weight on $\arg\min_i Q_\pi(s, e_i)$.

Projected Gradient Descent. Starting with some policy $\pi \in \Pi$, an iteration of the projected gradient descent algorithm with constant stepsize α updates to the solution of a regularized problem

$$\pi' = \arg\min_{\bar{\pi} \in \Pi} \langle \nabla \ell(\pi), \bar{\pi} \rangle + \frac{1}{2\alpha} \|\bar{\pi} - \pi\|_2^2 = \arg\min_{\bar{\pi} \in \Pi} \langle Q_\pi, \bar{\pi} \rangle_{\eta_\pi \times 1} + \frac{1}{2\alpha} \|\bar{\pi} - \pi\|_2^2.$$

As $\alpha \to \infty$ (the regularization term tends to zero), π' converges to the solution of (4.5), which is exactly the policy iteration update as noted above. For intermediate values of α, the projected gradient update decouples across states and takes the form:

$$\pi'_s = \text{Proj}_{2, \Delta^{k-1}}(\pi_s - \alpha Q_\pi(s, \cdot))$$

which is a gradient step followed by a projection onto the probability simplex. Note that from an implementation perspective, projections onto the probability simplex involves a computationally efficient ($O(k \log k)$) soft-thresholding operation [152].

Mirror-descent. The mirror descent method adapts to the geometry of the probability simplex by using a non-euclidean regularizer. We focus on using the Kullback Leibler (KL) divergence, a natural choice for the regularizer, under which an iteration of mirror descent updates policy π to π' as:

$$\pi' = \arg\min_{\bar{\pi} \in \Pi} \langle \nabla \ell(\pi), \bar{\pi} \rangle + \frac{1}{\alpha} \sum_{s=1}^{n} D_{\mathrm{KL}}(\bar{\pi}_s \,||\, \pi_s) = \arg\min_{\bar{\pi} \in \Pi} \langle Q_\pi, \bar{\pi} \rangle_{\eta_\pi \times 1} + \frac{1}{\alpha} \sum_{s=1}^{n} D_{\mathrm{KL}}(\bar{\pi}_s \,||\, \pi_s),$$

where KL divergence is defined as $D_{\mathrm{KL}}(p||q) = \sum_{i=1}^{k} p_i \log(p_i/q_i)$. It is well know that the solution to this optimization problem is the exponentiated gradient update [38, Section 6.3],

$$\pi'_{s,i} = \frac{\pi_{s,i} \cdot \exp\{-\alpha \eta_\pi(s) Q_\pi(s, e_i)\}}{\sum_{j=1}^{k} \pi_{s,j} \cdot \exp\{-\alpha \eta_\pi(s) Q_\pi(s, e_j)\}}.$$

Again, we can see that π' converges to a policy iteration update as $\alpha \to \infty$.

Natural policy gradient and TRPO. We consider the natural policy gradient (NPG) algorithm of [77] which is closely related to the widely used TRPO algorithm of [62]. We focus on NPG applied to the *softmax parameterization* for which it is actually an instance of mirror descent with a specific regularizer. In particular, beginning with some policy $\pi \in \Pi$, an iteration of NPG updates to π':

$$\begin{aligned}
\pi' &= \arg\min_{\bar{\pi} \in \Pi} \langle \nabla \ell(\pi), \bar{\pi} \rangle + \frac{1}{\alpha} \sum_{s=1}^{n} \eta_\pi(s) D_{\mathrm{KL}}(\bar{\pi}_s \,||\, \pi_s) \\
&= \arg\min_{\bar{\pi} \in \Pi} \langle Q_\pi, \bar{\pi} \rangle_{\eta_\pi \times 1} + \frac{1}{\alpha} \sum_{s=1}^{n} \eta_\pi(s) D_{\mathrm{KL}}(\bar{\pi}_s \,||\, \pi_s), \\
&= \left(\frac{\pi_{s,i} \cdot \exp\{-\alpha Q_\pi(s, e_i)\}}{\sum_{j=1}^{k} \pi_{s,j} \cdot \exp\{-\alpha Q_\pi(s, e_j)\}} \right)_{s \in \{1, \cdots, n\}, i \in \{1, \cdots, k\}}
\end{aligned} \tag{4.6}$$

Here, we have used a natural regularizer that penalizes changes to the the action distribution at states in proportion to their occupancy measure η_π. This yields a type of soft policy iteration update at each state. As discussed above, it is well known that this KL divergence regularized problem is solved by an exponential weights update [38, Section 6.3].

A potential source of confusion is that natural policy gradient is usually described as steepest descent in a variable metric defined by a certain fisher information matrix[1]. In this case, readers can check that the exponentiated update in (4.6) matches the explicit formula for the NPG update with softmax policies as given in [77] and [99].

The choice of stepsizes is an important issue for most first order methods. Each of the algorithms above can be applied with a sequence of stepsizes $\{\alpha_t\}_{t\geq 0}$ to produce a sequence of policies $\{\pi^t\}_{t\geq 0}$. We define one stepsize selection rule below.

Exact line search. At iteration t, the update rules for each of the algorithms described above actually specify a new policy π_α^{t+1} for a range of stepsizes, $\alpha \geq 0$. We consider an idealized stepsize rule using *exact line search*, which directly optimizes over this choice of stepsize at each iteration, selecting $\pi^{t+1} = \pi_{\alpha^*}^{t+1}$ where $\alpha^* = \arg\min_\alpha \ell(\pi_\alpha^{t+1})$ whenever this minimizer exists. More generally, we define

$$\pi^{t+1} = \arg\min_{\pi \in \Pi^{t+1}} \ell(\pi). \tag{4.7}$$

where $\Pi^{t+1} = \text{Closure}(\{\pi_\alpha^{t+1}\})$ denotes the closed curve of policies traced out by varying α. For Frank-Wolfe, $\Pi^{t+1} = \{\alpha\pi^t + (1-\alpha)\pi_+^t : \alpha \in [0,1]\}$ is the line segment connecting the current policy π^t and its policy iteration update π_+^t. Under NPG, $\Pi^{t+1} = \{\pi_\alpha^{t+1}\}$ is a curve where $\pi_0^{t+1} = \pi^t$ and $\pi_\alpha^{t+1} \to \pi_+^t$ as $\alpha \to \infty$. Since π_+^t is not attainable under any fixed α, this curve is not closed. By taking the closure, and defining line search via (4.7), certain formulas become cleaner. Of course, it is also possible to nearly solve (4.7) without taking the closure and obtain essentially the same

[1]This is equivalent to mirror descent under some conditions [153].

results[2].

Remark 6 (Policy parameterization and infima vs minima)**.** *Algorithms for policy gradients are usually presented for parameterized policies. For example, a policy gradient algorithm might search over the parameter $\theta \in \mathbb{R}^{n\times k}$ of a softmax policy $\pi_\theta \in \Pi$, defined by $\pi_\theta(s,i) \propto e^{\theta_{s,i}}$. Many of the algorithms described above could equivalently be viewed as searching over the parameter space. For example, NPG updates the softmax policy π_θ by solving*

$$\arg\min_{\overline{\theta}} \ \langle Q_{\pi_\theta}, \pi_{\overline{\theta}} \rangle_{\eta_{\pi_\theta}\times 1} + \frac{1}{\alpha}\sum_{s=1}^{n} \eta_{\pi_\theta}(s) D_{\mathrm{KL}}(\pi_{\overline{\theta}}(s,\cdot) \,||\, \pi_\theta(s,\cdot)).$$

That is, it solves the same minimization problem as described above, but over the parameterized policy class. The class of softmax policies $\Pi_\Theta := \{\pi_\theta : \theta \in \mathbb{R}^{n\times d}\}$ can approximate any policy over the simplex to arbitrary precision, so there is no practical distinction between defining Frank-Wolfe, Projected Gradient Descent, Mirror Descent or NPG iterations as optimizing over the policy parameter or over the policy $\pi \in \Pi$. But mathematical analysis is much cleaner over Π because it is closed. For example, it contains an optimal policy, whereas any softmax policy $\pi_\theta \in \Pi_\Theta$ can only come infinitesimally close to an optimal policy. In practice, optimization problems are never solved beyond machine precision, so we don't view the distinction between infimum and minimum to be relevant to the paper's main insights.

4.5 Main result: geometric convergence

So far, we have described variants of policy gradients for tabular MDPs. All of these algorithms essentially make policy iteration updates when their stepsizes are large. Intuitively, it makes sense to expect that their convergence behavior closely resemble results for policy iteration rather than the analysis of gradient descent. We quantify this precisely in Theorem 9 below.

Our first result confirms that all of the algorithms we presented in the previous section converge

[2]For example, if we select a parameter α that offers half the possible improvement, meaning $\ell(\pi^t) - \ell(\pi_\alpha^{t+1}) \leq (1/2)(\ell(\pi^t) - \inf_{\alpha'} \ell(\pi_{\alpha'}^{t+1}))$, then our results follows with some extra factors of 2 in the bounds. One essentially needs to modify Equation (4.8) in the proof and the rest is the same.

geometrically if stepsizes are set by exact line search on $\ell(\cdot)$. Again, the idea is that a policy gradient *is* a policy iteration update for an appropriate choice of stepsize. Our proof effectively shows that exact line search updates make at least as much progress in reducing $\ell(\cdot)$ as a policy iteration update. The mismatch between the policy gradient loss $\ell(\cdot)$, which governs the stepsize choice, and the maximum norm, which governs policy iteration convergence, is the source of the term $\min_{s\in\mathcal{S}} \rho(s)$ in the bound.

Our second and third results show that dependence on the initial distribution can be avoided by forcing the algorithm to use appropriately large stepsizes. The simplest result applies to the Frank-Wolfe algorithm with a constant stepsize, which we already showed to be exactly equivalent to a soft policy iteration update. For softmax policies and exact gradient evaluations, we show that NPG with an *adaptive stepsize sequence* will reach an ϵ optimal policy in $O(\log(1/\epsilon))$ iterations. The error term, ϵ, is inversely related to the stepsize and reflects the fact that NPG updates with finite stepsizes only approximately resemble the policy iteration updates. As we take stepsizes to infinity, we recover the same result as one would expect for policy iteration. As compared to the first result in part (a) which applies with exact line search, the result in part (c) is useful in the sense that it gives a quantification of how large the stepsizes need to be for linear convergence to hold.

Theorem 9 (Geometric convergence)**.** *Suppose one of the first-order algorithms in Section 4.4 is applied to minimize $\ell(\pi)$ over $\pi \in \Pi$ with stepsize sequence $(\alpha_t : t \in \{0, 1, 2, \cdots\})$. Let π^0 denote the initial policy and $(\pi^t : t \in \{0, 1, 2, \cdots\})$ denote the sequence of iterates. The following bounds apply:*

(a) ***Exact line search.*** *If either Frank-Wolfe, projected gradient descent, mirror descent, or NPG is applied with stepsizes chosen by exact line search as in* (4.7)*, then*

$$\|J_{\pi^t} - J^*\|_\infty \leq \left(1 - \min_{s\in\mathcal{S}} \rho(s)(1-\gamma)\right)^t \frac{\|J_{\pi^0} - J^*\|_\infty}{\min_{s\in\mathcal{S}} \rho(s)}.$$

(b) ***Constant stepsize Frank-Wolfe.*** *Under Frank-Wolfe with constant stepsize* $\alpha \in (0, 1]$,

$$\|J_{\pi^t} - J^*\|_\infty \leq (1 - \alpha(1-\gamma))^t \|J_{\pi^0} - J^*\|_\infty.$$

(c) ***Natural policy gradient with softmax policies and adaptive stepsize.*** *Fix any* $\epsilon > 0$. *Let* $\pi^t(s, i^*)$ *be the probability of the optimal action according to the current policy* π^t. *That is,* $i^* = \arg\min_i Q^{\pi^t}(s, i)$. *Suppose that NPG is performed with an adaptive stepsize sequence*

$$\alpha_t(s) \geq \frac{2}{(1-\gamma)\epsilon} \log\left(\frac{2}{\pi^t(s, i^*)}\right)$$

Then,

$$\|J_{\pi^t} - J^*\|_\infty \leq \left(\frac{1+\gamma}{2}\right)^t \|J_{\pi^0} - J^*\|_\infty + \epsilon.$$

Remark 7. *Note that for the softmax parameterization,* $\pi_\theta(s, i) > 0$ *for any* $\theta \in \mathbb{R}^{n \times k}$. *So,* $\pi^t(s, i^*) > 0$ *for all* t. *In fact, the result in part (a) suggests that* $\pi^t(s, i^*) \to 1$ *geometrically fast. Although an explicit proof is not given here, we expect a simple argument to show that* $\pi^t(s, i^*)$ *increases monotonically.*

Proof of Theorem 9. Throughout the proof, we use some standard properties of the Bellman operator as described in Section 4.2. We denote π^t_+ to be the policy iteration update to any policy $\pi^t \in \Pi$.

Part (a): Exact line-search: Under each algorithm and at each iteration t, the policy iteration update π^t_+ is contained in the class Π^{t+1} introduced in (4.7). Therefore, for each algorithm,

$$\ell(\pi^{t+1}) = \min_{\pi \in \Pi^t} \ell(\pi) \leq \ell(\pi^t_+) \tag{4.8}$$

Recall policy improvement property in (4.3), which shows $J^* \preceq J_{\pi_+^t} \preceq TJ_{\pi^t} \preceq J_{\pi^t}$. Denote $\rho_{\min} := \min_{s\in\mathcal{S}} \rho(s)$. We have,

$$\begin{aligned}
\ell(\pi^t) - \ell(\pi^{t+1}) \geq \ell(\pi^t) - \ell(\pi_+^t) &= \sum_s \rho(s) \left(J_{\pi^t}(s) - J_{\pi_+^t}(s) \right) \\
&\geq \rho_{\min} \| J_{\pi^t} - J_{\pi_+^t} \|_\infty \\
&\geq \rho_{\min} \| J_{\pi^t} - TJ_{\pi^t} \|_\infty \\
&= \rho_{\min} \| J_{\pi^t} - J^* - (TJ_{\pi^t} - J^*) \|_\infty \\
&\geq \rho_{\min} \left(\| J_{\pi^t} - J^* \|_\infty - \| \left(TJ_{\pi^t} - J^* \right) \|_\infty \right) \\
&= \rho_{\min} \left(\| J_{\pi^t} - J^* \|_\infty - \| TJ_{\pi^t} - TJ^* \|_\infty \right) \\
&\geq \rho_{\min} \left(1 - \gamma \right) \| J_{\pi^t} - J^* \|_\infty \\
&\geq \rho_{\min} \left(1 - \gamma \right) \left(\ell(\pi^t) - \ell(\pi^*) \right).
\end{aligned}$$

Rearranging terms gives,

$$\ell(\pi^{t+1}) - \ell(\pi^*) \leq (1 - \rho_{\min}(1-\gamma)) \left(\ell(\pi^t) - \ell(\pi^*) \right) \leq \cdots \leq (1 - \rho_{\min}(1-\gamma))^t \left(\ell(\pi^0) - \ell(\pi^*) \right),$$

where the later inequalities follow by inductively applying the first one. We immediately have the looser bound $\ell(\pi^{t+1}) - \ell(\pi^*) \leq (1 - \rho_{\min}(1-\gamma))^t \| J_{\pi^0} - J^* \|_\infty$. The final result follows from observing $\| J_{\pi^t} - J^* \|_\infty \leq \left(\ell(\pi^{t+1}) - \ell(\pi^*) \right) / \rho_{\min}$.

Part (b): Constant stepsize Frank-Wolfe: The proof makes simple modifications to the classic policy iteration analysis review in Section 4.2. Recall from Section 4.4 that a Frank-Wolfe update is equivalent to a soft-policy iteration update:

$$\pi^{t+1}(s) = (1-\alpha)\pi^t(s) + \alpha \pi_+^t(s)$$

where π_+^t is the policy iteration update to π^t. Thus, starting from a feasible policy $\pi^0 \in \Pi$, we always maintain feasibility for $\alpha \in (0, 1]$. By the linearity in (4.1), for any state s,

$$\begin{aligned} T_{\pi^{t+1}} J_{\pi^t}(s) &= g(s, \pi^{t+1}(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi^{t+1}(s)) J_{\pi^t}(s) \\ &= (1-\alpha)\left[g(s, \pi^t(s)) + \alpha g(s, \pi_+^t(s))\right] + \gamma \sum_{s' \in \mathcal{S}} \left\{(1-\alpha)P(s'|s, \pi^t(s)) + \alpha P(s'|s, \pi_+^t(s)\right\} J_{\pi^t}(s') \\ &= (1-\alpha) J_{\pi^t}(s) + \alpha T J_{\pi^t}(s), \end{aligned}$$

where the last step uses that $T_{\pi_+^t} J_{\pi^t} = T J_{\pi^t}$ by definition of the policy iteration update. Using $T J_{\pi^t} \preceq J_{\pi^t}$ as in (4.2), we get

$$T_{\pi^{t+1}} J_{\pi^t} = (1-\alpha) J_{\pi^t} + \alpha T J_{\pi^t} \preceq J_{\pi^t}.$$

Monotonicity of $T_{\pi^{t+1}}$ implies $J_{\pi^t} \succeq T_{\pi^{t+1}} J_{\pi^t} \succeq T_{\pi^{t+1}}^2 J_{\pi^t} \succeq \ldots \succeq J_{\pi^{t+1}}$ using that $J_{\pi^{t+1}} = \lim_{n \to \infty} T_{\pi^{t+1}}^n J_{\pi^t}$. Therefore,

$$J_{\pi^{t+1}} \preceq T_{\pi^{t+1}} J_{\pi^t} = (1-\alpha)\left[J_{\pi^t} - T J_{\pi^t}\right].$$

Subtracting J^* from both sides shows

$$J_{\pi^{t+1}} - J^* \preceq (1-\alpha)\left(J_{\pi^t} - J^*\right) + \alpha\left(T J_{\pi^t} - J^*\right).$$

Since the above inequality holds element wise,

$$\|J_{\pi^{t+1}} - J^*\|_\infty \leq (1-\alpha)\|J_{\pi^t} - J^*\|_\infty + \alpha\|T J_{\pi^t} - J^*\|_\infty \leq \left[(1-\alpha) + \gamma\alpha\right]\|J_{\pi^t} - J^*\|_\infty,$$

where we again use that $J^* = T J^*$ and $\|T J_{\pi^t} - T J^*\|_\infty \leq \gamma \|J_{\pi^t} - J^*\|_\infty$ as $T(\cdot)$ is a contraction. Iterating over the above equation gives us our final result:

$$\|J_{\pi^{t+1}} - J^*\|_\infty \leq (1 - \alpha(1-\gamma))\|J_{\pi^t} - J^*\|_\infty \leq (1-\alpha(1-\gamma))^t \|J_{\pi^0} - J^*\|_\infty.$$

Part (c): Proof for natural policy gradient with adaptive stepsizes: Recall that for $\theta \in \mathbb{R}^{n\times k}$, the softmax policy parameterization takes action i in state s with probability $\pi_\theta(s,i)$:

$$\pi_\theta(s,i) = \frac{e^{\theta_{s,i}}}{\sum_{j=1}^{k} e^{\theta_{s,j}}} \qquad i = 1,\ldots,k.$$

As shown in Section 4.4, the natural policy gradient (NPG) updates with a stepsize sequence $\{\alpha_t\}_{t\geq 0}$ take the simple form:

$$\pi^{t+1}(s,i) = \frac{\pi^t(s,i)\cdot e^{-\alpha_t(s)Q^t(s,i)}}{\sum_{j=1}^{k}\pi^t(s,j)\cdot e^{-\alpha_t(s)Q^t(s,j)}},$$

where we use the shorthand notation $\pi^t(\cdot)$ to denote $\pi_{\theta^t}(\cdot)$ and $Q^t(s,i)$ to denote $Q_{\pi_{\theta^t}}(s,i)$. For simplicity, we let $c := \frac{2}{(1-\gamma)}$ which implies, $\alpha_t(s) \geq \frac{c}{\epsilon}\log\left(\frac{2}{\pi^t(s,i^*)}\right)$.

Our proof strategy follows essentially shows that for any state $s \in \mathcal{S}$, the NPG update stepsize $\alpha_t(s)$ decreases the probability of *sub-optimal* actions by a multiplicative factor. Informally, the set of sub-optimal actions (per state) can be understood as the set of actions with action gap[3] larger than some threshold. Essentially, this shows the NPG update is equivalent to a soft policy iteration update upto some small additive error. We divide the proof into three steps.

Step 1: NPG update for *sub-optimal* actions: Fix some state $s \in \mathcal{S}$. Without loss of generality, we assume the following ordering on the Q-values: $Q^t(s,1) < Q^t(s,2) \ldots < Q^t(s,k)$ which implies that action 1 is optimal in state s under policy π^t. For error tolerance $\epsilon > 0$, define $O_t^-(s)$ and $O_t^+(s)$ as:

$$O_t^-(s) := \left\{i \mid Q^t(s,i) - Q^t(s,1) \geq \frac{\epsilon}{c}\right\}$$
$$O_t^+(s) := \left\{i \mid Q^t(s,i) - Q^t(s,1) < \frac{\epsilon}{c}\right\}$$

[3]The action gap of any action $i \in \{1,\ldots,k\}$ is the difference between Q-values of the action and Q-value of the optimal action.

The set $O_t^-(s)$ can be interpreted as the set of *sub-optimal* actions with the *action gap*, $Q^t(s,i) - Q^t(s,1)$, larger than the threshold ϵ. Similarly, $O_t^+(s)$ can be interpreted to be the set of *nearly optimal* actions according to policy π^t. The following lemma shows that the NPG updates decrease the probability of playing sub-optimal actions by a multiplicative factor.

Lemma 30. *For any state s,* $\frac{\pi^{t+1}(s,i)}{\pi^t(s,i)} \leq \frac{1}{2} \;\; \forall\, i \in O_t^-(s)$.

Proof. The proof follows a simple argument. By definition, for any $i \in O_t^-(s)$:

$$\begin{aligned} \left(Q^t(s,i) - Q^t(s,1)\right) &\geq \frac{\epsilon}{c} \\ \implies \alpha_t(s)\left(Q^t(s,i) - Q^t(s,1)\right) &\geq \log\frac{2}{\pi^t(s,1)} \end{aligned}$$

which follows by the definition, $\alpha_t(s) \geq \frac{c}{\epsilon}\log\frac{2}{\pi^t(s,1)}$ which implies $\frac{\epsilon}{c} \geq \frac{1}{\alpha_t(s)}\log\frac{2}{\pi^t(s,1)}$. Rearranging, we get

$$\log\left(\pi^t(s,1)e^{-\alpha_t(s)Q^t(s,1)}\right) + \log\left(\frac{1}{2}\right) \geq -\alpha_t(s)Q^t(s,i)$$

This implies,

$$\log\left(\sum_{j=1}^{k}\pi^t(s,j)e^{-\alpha_t(s)Q^t(s,j)}\right) + \log\left(\frac{1}{2}\right) \geq \log\left(\pi^t(s,1)e^{-\alpha_t(s)Q^t(s,1)}\right) + \log\left(\frac{1}{2}\right) \geq -\alpha_t(s)Q^t(s,i).$$

which holds as all the terms in the summation are positive, $\pi^t(s,j)e^{-\alpha_t(s)Q^t(s,j)} > 0 \;\; \forall\, j \in \{1,2,\ldots,k\}$ and $\log(\cdot)$ is a monotonic transformation. Our result holds by noting

$$\frac{1}{2}\sum_{j=1}^{k}\pi^t(s,j)e^{-\alpha_t(s)Q^t(s,j)} \geq e^{-\alpha_t(s)Q^t(s,i)} \quad \implies \quad \frac{\pi^{t+1}(s,i)}{\pi^t(s,i)} = \frac{e^{-\alpha_t(s)Q^t(s,i)}}{\sum_{j=1}^{k}\pi^t(s,j)e^{-\alpha_t(s)Q^t(s,j)}} \leq \frac{1}{2}.$$

□

Step 2: NPG updates as soft policy iteration: Recall that the policy iteration update, $\pi_+^t(s) = \arg\min_{i\in\{1,2,\ldots,k\}} \; Q^t(s,i)$, which puts the entire mass on the best action (according to Q-values) and zeros out the probability of playing other actions. On the other hand, Lemma 30 shows how

an NPG update with appropriate stepsize decays the probabilities of *sub-optimal* actions (in the set $O_t^-(s)$) by a multiplicative factor instead of zeroing them out[4]. This resembles a *soft* policy iteration update for the set of actions $O_t^-(s)$. We formalize this intuition in the following lemma which characterizes the progress made by an NPG update vis-a-vis a policy iteration update.

Lemma 31 (Progress quantification)**.** *Let $J_{\pi^t}(s)$ denote the cost-to-go function for policy π^t from any starting state s. Then,*

$$T_{\pi^{t+1}} J_{\pi^t}(s) - J_{\pi^t}(s) \leq \frac{1}{2} \cdot (T J_{\pi^t}(s) - J_{\pi^t}(s)) + \frac{\epsilon}{c}$$

Proof. Recall, we assumed that: $Q^t(s,1) < Q^t(s,2) \ldots < Q^t(s,k)$ which implies that the policy iteration update, π_t^+ puts the entire mass on action 1. That is, $\pi_+^t(s,1) = 1$ and $\pi_+^t(s,i) = 0 \ \forall \ i \neq 1$.

[4]This defintion of sub-optimal actions based on action gap threshold, ϵ/c, is essentially an artifact that we are taking gradient steps with finite stepsizes. As $\alpha_t(s) \to \infty \ \forall \ s$, NPG update equivalent to a soft-policy iteration update.

Consider,

$$
\begin{aligned}
T_{\pi^{t+1}} J_{\pi^t}(s) - T J_{\pi^t}(s) &= \langle \pi^{t+1}(s,\cdot) - \pi_t^+(s,\cdot), Q^t(s,\cdot) \rangle \\
&= (\pi^{t+1}(s,1) - 1) Q^t(s,1) + \sum_{j=2}^{k} \pi^{t+1}(s,j) Q^t(s,j) \\
&= -\sum_{j=2}^{k} \pi^{t+1}(s,j) Q^t(s,1) + \sum_{j=2}^{k} \pi^{t+1}(s,j) Q^t(s,j) \\
&= \sum_{j=2}^{k} \pi^{t+1}(s,j) \left(Q^t(s,j) - Q^t(s,1) \right) \\
&= \sum_{j \in O_t^-} \pi^{t+1}(s,j) \left(Q^t(s,j) - Q^t(s,1) \right) + \sum_{j \in O_t^+} \pi^{t+1}(s,j) \left(Q^t(s,j) - Q^t(s,1) \right) \\
&= \sum_{j \in O_t^-} \frac{\pi^{t+1}(s,j)}{\pi^t(s,j)} \pi^t(s,j) \left(Q^t(s,j) - Q^t(s,1) \right) + \sum_{j \in O_t^+} \pi^{t+1}(s,j) \underbrace{\left(Q^t(s,j) - Q^t(s,1) \right)}_{< \frac{\epsilon}{c}} \\
&\leq \frac{1}{2} \sum_{j \in O_t^-} \pi^t(s,j) \left(Q^t(s,j) - Q^t(s,1) \right) + \frac{\epsilon}{c} \\
&\leq \frac{1}{2} \left(\sum_{j=2}^{k} \pi^t(s,j) (Q^t(s,j) - Q^t(s,1)) \right) + \frac{\epsilon}{c} \\
&= \frac{1}{2} \left(\sum_{j=2}^{k} \pi^t(s,j) Q^t(s,j) - \sum_{j=2}^{k} \pi^t(s,j) Q^t(s,1) \right) + \frac{\epsilon}{c} \\
&= \frac{1}{2} \left(\left(\pi^t(s,j) - 1 \right) Q^t(s,1) + \sum_{j=2}^{k} \pi^t(s,j) Q^t(s,j) \right) + \frac{\epsilon}{c} \\
&= \frac{1}{2} \langle \pi^t(s,\cdot) - \pi_t^+(s,\cdot), Q^t(s,\cdot) \rangle + \frac{\epsilon}{c} \\
&= \frac{1}{2} \left(J_{\pi^t}(s) - T J_{\pi^t}(s) \right) + \frac{\epsilon}{c} \qquad (4.9)
\end{aligned}
$$

where we used that $\frac{\pi^{t+1}(s,j)}{\pi^t(s,j)} \leq \frac{1}{2} \; \forall \; j \in O_t^-(s)$ as shown above in Lemma 30 along with the fact that $\left(Q^t(s,j) - Q^t(s,1) \right) \leq \frac{\epsilon}{c} \; \forall \; j \in O_t^+(s)$ which follows by definition. Subtracting $J_{\pi^t}(s)$ from both sides in Equation (4.9) and rearranging terms gives our desired result

$$
T_{\pi^{t+1}} J_{\pi^t}(s) - J_{\pi^t}(s) \leq \frac{1}{2} \cdot (T J_{\pi^t}(s) - J_{\pi^t}(s)) + \frac{\epsilon}{c}
$$

□

Step 3: Completing the proof: Lemma 31 clearly quantifies the relationship between an NPG update with adaptive stepsize α_t and a soft policy iteration update with an additive error $\frac{\epsilon}{c}$. With this connection, we give a simple proof of geometric convergence for the natural policy gradient method. First, we claim that $J_{\pi^{t+1}}(s) \leq J_{\pi^t}(s)$. To see this, note that for any stepsize $\alpha(s)$, the NPG update for state s can be equivalently written as:

$$\pi^{t+1}(s) = \arg\min_{a \in \Delta^{k-1}} \left[Q^t(s,a) + \frac{\eta_{\pi^t(s)}}{\alpha(s)} D_{\mathrm{KL}}(a||\pi^t(s)) \right]$$

As $a = \pi^t(s)$ is feasible for the optimization problem above,

$$T_{\pi^{t+1}} J_{\pi^t}(s) = Q^t(s, \pi^{t+1}(s)) \leq Q^t(s, \pi^t(s)) = J_{\pi^t}(s)$$

By monotonicity property of $T_{\pi^{t+1}}$, we have $J_{\pi^t} \geq T_{\pi^{t+1}} J_{\pi^t} \geq T^2_{\pi^{t+1}} J_{\pi^t} \geq \ldots \geq J_{\pi^{t+1}}$ by noting that $J_{\pi^{t+1}} = \lim_{n \to \infty} T^n_{\pi^{t+1}} J_{\pi^t}$. From Lemma 31, we get that

$$J_{\pi^{t+1}} - J_{\pi^t} \leq T_{\pi^{t+1}} J_{\pi^t} - J_{\pi^t} \leq \frac{1}{2} \cdot (T J_{\pi^t} - J_{\pi^t}) + \frac{\epsilon}{c}$$

Subtracting J^* from both sides and rearranging terms,

$$J_{\pi^{t+1}} - J^* \leq \frac{1}{2} J_{\pi^t} + \frac{1}{2} T J_{\pi^t} - J^* + \frac{\epsilon}{c} = \frac{1}{2} (J_{\pi^t} - J^*) + \frac{1}{2} (T J_{\pi^t} - J^*) + \frac{\epsilon}{c}.$$

As the above inequality holds element wise, this implies,

$$\|J_{\pi^{t+1}} - J^*\|_\infty \leq \frac{1}{2} \|J_{\pi^t} - J^*\|_\infty + \frac{1}{2} \|T J_{\pi^t} - J^*\|_\infty + \frac{\epsilon}{c} \leq \left[\frac{1}{2} + \frac{\gamma}{2} \right] \|J_{\pi^t} - J^*\|_\infty + \frac{\epsilon}{c}$$

where we used that $\|TJ_{\pi^t} - J^*\|_\infty = \|TJ_{\pi^t} - TJ^*\|_\infty \leq \gamma \|J_{\pi^t} - J^*\|_\infty$ as shown in (4.4). Rewriting $\left(\frac{1}{2} + \frac{\gamma}{2}\right) = \left(1 - \frac{1}{2}(1-\gamma)\right)$ and iterating over the above equation gives us our final result.

$$\begin{aligned}
\|J_{\pi^{t+1}} - J^*\|_\infty &\leq \left(1 - \frac{(1-\gamma)}{2}\right) \|J_{\pi^t} - J^*\|_\infty + \frac{\epsilon}{c} \\
&\leq \left(1 - \frac{(1-\gamma)}{2}\right)^t \|J_{\pi^0} - J^*\|_\infty + \frac{\epsilon}{c} \sum_{i=0}^{t-1} \left(1 - \frac{(1-\gamma)}{2}\right)^i \\
&\leq \left(1 - \frac{(1-\gamma)}{2}\right)^t \|J_{\pi^0} - J^*\|_\infty + \frac{2}{(1-\gamma)} \frac{\epsilon}{c}.
\end{aligned}$$

□

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. 2018.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[3] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *arXiv preprint arXiv:2002.00444*, 2020.

[4] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[5] M. Popova, O. Isayev, and A. Tropsha, "Deep reinforcement learning for de novo drug design," *Science advances*, vol. 4, no. 7, eaap7885, 2018.

[6] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.

[7] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," in *Advances in Neural Information Processing Systems 7*, 1994, pp. 703–710.

[8] L. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Machine Learning Proceedings*, 1995, pp. 30–37.

[9] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674 –690, 1997.

[10] N. Korda and P. L.A., "On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 626–634.

[11] C. Lakshminarayanan and C. Szepesvári, "Finite Time Bounds for Temporal Difference Learning with Function Approximation: Problems with some "state-of-the-art" results," *https://sites.ualberta.ca/ szepesva/papers/TD-issues17.pdf*, 2017.

[12] R. S. Sutton, C. Szepesvári, and H. R. Maei, "A Convergent O(n) Temporal-difference Algorithm for Off-policy Learning with Linear Function Approximation," in *Advances in Neural Information Processing Systems*, 2009, pp. 1609–1616.

[13] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. 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, pp. 993–1000.

[14] B. Liu, J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik, "Finite-sample Analysis of Proximal Gradient TD Algorithms," in *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, 2015, pp. 504–513.

[15] A. Touati, P.-L. Bacon, D. Precup, and P. Vincent, "Convergent TREE BACKUP and RETRACE with function approximation," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 4962–4971.

[16] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, "Finite Sample Analysis of Two-Timescale Stochastic Approximation with Applications to Reinforcement Learning," in *Proceedings of the 31st Conference On Learning Theory*, 2018, pp. 1199–1233.

[17] C. Lakshminarayanan and C. Szepesvári, "Linear stochastic approximation: How far does constant step-size and iterate averaging go?" In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1347–1355.

[18] H. Kushner, "Stochastic approximation: A survey," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 87–96, 2010.

[19] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 1998.

[20] J. N. Tsitsiklis and B. Van Roy, "Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives," *IEEE Transactions on Automatic Control*, vol. 44, no. 10, pp. 1840 –1851, 1999.

[21] L. Andersen and M. Broadie, "Primal-dual simulation algorithm for pricing multidimensional american options," *Management Science*, vol. 50, no. 9, pp. 1222–1234, 2004.

[22] M. B. Haugh and L. Kogan, "Pricing American options: A duality approach," *Operations Research*, vol. 52, no. 2, pp. 258–270, 2004.

[23] V. V. Desai, V. F. Farias, and C. C. Moallemi, "Pathwise optimization for optimal stopping problems," *Management Science*, vol. 58, no. 12, pp. 2292–2308, 2012.

[24] D. A. Goldberg and Y. Chen, "Beating the curse of dimensionality in options pricing and optimal stopping," *arXiv preprint arXiv:1807.02227*, 2018.

[25] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, "Finite Sample Analyses for TD(0) With Function Approximation," in *AAAI*, 2018.

[26] D. Ruppert, "Efficient estimations from a slowly convergent robbins-monro process," Cornell University Operations Research and Industrial Engineering, Tech. Rep., 1988.

[27] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[28] L. Györfi and H. Walk, "On the averaged stochastic approximation for linear regression," *SIAM Journal on Control and Optimization*, vol. 34, no. 1, pp. 31–61, 1996.

[29] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 773–781.

[30] R. E. Schapire and M. K. Warmuth, "On the worst-case analysis of temporal-difference learning algorithms," *Machine Learning*, vol. 22, no. 1-3, pp. 95–121, 1996.

[31] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

[32] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

[33] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media, 2012, vol. 22.

[34] V. S. Borkar and S. P. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 447–469, 2000.

[35] V. R. Konda, "Actor-Critic Algorithms," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[36] A. M. Devraj and S. P. Meyn, "Zap Q-Learning," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 2235–2244.

[37] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[38] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[39] P. Jain and P. Kar, "Non-convex optimization for machine learning," *Foundations and Trends® in Machine Learning*, vol. 10, no. 3-4, pp. 142–336, 2017.

[40] A. Antos, C. Szepesvári, and R. Munos, "Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path," *Machine Learning*, vol. 71, no. 1, pp. 89–129, 2008.

[41] A. Lazaric, M. Ghavamzadeh, and R. Munos, "Finite-sample analysis of LSTD," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 615–622.

[42] M. Ghavamzadeh, A. Lazaric, O. Maillard, and R. Munos, "LSTD with Random Projections," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 721–729.

[43] B. Á. Pires and C. Szepesvári, "Statistical linear estimation with penalized estimators: An application to reinforcement learning," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 1755–1762.

[44] P. L.A., N. Korda, and R. Munos, "Fast LSTD using stochastic approximation: Finite time analysis and application to traffic control," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2013, pp. 66–81.

[45] S. Tu and B. Recht, "Least-squares temporal difference learning for the linear quadratic regulator," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 5012–5021.

[46] H. Yu and D. P. Bertsekas, "Convergence results for some temporal difference methods based on least squares," *IEEE Transactions on Automatic Control*, vol. 54, no. 7, pp. 1515–1531, 2009.

[47] S. V. Macua, J. Chen, S. Zazo, and A. H. Sayed, "Distributed policy evaluation under multiple behavior strategies," *IEEE Transactions on Automatic Control*, vol. 60, no. 5, pp. 1260–1274, 2014.

[48] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, 2012, vol. 2.

[49] C. Dann, G. Neumann, and J. Peters, "Policy evaluation with temporal differences: A survey and comparison," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 809–883, 2014.

[50] D. P. De Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Operations research*, vol. 51, no. 6, pp. 850–865, 2003.

[51] D. P. Bertsekas and S. Shreve, *Stochastic optimal control: the discrete-time case*. Athena Scientific Belmont, MA, 2004.

[52] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[53] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[54] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine learning*, vol. 22, no. 1-3, pp. 33–57, 1996.

[55] J. C. Duchi, "Introductory lectures on stochastic optimization," *The Mathematics of Data*, vol. 25, p. 99, 2018.

[56] S. Lacoste-Julien, M. Schmidt, and F. Bach, "A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method," *arXiv preprint arXiv:1212.2002*, 2012.

[57] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Society, 2017, vol. 107.

[58] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[59] G. O. Roberts and J. S. Rosenthal, "General state space markov chains and mcmc algorithms," *Probability surveys*, vol. 1, pp. 20–71, 2004.

[60] H. van Seijen and R. S. Sutton, "True online TD(λ)," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 692–700.

[61] B. Van Roy, "Learning and value function approximation in complex decision processes," Ph.D. dissertation, Massachusetts Institute of Technology, 1998.

[62] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, 2015, pp. 1889–1897.

[63] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[64] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[65] P. Marbach and J. N. Tsitsiklis, "Simulation-based optimization of markov reward processes," *IEEE Transactions on Automatic Control*, vol. 46, no. 2, pp. 191–209, 2001.

[66] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems*, 2000, pp. 1057–1063.

[67] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[68] J. Peters and S. Schaal, "Policy gradient methods for robotics," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2006, pp. 2219–2225.

[69] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[70] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1467–1476.

[71] S. Kunnumkal and H. Topaloglu, "Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems," *Operations Research*, vol. 56, no. 3, pp. 646–664, 2008.

[72] B. T. Polyak, "Gradient methods for the minimisation of functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 864–878, 1963.

[73] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[74] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2016, pp. 795–811.

[75] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[76] S. Amari, "Natural gradient works efficiently in learning," *Neural computation*, vol. 10, no. 2, pp. 251–276, 1998.

[77] S. M. Kakade, "A natural policy gradient," in *Advances in Neural Information Processing Systems*, 2002, pp. 1531–1538.

[78] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *Proceedings of the 19th International Conference on Machine Learning*, vol. 2, 2002, pp. 267–274.

[79] O. Nachum, M. Norouzi, and D. Schuurmans, "Improving policy gradient by exploring under-appreciated rewards," *CoRR*, vol. abs/1611.09321, 2017.

[80] M. Plappert, R. Houthooft, P. Dhariwal, S. Sidor, R. Y. Chen, X. Chen, T. Asfour, P. Abbeel, and M. Andrychowicz, "Parameter space noise for exploration," *arXiv preprint arXiv:1706.01905*, 2017.

[81] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*, 2000, pp. 1008–1014.

[82] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 1928–1937.

[83] M. Riedmiller, J. Peters, and S. Schaal, "Evaluation of policy gradient methods and variants on the cart-pole benchmark," in *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, IEEE, 2007, pp. 254–261.

[84] H. Mania, A. Guy, and B. Recht, "Simple random search of static linear policies is competitive for reinforcement learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 1800–1809.

[85] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," *arXiv preprint arXiv:1703.03864*, 2017.

[86] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996, vol. 5.

[87] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

[88] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on Learning Theory*, 2016, pp. 1246–1257.

[89] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, "Finding approximate local minima faster than gradient descent," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, ACM, 2017, pp. 1195–1199.

[90] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1724–1732.

[91] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for nonconvex optimization," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1751–1772, 2018.

[92] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, 2015, pp. 797–842.

[93] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere i: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[94] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[95] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[96] K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems*, 2016, pp. 586–594.

[97] P. Thomas, "Bias in natural actor-critic algorithms," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 441–448.

[98] B. Scherrer and M. Geist, "Local policy search in a convex space and conservative policy iteration as boosted policy search," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2014, pp. 35–50.

[99] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and approximation with policy gradient methods in markov decision processes," *arXiv preprint arXiv:1908.00261*, 2019.

[100] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[101] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995.

[102] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*. Springer Science & Business Media, 2012, vol. 30.

[103] S. E. Shreve and D. P. Bertsekas, "Universally measurable policies in dynamic programming," *Mathematics of Operations Research*, vol. 4, no. 1, pp. 15–30, 1979.

[104] J. Fu, A. Singh, D. Ghosh, L. Yang, and S. Levine, "Variational inverse control with events: A general framework for data-driven reward definition," in *Advances in Neural Information Processing Systems*, 2018, pp. 8538–8547.

[105] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[106] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information Fusion*, vol. 58, pp. 52–68, 2020.

[107] H. Zhu, J. Yu, A. Gupta, D. Shah, K. Hartikainen, A. Singh, V. Kumar, and S. Levine, "The ingredients of real-world robotic reinforcement learning," *arXiv preprint arXiv:2004.12570*, 2020.

[108] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[109] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[110] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075, 2014.

[111] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*, 2014, pp. 1646–1654.

[112] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 314–323.

[113] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Mathematical Programming*, vol. 156, no. 1-2, pp. 59–99, 2016.

[114] S. J. Reddi, S. Sra, B. Poczos, and A. J. Smola, "Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 1145–1153.

[115] S. J. Reddi, S. Sra, B. Póczos, and A. Smola, "Stochastic frank-wolfe methods for nonconvex optimization," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2016, pp. 1244–1251.

[116] D. Davis and B. Grimmer, "Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems," *SIAM Journal on Optimization*, vol. 29, no. 3, pp. 1908–1930, 2019.

[117] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.

[118] A. Beck, *First-order methods in optimization*. SIAM, 2017, vol. 25.

[119] A. Beck, "Convergence rate analysis of gradient based algorithms," Ph.D. dissertation, Tel-Aviv University, 2002.

[120] D. P. Bertsekas, "Approximate policy iteration: A survey and some new methods," *Journal of Control Theory and Applications*, vol. 9, no. 3, pp. 310–335, 2011.

[121] L. C. Evans, "An introduction to mathematical optimal control theory," *Lecture Notes, University of California, Department of Mathematics, Berkeley*, 2005.

[122] D. Kleinman, "On an iterative technique for riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, pp. 114 –115, 1 1968.

[123] G Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.

[124] H. T. Toivonen, "A globally convergent algorithm for the optimal constant output feedback problem," *International Journal of Control*, vol. 41, no. 6, pp. 1589–1599, 1985.

[125] T. Rautert and E. W. Sachs, "Computational design of optimal output feedback controllers," *SIAM Journal on Optimization*, vol. 7, no. 3, pp. 837–852, 1997.

[126] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedigs of the 31st International Conference on Machine Learning*, 2014.

[127] K. Ciosek and S. Whiteson, "Expected policy gradients for reinforcement learning.," *Journal of Machine Learning Research*, vol. 21, no. 52, pp. 1–51, 2020.

[128] G. C. Pflug, "Derivatives of probability measures-concepts and applications to the optimization of stochastic systems," in *Discrete Event Systems: Models and Applications*, Springer, 1988, pp. 252–274.

[129] G. C. Pflug, "On-line optimization of simulated markovian processes," *Mathematics of Operations Research*, vol. 15, no. 3, pp. 381–395, 1990.

[130] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, "Monte carlo gradient estimation in machine learning," *arXiv preprint arXiv:1906.10652*, 2019.

[131] M. C. Fu, "Gradient estimation," *Handbooks in operations research and management science*, vol. 13, pp. 575–616, 2006.

[132] S. Asmussen and P. W. Glynn, *Stochastic simulation: algorithms and analysis*. Springer Science & Business Media, 2007, vol. 57.

[133] I. Osband, B. Van Roy, D. Russo, and Z. Wen, "Deep exploration via randomized value functions," *arXiv preprint arXiv:1703.07608*, 2017.

[134] R. Munos, "Error bounds for approximate policy iteration," in *Proceedings of the 20th International Conference on Machine Learning*, vol. 3, 2003, pp. 560–567.

[135] R. Munos, "Performance bounds in l_p-norm for approximate value iteration," *SIAM journal on control and optimization*, vol. 46, no. 2, pp. 541–561, 2007.

[136] R. Munos and C. Szepesvári, "Finite-time bounds for fitted value iteration," *Journal of Machine Learning Research*, vol. 9, no. May, pp. 815–857, 2008.

[137] A. Farahmand, C. Szepesvári, and R. Munos, "Error propagation for approximate policy and value iteration," in *Advances in Neural Information Processing Systems*, 2010, pp. 568–576.

[138] M. Geist, B. Piot, and O. Pietquin, "Is the bellman residual a bad proxy?" In *Advances in Neural Information Processing Systems*, 2017, pp. 3205–3214.

[139] B. Scherrer, "Approximate policy iteration schemes: A comparison," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 1314–1322.

[140] J. N. Tsitsiklis and B. Van Roy, "Regression methods for pricing complex american-style options," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 694–703, 2001.

[141] Y. Nesterov, *Lectures on convex optimization*. Springer, 2018, vol. 137.

[142] P. L'Ecuyer and P. W. Glynn, "Stochastic optimization by simulation: Convergence proofs for the gi/g/1 queue in steady-state," *Management Science*, vol. 40, no. 11, pp. 1562–1578, 1994.

[143] I. Karaesmen and G. Van Ryzin, "Overbooking with substitutable inventory classes," *Operations Research*, vol. 52, no. 1, pp. 83–104, 2004.

[144] G. Van Ryzin and G. Vulcano, "Simulation-based optimization of virtual nesting controls for network revenue management," *Operations Research*, vol. 56, no. 4, pp. 865–880, 2008.

[145] A. Rajeswaran, K. Lowrey, E. V. Todorov, and S. M. Kakade, "Towards generalization and simplicity in continuous control," in *Advances in Neural Information Processing Systems*, 2017, pp. 6550–6561.

[146] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Reinforcement learning with soft state aggregation," in *Advances in neural information processing systems*, 1995, pp. 361–368.

[147] N. Ferns, P. Panangaden, and D. Precup, "Metrics for finite markov decision processes," in *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, AUAI Press, 2004, pp. 162–169.

[148] D. P. Bertsekas, "Feature-based aggregation and deep reinforcement learning: A survey and some new implementations," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 1–31, 2019.

[149] L. Shani, Y. Efroni, and S. Mannor, "Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps," *arXiv preprint arXiv:1909.02769*, 2019.

[150] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *arXiv preprint arXiv:2007.06558*, 2020.

[151] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," *arXiv preprint arXiv:2005.06392*, 2020.

[152] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l_1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 272–279.

[153] G. Raskutti and S. Mukherjee, "The information geometry of mirror descent," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1451–1457, 2015.

[154] R. A. Howard, *Dynamic programming and markov processes*. John Wiley, 1960.

[155] S. B. Thrun, "Efficient exploration in reinforcement learning," Technical Report CMU-CS-92-102, School of Computer Science, Carnegie Mellon, Tech. Rep., 1992.

[156] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for markov decision processes," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008.

[157] I. Osband, D. Russo, and B. Van Roy, "More efficient reinforcement learning via posterior sampling," in *Advances in Neural Information Processing Systems*, 2013, pp. 3003–3011.

[158] S. Koenig and R. G. Simmons, "Complexity analysis of real-time reinforcement learning," in *AAAI*, 1993, pp. 99–107.

[159] M. G. Azar, R. Munos, M. Ghavamzadeh, and H. Kappen, "Reinforcement learning with a near optimal rate of convergence," 2011.

[160] A. L. Peressini, F. E. Sullivan, and J. J. Uhl, *The mathematics of nonlinear programming*. Springer-Verlag, New York, 1988.

[161] Y. Fang, K. A. Loparo, and X. Feng, "Inequalities for the trace of matrix product," *IEEE Transactions on Automatic Control*, vol. 39, no. 12, pp. 2489–2490, 1994.

[162] G. J. Gordon, "Stable function approximation in dynamic programming," in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 261–268.

[163] J. N. Tsitsiklis and B. Van Roy, "Feature-based methods for large scale dynamic programming," *Machine Learning*, vol. 22, no. 1-3, pp. 59–94, 1996.

[164] B. Van Roy, "Performance loss bounds for approximate value iteration with state aggregation," *Mathematics of Operations Research*, vol. 31, no. 2, pp. 234–244, 2006.

[165] W. Whitt, "Approximations of dynamic programs," *Mathematics of Operations Research*, vol. 3, no. 3, pp. 231–243, 1978.

[166] J. Rust, "Using randomization to break the curse of dimensionality," *Econometrica: Journal of the Econometric Society*, pp. 487–516, 1997.

[167] R. Ortner and D. Ryabko, "Online regret bounds for undiscounted continuous reinforcement learning," in *Advances in Neural Information Processing Systems*, 2012, pp. 1763–1771.

Appendix A: Proofs for Chapter 2

A.1 Analysis of Projected TD(0) under Markov chain sampling model

In this section, we complete the proof of Theorem 3. The first subsection restates the theorem, as well as the two key lemmas from Section 2.9 that underly the proof. The second subsection contains a proof of Theorem 3. Finally, Subsection A.1.3 contains the proof of a technical result, Lemma 10, which was omitted from the main text but we need for the proof.

A.1.1 Restatement of the theorem and key lemmas from the main text

Theorem 3. *Suppose the Projected TD algorithm is applied with parameter $R \geq \|\theta^*\|_2$ under the Markov chain observation model with Assumption 1. Set $G = (r_{\max} + 2R)$. Then the following claims hold.*

(a) With a constant step-size sequence $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left(9 + 12\tau^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\gamma)}.$$

(b) With a constant step-size sequence $\alpha_0 = \cdots = \alpha_T < 1/(2\omega(1-\gamma))$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_T}\right\|_D^2\right] \leq \left(e^{-2\alpha_0(1-\gamma)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \alpha_0\left(\frac{G^2\left(9 + 12\tau^{\text{mix}}(\alpha_0)\right)}{2(1-\gamma)\omega}\right).$$

(c) With a decaying step-size sequence $\alpha_t = 1/(\omega(t+1)(1-\gamma))$ for all $t \in \mathbb{N}_0$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{G^2\left(9 + 24\tau^{\text{mix}}(\alpha_T)\right)}{T(1-\gamma)^2\omega}\left(1 + \log T\right),$$

The key to our proof is the following lemmas, which were establish in Section 2.9. Recall the definition of the semi-gradient error $\zeta_t(\theta) \equiv (g_t(\theta) - \bar{g}(\theta))^\top (\theta - \theta^*)$.

Lemma 8. *With probability 1, for every $t \in \mathbb{N}_0$,*

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2.$$

Lemma 11 (Bound on semi-gradient bias)**.** *Consider a non-increasing step-size sequence, $\alpha_0 \geq \alpha_1 \ldots \geq \alpha_T$. Fix any $t < T$, and set $t^* \equiv \max\{0, t - \tau^{\text{mix}}(\alpha_T)\}$. Then,*

$$\mathbb{E}\left[\zeta_t(\theta_t)\right] \leq G^2 \left(4 + 6\tau^{\text{mix}}(\alpha_T)\right) \alpha_{t^*}.$$

The following bound also holds:

$$\mathbb{E}\left[\zeta_t(\theta_t)\right] \leq 6G^2 \sum_{i=0}^{t-1} \alpha_i.$$

A.1.2 Proof of Theorem 3.

We now complete the proof of Theorem 3. The proof directly uses Lemma 8 and Lemma 11.

Proof. From Lemma 8, we have

$$\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] \leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha_t(1-\gamma)\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] + 2\alpha_t\mathbb{E}\left[\zeta_t(\theta_t)\right] + \alpha_t^2 G^2. \quad \text{(A.1)}$$

Proof of part (a): We first show the analysis for a constant step-size and iterate averaging. Considering $\alpha_t = \alpha_0 = 1/\sqrt{T}$ in Equation (A.1), rearranging terms and summing from $t = 0$ to $t = T-1$, we get

$$2\alpha_0(1-\gamma)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] \leq \sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - \mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right]\right) + G^2 + 2\alpha_0\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right].$$

Using Lemma 11 (in which $\alpha_{t^*} = \alpha_0$ in this case) and simplifying, we find

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2}{2\alpha_0(1-\gamma)} + \frac{T \cdot 2G^2(2 + 3\tau^{\text{mix}}(1/\sqrt{T}))\alpha_0}{(1-\gamma)}$$
$$= \frac{\sqrt{T}\left(\|\theta^* - \theta_0\|_2^2 + G^2\right)}{2(1-\gamma)} + \frac{\sqrt{T} \cdot 2G^2(2 + 3\tau^{\text{mix}}(1/\sqrt{T}))}{(1-\gamma)}.$$

This gives us our desired result,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left(9 + 12\tau^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\gamma)}.$$

Proof of part (b): The proof is analogous to part (b) of Theorem 2. Consider a constant step-size of $\alpha_0 < 1/(2\omega(1-\gamma))$. Starting with Equation (A.1) and applying Lemma 1, which showed $\|V_{\theta^*} - V_\theta\|_D^2 \geq \omega\|\theta^* - \theta\|_2^2$ for all θ, we get

$$\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] \leq (1 - 2\alpha_0(1-\gamma)\omega)\,\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + \alpha_0^2 G^2 + 2\alpha_0 \mathbb{E}\left[\zeta_t(\theta_t)\right]$$
$$\leq (1 - 2\alpha_0(1-\gamma)\omega)\,\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + \alpha_0^2 G^2\left(9 + 12\tau^{\text{mix}}(\alpha_0)\right),$$

where we used Lemma 11 to go to the second inequality. Iterating over this inequality gives us our final result. For any $T \in \mathbb{N}_0$,

$$\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] \leq (1 - 2\alpha_0(1-\gamma)\omega)^T \|\theta^* - \theta_0\|_2^2 + \alpha_0^2 G^2\left(9 + 12\tau^{\text{mix}}(\alpha_0)\right)\sum_{t=0}^{\infty}(1 - 2\alpha_0(1-\gamma)\omega)^t$$
$$\leq \left(e^{-2\alpha_0(1-\gamma)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \frac{\alpha_0 G^2\left(9 + 12\tau^{\text{mix}}(\alpha_0)\right)}{2(1-\gamma)\omega}.$$

Final inequality follows by solving the geometric series and using that $(1 - 2\alpha_0(1-\gamma)\omega) \leq e^{-2\alpha_0(1-\gamma)\omega}$ along with Lemma 1.

Proof of part (c): We now show the analysis for a linearly decaying step-size using Equation (A.1) as our starting point. We again use Lemma 1, which showed $\|V_{\theta^*} - V_\theta\|_D^2 \geq \omega\|\theta^* - \theta\|_2^2$ for

all θ, to get,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] \leq \frac{1}{(1-\gamma)\alpha_t}\left((1-(1-\gamma)\omega\alpha_t)\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - \mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] + \alpha_t^2 G^2\right) + \frac{2}{(1-\gamma)}\mathbb{E}\left[\zeta_t(\theta_t)\right].$$

Consider a decaying step-size $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$, simplify and sum from $t = 0$ to $T-1$ to get

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] \leq \underbrace{-\omega T\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right]}_{<0} + \frac{G^2}{\omega(1-\gamma)^2}\sum_{t=0}^{T-1}\frac{1}{t+1} + \frac{2}{(1-\gamma)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right]. \tag{A.2}$$

To simplify notation, for the remainder of the proof put $\tau = \tau^{\text{mix}}(\alpha_T)$. We can decompose the sum of semi-gradient errors as

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right] = \sum_{t=0}^{\tau}\mathbb{E}\left[\zeta_t(\theta_t)\right] + \sum_{t=\tau+1}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right]. \tag{A.3}$$

We will upper bound each term. In each case we use that, since $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$,

$$\sum_{t=0}^{T-1}\alpha_t = \frac{1}{\omega(1-\gamma)}\sum_{t=0}^{T-1}\frac{1}{(t+1)} \leq \frac{1+\log T}{\omega(1-\gamma)}.$$

Combining this with Lemma 11 gives,

$$\sum_{t=0}^{\tau}\mathbb{E}\left[\zeta_t(\theta_t)\right] \leq \sum_{t=0}^{\tau}\left(6G^2\sum_{i=0}^{t-1}\alpha_i\right) \leq \tau\left(6G^2\sum_{i=0}^{T-1}\alpha_i\right) \leq \frac{6G^2\tau}{\omega(1-\gamma)}(1+\log T).$$

Similarly, using Lemma 11, we have

$$\sum_{t=\tau+1}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right] \leq 2G^2(2+3\tau)\sum_{t=\tau+1}^{T-1}\alpha_{t-\tau} \leq 2G^2(2+3\tau)\sum_{t=1}^{T-1}\alpha_t \leq \frac{2G^2(2+3\tau)}{\omega(1-\gamma)}(1+\log T).$$

Combining the two parts, we get

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t)\right] \leq \frac{4G^2\,(1+3\tau)}{\omega(1-\gamma)}(1+\log T).$$

Using this in conjunction with Equation (A.2) we get,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|V_{\theta_t} - V_{\theta^*}\right\|_D^2\right] \leq \frac{G^2}{\omega T(1-\gamma)^2}\,(1+\log T) + \frac{2}{T(1-\gamma)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t)\right].$$

Simplifying and substituting $\tau = \tau^{\text{mix}}(\alpha_T)$, we give final result.

$$\begin{aligned}\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] &\leq \frac{G^2}{\omega T(1-\gamma)^2}\,(1+\log T) + \frac{8G^2\left(1+3\tau^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\gamma)^2}(1+\log T)\\ &\leq \frac{G^2\left(9+24\tau^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\gamma)^2}\,(1+\log T)\,.\end{aligned}$$

□

We remark that Equation (A.2) also gives us a convergence rate of $O(\log T/T)$ for the iterate θ_T itself (and hence on the value function V_{θ_T}) but the bound degrades by a factor of ω. In particular, we have

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_T}\right\|_D^2\right] \leq \mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] \leq \frac{G^2\left(9+24\tau^{\text{mix}}(\alpha_T)\right)}{\omega^2 T(1-\gamma)^2}\,(1+\log T)\,. \tag{A.4}$$

where the first inequality follows using Lemma 1.

A.1.3 Proof of Lemma 10

Lemma 10 (Semi-gradient error is bounded and Lipschitz)**.** *With probability 1,*

$$|\zeta_t(\theta)| \leq 2G^2 \qquad \forall\,\theta \in \Theta_R$$

and

$$|\zeta_t(\theta) - \zeta_t(\theta')| \leq 6G\|(\theta - \theta')\|_2 \qquad \forall\, \theta, \theta' \in \Theta_R.$$

Proof. The first claim follows from a simple argument using Lemma 6.

$$|\zeta_t(\theta)| = \left|(g_t(\theta) - \bar{g}(\theta))^\top (\theta - \theta^*)\right| \leq (\|g_t(\theta)\|_2 + \|\bar{g}(\theta)\|_2)\,(\|\theta\|_2 + \|\theta^*\|_2) \leq 4GR \leq 2G^2,$$

where the first inequality follows from the triangle inequality and the Cauchy-Schwartz inequality, and the final inequality uses that $R \leq G/2$ by definition of $G = r_{\max} + 2R$.

To establish the second claim, consider the following inequality for any vectors (a_1, b_1, a_2, b_2):

$$\left|a_1^\top b_1 - a_2^\top b_2\right| = \left|a_1^\top (b_1 - b_2) + b_2^\top (a_1 - a_2)\right| \leq \|a_1\|\|b_1 - b_2\| + \|b_2\|\|a_1 - a_2\|.$$

This follows as a direct application of Cauchy-Schwartz. It implies that for any $\theta, \theta' \in \Theta_R$,

$$\begin{aligned}
|\zeta_t(\theta) - \zeta_t(\theta')| &= \left|(g_t(\theta) - \bar{g}(\theta))^\top (\theta - \theta^*) - (g_t(\theta') - \bar{g}(\theta'))^\top (\theta' - \theta^*)\right| \\
&\leq \|g_t(\theta) - \bar{g}(\theta)\|_2 \|\theta - \theta'\|_2 + \|\theta' - \theta^*\|_2 \|(g_t(\theta) - \bar{g}(\theta)) - (g_t(\theta') - \bar{g}(\theta'))\|_2 \\
&\leq 2G\|\theta - \theta'\|_2 + 2R\,(\|g_t(\theta) - g_t(\theta')\|_2 + \|\bar{g}(\theta) - \bar{g}(\theta')\|_2) \\
&\leq 2G\|\theta - \theta'\|_2 + 8R\|\theta - \theta'\|_2 \\
&\leq 6G\|\theta - \theta'\|_2.
\end{aligned}$$

where we used that $R \leq G/2$ by the definition of G. We also used that both $g_t(\cdot)$ and $\bar{g}(\cdot)$ are 2-Lipschitz functions which is easy to see. Starting with $g_t(\theta) = (r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta)\phi(s_t)$, consider

$$\begin{aligned}
\|g_t(\theta) - g_t(\theta')\|_2 &= \left\|\phi(s_t)\left(\gamma\phi(s_t') - \phi(s_t)\right)^\top (\theta - \theta')\right\|_2 \\
&\leq \|\phi(s_t)\|_2 \left\|\left(\gamma\phi(s_t') - \phi(s_t)\right)\right\|_2 \|(\theta - \theta')\|_2 \\
&\leq 2\|(\theta - \theta')\|_2.
\end{aligned}$$

Similarly, following Equation (2.2), we have $\|\bar{g}(\theta) - \bar{g}(\theta')\|_2 = \left\|\mathbb{E}\left[\phi\left(\gamma\phi' - \phi\right)\right]^\top(\theta - \theta')\right\|_2$, where $\phi = \phi(s)$ is the feature vector of a random initial state $s \sim \pi$, $\phi' = \phi(s')$ is the feature vector of a random next state drawn according to $s' \sim \mathcal{P}(\cdot \mid s)$. Therefore,

$$\|\bar{g}(\theta) - \bar{g}(\theta')\|_2 \leq \left\|\phi\left(\gamma\phi' - \phi\right)^\top(\theta - \theta')\right\| \leq 2\|(\theta - \theta')\|_2.$$

□

A.2 Analysis of Projected TD(λ) under Markov chain sampling model

In this section, we give a detailed proof of the convergence bounds presented in Theorem 4. Subsection A.2.1 details our proof strategy along with key lemmas which come together in Subsection A.2.2 to establish the results. We begin by providing mathematical expressions for TD(λ) updates.

Stationary distribution of TD(λ) updates: Recall that the projected TD(λ) update at time t is given by:

$$\theta_{t+1} = \Pi_{2,R}\left(\theta_t + \alpha_t x_t(\theta_t, z_{0:t})\right)$$

where $\Pi_{2,R}(\cdot)$ denotes the projection operator onto a norm ball of radius $R < \infty$ and $x_t(\theta_t, z_{0:t})$ is the update direction. Let us now give explicit mathematical expressions for $x_t(\theta, z_{0:t})$ and its steady-state mean $\bar{x}(\theta)$. Note that these are analogous to the expressions for the negative semi-gradient $g_t(\theta)$ and its steady-state expectation $\bar{g}(\theta)$ for TD(0). At time t, as a general function of (non-random) θ and the tuple $O_t = (s_t, r_t, s_t')$ along with the eligibility trace term $z_{0:t}$, we have

$$x_t(\theta, z_{0:t}) = \left(r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right) z_{0:t} = \delta_t(\theta) z_{0:t} \qquad \forall\ \theta \in \mathbb{R}^d.$$

The asymptotic convergence of TD(λ) is closely related to the expected value of $x_t(\theta, z_{0:t})$ under the steady-state behavior of $(O_t, z_{0:t})$,

$$\bar{x}(\theta) = \lim_{t\to\infty} \mathbb{E}\left[\delta_t(\theta) z_{0:t}\right].$$

Rather than take this limit, it will be helpful in our analysis to think of an equivalent *backward view* by constructing a stationary process with mean $\bar{x}(\theta)$. Consider a stationary sequence of states $(\ldots, s_{-1}, s_0, s_1, \ldots)$ and set $z_{-\infty:t} = \sum_{k=0}^{\infty}(\gamma\lambda)^k\phi(s_{t-k})$. Then the sequence $(x_0(\theta, z_{-\infty:0}), x_1(\theta, z_{-\infty:1}), \ldots)$

is stationary, and we have

$$\bar{x}(\theta) = \mathbb{E}\left[\delta_t(\theta) z_{-\infty:t}\right]. \tag{A.5}$$

It should be emphasized that $\bar{x}(\theta)$ and the states $(\ldots, s_{-2}, s_{-1})$ are introduced only for the purposes of our analysis and are never used by the algorithm itself. However, this turns out to be quite useful as it is easy to show [61] that

$$\bar{x}(\theta) = \Phi^\top D\left(T_\mu^{(\lambda)}\Phi\theta - \Phi\theta\right), \tag{A.6}$$

where Φ is the feature matrix and $(T_\mu^{(\lambda)}\Phi\theta - \Phi\theta)$ denotes the Bellman error defined with respect to the Bellman operator $T_\mu^{(\lambda)}(\cdot)$, corresponding to a policy μ. Careful readers will notice the stark similarity between Equation (A.6) and Equation (2.3). Exploiting the property that $\Pi_D T_\mu^\lambda(\cdot)$ is also a contraction operator, one can easily show a result equivalent to Lemma 3, thus quantifying the progress we make by taking steps in the direction of $\bar{x}(\theta)$. The rest of our proof essentially shows how to control for the observation noise, i.e. the fact that we use $x_t(\theta, z_{0:t})$ rather than $\bar{x}(\theta)$ to make updates. To remind the readers of the results, we first restate Theorem 4 below.

Theorem 4. *Suppose the Projected TD(λ) algorithm is applied with parameter $R \geq \|\theta^*\|_2$ under the Markov chain observation model with Assumption 1. Set $B = \frac{(r_{\max}+2R)}{(1-\gamma\lambda)}$. Then the following claims hold.*

(a) With a constant step-size $\alpha_t = \alpha_0 = 1/\sqrt{T}$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2\left(13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\kappa)}.$$

(b) With a constant step-size $\alpha_t = \alpha_0 < 1/(2\omega(1-\kappa))$ and $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_T}\right\|_D^2\right] \leq \left(e^{-2\alpha_0(1-\kappa)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \alpha_0\left(\frac{B^2\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right)}{2(1-\kappa)\omega}\right).$$

(c) With a decaying step-size $\alpha_t = 1/(\omega(t+1)(1-\kappa))$,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{B^2\left(13 + 52\tau_\lambda^{\text{mix}}(\alpha_T)\right)}{T(1-\kappa)^2\omega}\left(1 + \log T\right).$$

A.2.1 Proof strategy and key lemmas

We now describe our proof strategy and give key lemmas used to establish Theorem 4. Throughout, we consider the Markov chain observation model with Assumption 1 and study the Projected TD (λ) algorithm applied with parameter $R \geq \|\theta^*\|_2$ and step-size sequence $(\alpha_0, \ldots, \alpha_T)$. To simplify our exposition, we introduce some notation below.

Notation: We specify the notation used throughout this section. Define the set $\Theta_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$, so $\theta_t \in \Theta_R$ for each t because of the algorithm's projection step. Next, we generically define $z_{l:t} = \sum_{k=0}^{t-l}(\gamma\lambda)^k\phi(s_{t-k})$ for any lower limit $l \leq t$. Thus, $z_{l:t}$ denotes the eligibility trace as a function of the states $(s_l, \ldots, s_t)$. Next, we define $\zeta_t(\theta, z_{l:t})$ as a general function of θ and $z_{l:t}$,

$$\zeta_t(\theta, z_{l:t}) = \left(\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\right)^\top(\theta - \theta^*). \tag{A.7}$$

Here, the subscript t in ζ_t encodes the dependence on the tuple $O_t = (s_t, r_t, s_t')$ which is used to compute the Bellman error, $\delta_t(\cdot)$ at time t. Finally, we set $B := (r_{\max} + 2R)/(1-\gamma\lambda)$ which implies $B > 2R$, a fact we use many times in our proofs to simplify constant terms. As a reminder, note that our bounds depend on the mixing time, which we defined in Section 2.10 as

$$\tau_\lambda^{\text{mix}}(\epsilon) = \max\{\tau^{\text{MC}}(\epsilon), \tau^{\text{Algo}}(\epsilon)\},$$

where $\tau^{\text{MC}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \epsilon\}$ and $\tau^{\text{Algo}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid (\gamma\lambda)^t \leq \epsilon\}$.

Proof outline: The analysis for TD(λ) can be broadly divided into three parts and closely mimics the steps used to prove TD(0) results.

1. As a first step, we do an error decomposition, similar to the result shown in Lemma 8. This is enabled by two key lemmas, which are analogues of Lemma 3 and Lemma 6 for Projected TD(0). The first one spells out a clear relationship of how the updates following $\bar{x}(\theta)$ point in the descent direction of $\|\theta^* - \theta\|_2^2$ while the second one upper bounds the norm of the update direction, $x_t(\theta, z_{0:t})$, by the constant B (as defined above).

2. The error decomposition that we obtain from Step 1 can be stated as:

$$\mathbb{E}[\|\theta^* - \theta_{t+1}\|_2^2] \leq \mathbb{E}[\|\theta^* - \theta_t\|_2^2] - 2\alpha_t(1-\kappa)\mathbb{E}[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2] + 2\alpha_t\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] + \alpha_t^2 B^2.$$

   In the second step, we establish an upper bound on the bias term, $\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]$, which is the main challenge in our proof. Recall that the dependent nature of the state transitions may result in strong coupling between the tuples O_{t-1} and O_t under the Markov chain observation model. Therefore, this bias in update direction can potentially be non-zero. Presence of the eligibility trace term, $z_{0:t}$, which is a function of the entire history of states, $(s_0, \ldots, s_t)$, further complicates the analysis by introducing subtle dependencies.

   To control for this, we use information-theoretic techniques shown in Lemma 9 which exploit the geometric ergodicity of the MDP, along with the geometric weighting of state features in the eligibility trace term. Our result essentially shows that the bias scales the noise in update direction by a factor of the mixing time. Mathematically, for a constant step-size α, we show that $\mathbb{E}\left[\alpha\zeta_t(\theta_t, z_{0:t})\right] \approx B^2(6+12\tau_\lambda^{\text{mix}}(\alpha))\alpha^2$. We show a similar result for decaying step-sizes as well.

3. In the final step, we combine the error decomposition from Step 1 and the bound on the bias from Step 2, to establish finite time bounds on the performance of Projected TD(λ) for different step-size choices. We closely mimic the analysis of [53] for a constant, aggressive step-size of $(1/\sqrt{T})$ and the proof ideas of [56] for decaying step-sizes.

Error decomposition under Projected TD(λ)

We first state two important lemmas below which enable the error decomposition shown in Lemma 34.

Lemma 32. ***[Tsitsiklis and Van Roy [9]]*** *Let V_{θ^*} be the unique fixed point of $\Pi_D T_\mu^{(\lambda)}(\cdot)$ i.e. $V_{\theta^*} = \Pi T_\mu^{(\lambda)} V_{\theta^*}$. For any $\theta \in \mathbb{R}^d$,*

$$(\theta^* - \theta)^\top \bar{x}(\theta) \geq (1 - \kappa)\|V_{\theta^*} - V_\theta\|_D^2.$$

Proof. We use the definition of $\bar{x}(\theta) = \langle \Phi^\top, T_\mu^{(\lambda)}\Phi\theta - \Phi\theta\rangle_D$ as shown in Equation (A.6) along with the fact that $\Pi_D T_\mu^{(\lambda)}(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus κ. See Appendix A.3 for a complete proof. □

Lemma 33. *For all $\theta \in \Theta_R$, $\|x_t(\theta, z_{0:t})\|_2 \leq B$ with probability 1. Additionally, $\|\bar{x}(\theta)\|_2 \leq B$.*

Proof. See Subsection A.2.3 for a complete proof. □

The above two lemmas can be easily combined to establish a recursion for the error under projected TD(λ) that holds for each sample path.

Lemma 34. *With probability 1, for every $t \in \mathbb{N}_0$,*

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1 - \kappa)\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2 + 2\alpha_t\zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 B^2.$$

Proof. The Projected TD(λ) algorithm updates the parameter as: $\theta_{t+1} = \Pi_{2,R}[\theta_t + \alpha_t x_t(\theta_t, z_{0:t})] \;\; \forall \; t \in$

$\mathbb{N}_0$. This implies,

$$\begin{aligned}
\|\theta^* - \theta_{t+1}\|_2^2 &= \left\|\theta^* - \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta, z_{0:t}))\right\|_2^2 \\
&= \left\|\Pi_{2,R}(\theta^*) - \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))\right\|_2^2 \\
&\leq \|\theta^* - \theta_t - \alpha_t x_t(\theta_t, z_{0:t})\|_2^2 \\
&= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t x_t(\theta_t, z_{0:t})^\top(\theta^* - \theta_t) + \alpha_t^2\|x_t(\theta_t, z_{0:t})\|_2^2 \\
&\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t x_t(\theta_t, z_{0:t})^\top(\theta^* - \theta_t) + \alpha_t^2 B^2 \\
&= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t \bar{x}(\theta_t)^\top(\theta^* - \theta_t) + 2\alpha_t\zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 G^2. \\
&\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\kappa)\|V_{\theta^*} - V_\theta\|_D^2 + 2\alpha_t\zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 B^2.
\end{aligned}$$

The first inequality used that orthogonal projection operators onto a convex set are non-expansive, the second used Lemma 33 together with the fact $\|\theta_t\|_2 \leq R$ due to projection, and the third used Lemma 32. Note that we used $\zeta_t(\theta_t, z_{0:t})$ to simplify the notation for the error in the update direction. Recall the definition of the error function from Equation (A.7) which implies,

$$\zeta_t(\theta_t, z_{0:t}) = (\delta_t(\theta_t)z_{0:t} - \bar{x}(\theta_t))^\top(\theta_t - \theta^*) = (x_t(\theta_t, z_{0:t}) - \bar{x}(\theta_t))^\top(\theta_t - \theta^*).$$

□

Upper bound on the bias in update direction.

We give an upper bound on the expected error in the update direction, $\mathbb{E}[\zeta_t(\theta_t, z_{0:t})]$, which as explained above, is the key challenge for our analysis. For this, we first establish some basic regularity properties of the error function $\zeta_t(\cdot,\cdot)$ in Lemma 35 below. In particular, part (a) shows boundedness, part (b) shows that it is Lipschitz in the first argument and part (c) bounds the error due to truncation of the eligibility trace. Recall that $z_{l:t}$ denotes the eligibility trace as a function of the states $(s_l, \ldots, s_t)$.

Lemma 35. *Consider any $l \leq t$ and any $\theta, \theta' \in \Theta_R$. With probability 1,*

(a) $|\zeta_t(\theta, z_{l:t})| \leq 2B^2$.

(b) $|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| \leq 6B\|(\theta - \theta')\|_2$.

(c) The following two bounds also hold,

$$|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| \leq B^2(\gamma\lambda)^\tau \text{ for all } \tau \leq t,$$
$$|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| \leq B^2(\gamma\lambda)^t.$$

Proof. We essentially use the uniform bound on $x_t(\theta, z_{0:t})$ and $\bar{x}(\theta)$ as stated in Lemma 33 to show this result. See Subsection A.2.3 for a detailed proof. □

Lemma 35 can be combined with Lemma 9 to give an upper bound on the bias term, $\mathbb{E}[\zeta_t(\theta_t, z_{0:t})]$, as shown below.

Lemma 36. *Consider a non-increasing step-size sequence, $\alpha_0 \geq \alpha_1 \ldots \geq \alpha_T$. Then the following hold.*

(a) For $2\tau_\lambda^{\text{mix}}(\alpha_T) < t \leq T$,

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T))\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)}.$$

(b) For $0 \leq t \leq 2\tau_\lambda^{\text{mix}}(\alpha_T)$,

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2\left(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T)\right)\alpha_0 + B^2(\gamma\lambda)^t.$$

(c) For all $t \in \mathbb{N}_0$,

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t$$

Proof. We proceed in two cases below. Throughout the proof, results from Lemma 35 are applied using the fact that $\theta_t \in \Theta_R$, because of the algorithm's projection step.

Case (a): Let $t > 2\tau$ and consider the following decomposition for all $\tau \in \{0, 1, \ldots, t/2\}$. We show an upper bound on each of the three terms separately.

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq \left|\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})]\right| + \left|\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]\right| + \left|\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]\right|.$$

Step 1: Use regularity properties of the error function to bound first two terms.

We relate $\zeta_t(\theta_t, z_{0:t})$ and $\zeta_t(\theta_{t-\tau}, z_{0:t})$ using the Lipschitz property shown in part (b) of Lemma 35 to get,

$$\left|\zeta_t(\theta_t, z_{0:t}) - \zeta_t(\theta_{t-2\tau}, z_{0:t})\right| = 6B\|\theta_t - \theta_{t-2\tau}\|_2 \leq 6B^2 \sum_{i=t-2\tau}^{t-1} \alpha_i. \tag{A.8}$$

Taking expectations on both sides gives us the desired bound on the first term. The last inequality used the norm bound on update direction as shown in Lemma 33 to simplify,

$$\|\theta_t - \theta_{t-2\tau}\|_2 \leq \sum_{i=t-2\tau}^{t-1} \left\|\Pi_{2,R}\left(\theta_{i+1} + \alpha_i x_i(\theta_i, z_{0:i})\right) - \theta_i\right\|_2 \leq \sum_{i=t-2\tau}^{t-1} \alpha_i \|x_i(\theta_i, z_{0:i})\|_2 \leq B \sum_{i=t-2\tau}^{t-1} \alpha_i.$$

Similarly, by part (c) of Lemma 35, we have a bound on the second term.

$$\left|\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]\right| \leq B^2(\gamma\lambda)^\tau. \tag{A.9}$$

Step 2: Use information-theoretic arguments to upper bound $\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]$.

We will essentially use Lemma 9 to upper bound $\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]$. We first introduce some notation to highlight subtle dependency issues. Note that $\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})$ is a function of $(\theta_{t-2\tau}, s_{t-\tau}, \ldots, s_{t-1}, O_t)$.

To simplify, let $Y_{t-\tau:t} = (s_{t-\tau}, \ldots, s_{t-1}, O_t)$. Define,

$$f(\theta_{t-2\tau}, Y_{t-\tau:t}) := \zeta_t(\theta_{t-2\tau}, z_{t-\tau:t}).$$

Consider random variables $\theta'_{t-2\tau}$ and $Y'_{t-\tau:t}$ drawn independently from the marginal distributions of $\theta_{t-2\tau}$ and $Y_{t-\tau:t}$, so $\mathbb{P}(\theta'_{t-2\tau} = \cdot, Y'_{t-\tau:t} = \cdot) = \mathbb{P}(\theta_{t-\tau} = \cdot) \otimes \mathbb{P}(Y_{t-\tau:t} = \cdot)$. By Lemma 35 we have that $|f(\theta, Y_{t-\tau:t})| \leq 2B^2$ for all $\theta \in \Theta_R$ with probability 1. As

$$\theta_{t-2\tau} \to s_{t-2\tau} \to s_{t-\tau} \to s_t \to O_t$$

form a Markov chain, a direct application of Lemma 9 gives us:

$$\left|\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})] - \mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]\right| \leq 4B^2 m\rho^\tau. \tag{A.10}$$

We also have the following bound for all fixed $\theta \in \Theta_R$. Using $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta) z_{-\infty:t}]$, we get

$$\mathbb{E}[f(\theta, Y_{t-\tau:t})] = \left(\mathbb{E}[\delta_t(\theta) z_{t-\tau:t}] - \bar{x}(\theta)\right)^\top (\theta - \theta^*) \leq \left|\left(\delta_t(\theta) z_{-\infty:t-\tau}\right)^\top (\theta - \theta^*)\right| \leq B^2(\gamma\lambda)^\tau \tag{A.11}$$

Combining the above with Equation (A.10), we get

$$\begin{aligned}
|\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]| &= |\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})]| \\
&\leq \left|\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})] - \mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]\right| + \left|\mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]\right| \\
&\leq 4B^2 m\rho^\tau + \left|\mathbb{E}[\mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})|\theta'_{t-2\tau}]]\right| \\
&\leq 4B^2 m\rho^\tau + B^2(\gamma\lambda)^\tau.
\end{aligned} \tag{A.12}$$

Step 3. Combine terms to show part (a) of our claim.

Taking $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$ and combining Equations (A.8), (A.9) and (A.12) establishes the first claim.

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2 \sum_{i=t-2\tau}^{t-1} \alpha_i + 4B^2 m\rho^\tau + 2B^2(\gamma\lambda)^\tau \leq 12B^2\tau_\lambda^{\text{mix}}(\alpha_T)\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)} + 6B^2\alpha_T$$
$$\leq 6B^2(1 + 2\tau^{\text{mix}}(\alpha_T))\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)}.$$

Here we used that letting $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$ implies: $\max\{m\rho^\tau, (\gamma\lambda)^\tau\} \leq \alpha_T$. Two additional facts which we also use follow from a non-increasing step-size sequence, $\sum_{i=t-2\tau}^{t-1} \alpha_i \leq 2\tau\alpha_{t-2\tau}$ and $\alpha_T \leq \alpha_{t-2\tau}$.

Case (b): Consider the following decomposition for all $t \in \mathbb{N}_0$,

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq |\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] - \mathbb{E}[\zeta_t(\theta_0, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]|.$$

Step 1: Use regularity properties of the error function to upper bound the first two terms.

Using parts (b), (c) of Lemma 35 and following the arguments shown in Step 1, 2 of case (a) above, we get

$$|\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] - \mathbb{E}[\zeta_t(\theta_0, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]| \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t. \quad \text{(A.13)}$$

Step 2: Characterizing $\mathbb{E}[\zeta_t(\theta, z_{-\infty:t})]$ for any fixed (non-random) θ.

Recall the definition of $\bar{x}(\theta)$ from Equation (A.5). For any fixed θ, we have $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta) z_{-\infty:t}]$. Therefore,

$$\mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})] = \left(\mathbb{E}[\delta_t(\theta_0) z_{-\infty:t}] - \bar{x}(\theta_0)\right)^\top (\theta_0 - \theta^*) = 0. \quad \text{(A.14)}$$

Step 3. Combine terms to show parts (b), (c) of our claim.

Combining Equations (A.13) and (A.14) establishes part (c) which states,

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t \quad \forall\ t \in \mathbb{N}_0.$$

We establish part (b) by using that the step-size sequence is non-increasing which implies: $\sum_{i=0}^{t-1} \alpha_i \leq t\alpha_0$. For all $t \leq 2\tau_\lambda^{\text{mix}}(\alpha_T)$, we have the following loose upper bound.

$$\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \ \leq\ 6B^2 t\alpha_0 + B^2(\gamma\lambda)^t \ \leq\ 6B^2\left(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T))\right)\alpha_0 + B^2(\gamma\lambda)^t.$$

□

A.2.2 Proof of Theorem 4

In this subsection, we establish convergence bounds for Projected TD(λ) as stated in Theorem 4 using Lemmas 34 and 36. From Lemma 34 we have,

$$\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] \leq \mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - 2\alpha_t(1-\kappa)\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] + 2\alpha_t\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] + \alpha_t^2 B^2 \quad \text{(A.15)}$$

Equation (A.15) will be used as a starting point for analyzing different step-size choices.

Proof of part (a): Fix a constant step-size of $\alpha_0 = \ldots = \alpha_t = 1/\sqrt{T}$ in Equation (A.15), rearrange terms and sum from $t = 0$ to $t = T - 1$, we get

$$2\alpha_0(1-\kappa)\sum_{t=0}^{T-1}\mathbb{E}\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2 \leq \sum_{t=0}^{T-1}\left(\mathbb{E}\|\theta^* - \theta_t\|_2^2 - \mathbb{E}\|\theta^* - \theta_{t+1}\|_2^2\right) + B^2 + 2\alpha_0\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right].$$

Using Lemma 36 where $\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)} = \alpha_0$ along with the fact that $(\gamma\lambda) < 1$, we simplify to get

$$\sum_{t=0}^{T-1} \mathbb{E}\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2 \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2}{2\alpha_0(1-\kappa)} + \frac{6B^2T(1+2\tau_\lambda^{\text{mix}}(1/\sqrt{T}))\alpha_0}{(1-\kappa)} + \frac{1}{(1-\kappa)} \sum_{t=0}^{2\tau_\lambda^{\text{mix}}(\frac{1}{\sqrt{T}})} B^2(\gamma\lambda)^t$$
$$\leq \frac{\sqrt{T}\left(\|\theta^* - \theta_0\|_2^2 + B^2\right)}{2(1-\kappa)} + \frac{6B^2\sqrt{T}(1+2\tau_\lambda^{\text{mix}}(1/\sqrt{T}))}{(1-\kappa)} + \frac{2B^2\tau_\lambda^{\text{mix}}(1/\sqrt{T})}{(1-\kappa)}.$$

Adding these terms, we conclude

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2\left(13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T})\right)}{2\sqrt{T}(1-\kappa)}.$$

Proof of part (b): For a constant step-size of $\alpha_0 < 1/(2\omega(1-\kappa))$, we show that the expected distance between the iterate θ_T and the TD(λ) limit point, θ^* converges at an exponential rate below some level that depends on the choice of step-size and λ. Starting with Equation (A.15) and applying Lemma 1 which shows that $\|V_{\theta^*} - V_\theta\|_D^2 \geq w\|\theta^* - \theta\|_2^2$ for any θ, we have that for all $t > 2\tau_\lambda^{\text{mix}}(\alpha_0)$,

$$\mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] \leq (1 - 2\alpha_0(1-\kappa)\omega)\,\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + \alpha_0^2B^2 + 2\alpha_0\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]$$
$$\leq (1 - 2\alpha_0(1-\kappa)\omega)\,\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] + \alpha_0^2B^2\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right),$$

where we used part (a) of Lemma 36 for the second inequality. Iterating over it gives us our final result. For any $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$,

$$\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] \leq (1 - 2\alpha_0(1-\kappa)\omega)^T\,\|\theta^* - \theta_0\|_2^2 + \alpha_0^2B^2\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right)\sum_{t=0}^{\infty}(1 - 2\alpha_0(1-\kappa)\omega)^t$$
$$\leq \left(e^{-2\alpha_0(1-\kappa)\omega T}\right)\|\theta^* - \theta_0\|_2^2 + \frac{B^2\alpha_0\left(13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)\right)}{2(1-\kappa)\omega}.$$

Final inequality follows by solving the geometric series and using that $(1 - 2\alpha_0(1-\kappa)\omega) \leq e^{-2\alpha_0(1-\kappa)\omega}$ along with Lemma 1.

Proof of part (c): Consider a decaying step-size of $\alpha_t = 1/(\omega(t+1)(1-\kappa))$. We start with Equation (A.15) and use Lemma 1 which showed $\mathbb{E}\left[\|V_{\theta^*} - V_\theta\|_D^2\right] \geq w\mathbb{E}\left[\|\theta^* - \theta\|_2^2\right]$ for all θ to get,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] \leq \frac{1}{(1-\kappa)\alpha_t}\left((1-(1-\kappa)\omega\alpha_t)\mathbb{E}\left[\|\theta^* - \theta_t\|_2^2\right] - \mathbb{E}\left[\|\theta^* - \theta_{t+1}\|_2^2\right] + \alpha_t^2 B^2\right) + \frac{2}{(1-\kappa)}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right].$$

Substituting $\alpha_t = \frac{1}{\omega(t+1)(1-\kappa)}$, simplify and sum from $t = 0$ to $T-1$ to get,

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_t}\right\|_D^2\right] &\leq -\omega T\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right] + \frac{B^2}{\omega(1-\kappa)^2}\sum_{t=0}^{T-1}\frac{1}{t+1} + \frac{2}{(1-\kappa)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \\ &\leq \underbrace{-\omega T\mathbb{E}\left[\|\theta^* - \theta_T\|_2^2\right]}_{<0} + \frac{B^2(1+\log T)}{\omega(1-\kappa)^2} + \frac{2}{(1-\kappa)}\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right],\end{aligned} \tag{A.16}$$

where we use that $\sum_{t=0}^{T-1}\frac{1}{t+1} \leq (1+\log T)$. To simplify notation, we put $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$ for the remainder of the proof. We use Lemma 36 to upper bound the total bias, $\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]$ which can be decomposed as:

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] = \sum_{t=0}^{2\tau}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] + \sum_{t=2\tau+1}^{T-1}\mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right]. \tag{A.17}$$

First, note that for a decaying step-size $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$ we have

$$\sum_{t=0}^{T-1}\alpha_t = \frac{1}{\omega(1-\gamma)}\sum_{t=0}^{T-1}\frac{1}{(t+1)} \leq \frac{1+\log T}{\omega(1-\gamma)}.$$

We will combine this with Lemma 36 to upper bound each term separately. First,

$$\sum_{t=0}^{2\tau} \mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq \sum_{t=0}^{2\tau} \left(6B^2 \sum_{i=0}^{t-1} \alpha_i\right) + \sum_{t=0}^{2\tau} B^2(\gamma\lambda)^t$$
$$\leq \frac{6B^2}{\omega(1-\kappa)} \sum_{t=0}^{2\tau} \sum_{i=0}^{T-1} \frac{1}{(i+1)} + 2B^2\tau \;\leq\; \frac{14B^2\tau}{\omega(1-\kappa)}(1+\log T),$$

where we used the fact that $\omega, \kappa, (\gamma\lambda) < 1$. Similarly,

$$\sum_{t=2\tau+1}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq 6B^2(1+2\tau) \sum_{t=2\tau+1}^{T-1} \alpha_{t-2\tau} \leq 6B^2(1+2\tau) \sum_{t=0}^{T-1} \alpha_t \leq \frac{6B^2(1+2\tau)}{\omega(1-\kappa)} \left(1+\log T\right).$$

Combining the two parts, we get

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right] \leq \frac{B^2(6+26\tau)}{\omega(1-\kappa)}(1+\log T).$$

Using this in conjunction with Equation (A.16) we get,

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|V_{\theta_t} - V_{\theta^*}\right\|_D^2\right] \leq \frac{B^2\left(1+\log T\right)}{\omega T(1-\kappa)^2} + \frac{2}{T(1-\kappa)} \sum_{t=0}^{T-1} \mathbb{E}\left[\zeta_t(\theta_t, z_{0:t})\right].$$

Simplifying and putting back $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$, we get our final result.

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2\right] \leq \frac{B^2}{\omega T(1-\kappa)^2}\left(1+\log T\right) + \frac{2B^2\left(6+26\tau_\lambda^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\kappa)^2}(1+\log T)$$
$$\leq \frac{B^2\left(13+52\tau^{\text{mix}}(\alpha_T)\right)}{\omega T(1-\kappa)^2}(1+\log T).$$

We remark that Equation (A.16) implies a convergence rate of $\tilde{O}(\log T/T)$ for the iterate θ_T (and hence the value function V_{θ_T}) itself but the bounds degrade by a factor of ω. In particular, we have

$$\mathbb{E}\left[\left\|V_{\theta^*} - V_{\theta_T}\right\|_D^2\right] \leq \mathbb{E}\left[\left\|\theta^* - \theta_T\right\|_2^2\right] \leq \frac{B^2\left(13+52\tau_\lambda^{\text{mix}}(\alpha_T)\right)}{\omega^2 T(1-\kappa)^2}\left(1+\log T\right) \tag{A.18}$$

where the first inequality follows using Lemma 1.

A.2.3 Proof of supporting lemmas.

In this subsection, we provide standalone proofs of Lemma 33 and 35 used above.

Lemma 33. *For all $\theta \in \Theta_R$, $\|x_t(\theta, z_{0:t})\|_2 \leq B$ with probability 1. Additionally, $\|\bar{x}(\theta)\|_2 \leq B$.*

Proof. We start with the mathematical expression for $x_t(\theta, z_{0:t})$.

$$x_t(\theta, z_{0:t}) = \delta_t(\theta) z_{0:t} \implies \|x_t(\theta, z_{0:t})\|_2 = |\delta_t(\theta)| \|z_{0:t}\|_2.$$

We give an upper bound on both $|\delta_t(\theta)|$ and $\|z_{0:t}\|_2$. Starting with the definition of $\delta_t(\theta)$ and using that $\|\phi(s_t)\|_2 \leq 1 \ \forall \ t$ along with $\|\theta\|_2 \leq R$, we get

$$|\delta_t(\theta)| = \left|r_t + \gamma \phi(s_t')^\top \theta - \phi(s_t)^\top \theta\right| \leq r_{\max} + \left\|\phi(s_t')\right\|_2 \|\theta\|_2 + \|\phi(s_t)\|_2 \|\theta\|_2 \leq (r_{\max} + 2R) \, .$$

Next,

$$\|z_{0:t}\|_2^2 = \left\| \sum_{k=0}^{t} (\gamma\lambda)^k \phi(s_{t-k}) \right\|_2^2 \leq \left(\sum_{k=0}^{t} (\gamma\lambda)^k \right)^2 \leq \left(\sum_{k=0}^{\infty} (\gamma\lambda)^k \right)^2 = \frac{1}{(1-\gamma\lambda)^2}.$$

Combining these two implies the first part of our claim.

$$\|x_t(\theta, z_{0:t})\|_2 = |\delta_t(\theta)| \|z_{0:t}\|_2 \leq \frac{(r_{\max} + 2R)}{(1-\gamma\lambda)} = B.$$

Note that can easily show an upper bound $\|\delta_t(\theta) z_{l:t}\|_2 \leq B$ for any pair $(\theta, z_{l:t})$ with $l \leq t$. Consider,

$$\|z_{l:t}\|_2^2 \leq \|z_{-\infty:t}\|_2^2 \leq \left(\sum_{k=0}^{\infty} (\gamma\lambda)^k \right)^2 = \frac{1}{(1-\gamma\lambda)^2}$$
$$\implies \|\delta_t(\theta) z_{l:t}\|_2 = |\delta_t(\theta)| \|z_{l:t}\|_2 \leq \frac{(r_{\max} + 2R)}{(1-\gamma\lambda)} = B.$$

Taking $l \to -\infty$ implies that $\|\delta_t(\theta) z_{-\infty:t}\|_2 \leq B$. As $\bar{x}(\theta) = \mathbb{E}\left[\delta_t(\theta) z_{-\infty:t}\right]$, we also have a uniform

norm bound on the expected updates, $\|\bar{x}(\theta)\|_2 \leq B$, as claimed. □

Lemma 35. *Consider any $l \leq t$ and any $\theta, \theta' \in \Theta_R$. With probability 1,*

(a) $|\zeta_t(\theta, z_{l:t})| \leq 2B^2$.

(b) $|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| \leq 6B\|(\theta - \theta')\|_2$.

(c) The following two bounds also hold,

$$|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| \leq B^2(\gamma\lambda)^\tau \text{ for all } \tau \leq t,$$
$$|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| \leq B^2(\gamma\lambda)^t.$$

Proof. Throughout, we use the assumption that basis vectors are normalized i.e. $\|\phi(s_t)\|_2 \leq 1 \ \forall \ t$.

Part (a): We show a uniform norm bound on $\zeta_t(\theta, z_{l:t}) \ \forall \ \theta \in \Theta_R$. First consider the following:

$$\begin{aligned}
\|\delta_t(\theta)z_{l:t}\|_2 = |\delta_t(\theta)|\|z_{l:t}\|_2 &\leq \left|r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right|\left\|\sum_{k=0}^{t-l}(\gamma\lambda)^k\phi(s_{t-k})\right\|_2 \\
&\leq \left|r_t + \left\|\phi(s_t')\right\|_2\|\theta\|_2 + \|\phi(s_t)\|_2\|\theta\|_2\right|\left\|\sum_{k=0}^{\infty}(\gamma\lambda)^k\phi(s_{t-k})\right\|_2 \\
&\leq \frac{(r_{\max} + 2R)}{(1-\gamma\lambda)} = B.
\end{aligned}$$

Using this along with the fact that $\|\theta - \theta^*\|_2 \leq 2R \leq B$ and $\|\bar{x}(\theta)\|_2 \leq B$ for all $\theta \in \Theta_R$, we get

$$\begin{aligned}
|\zeta_t(\theta, z_{l:t})| = \left|(\delta_t(\theta)z_{l:t} - \bar{x}(\theta))^\top(\theta - \theta^*)\right| &\leq \|\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\|_2\|(\theta - \theta^*)\|_2 \\
&\leq (\|\delta_t(\theta)z_{l:t}\|_2 + \|\bar{x}(\theta)\|_2)\,\|(\theta - \theta^*)\|_2 \\
&\leq 2B\|(\theta - \theta^*)\|_2 \leq 2B^2.
\end{aligned}$$

Part (b): To show that $\zeta_t(\cdot, z_{l:t})$ is L-Lipschitz, consider the following inequality for any four vectors (a_1, b_1, a_2, b_2), which follows as a direct application of Cauchy-Schwartz.

$$\left|a_1^\top b_1 - a_2^\top b_2\right| = \left|a_1^\top (b_1 - b_2) + b_2^\top (a_1 - a_2)\right| \leq \|a_1\|_2 \|b_1 - b_2\|_2 + \|b_2\|_2 \|a_1 - a_2\|_2.$$

This implies,

$$\begin{aligned}
|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| &= \left|(\delta_t(\theta) z_{l:t} - \bar{x}(\theta))^\top (\theta - \theta^*) - (\delta_t(\theta') z_{l:t} - \bar{x}(\theta'))^\top (\theta' - \theta^*)\right| \\
&\leq \|\delta_t(\theta) z_{l:t} - \bar{x}(\theta)\|_2 \|\theta - \theta'\|_2 + \|\theta' - \theta^*\|_2 \|(\delta_t(\theta) z_{l:t} - \bar{x}(\theta)) - (\delta_t(\theta') z_{l:t} - \bar{x}(\theta'))\|_2 \\
&\leq 2B\|\theta - \theta'\|_2 + 2R\left[\|z_{l:t}\,(\delta_t(\theta) - \delta_t(\theta'))\|_2 + \|\bar{x}(\theta) - \bar{x}(\theta')\|_2\right] \\
&\leq 2B\|\theta - \theta'\|_2 + \frac{8R}{(1 - \gamma\lambda)}\|\theta - \theta'\|_2 \\
&\leq 6B\|\theta - \theta'\|_2,
\end{aligned}$$

where the last inequality follows as $\frac{R}{1-\gamma\lambda} \leq B/2$ by definition. In the penultimate inequality, we used that $\|z_{l:t}\,(\delta_t(\theta) - \delta_t(\theta'))\|_2 \leq \frac{2}{(1-\gamma\lambda)}\|\theta - \theta'\|_2$ which is easy to prove. Consider,

$$\begin{aligned}
\|z_{l:t}\,(\delta_t(\theta) - \delta_t(\theta'))\|_2 &\leq \|z_{l:t}\|_2 |(\delta_t(\theta) - \delta_t(\theta'))| \\
&\leq \left\|\sum_{k=0}^{\infty} (\gamma\lambda)^k \phi(s_{t-k})\right\|_2 |(\delta_t(\theta) - \delta_t(\theta'))| \\
&\leq \frac{1}{(1 - \gamma\lambda)}\left|\left(\gamma\phi(s_t') - \phi(s_t)\right)^\top (\theta - \theta')\right| \\
&\leq \frac{\left(\left\|\phi(s_t')\right\|_2 + \|\phi(s_t)\|_2\right)}{(1 - \gamma\lambda)}\|\theta - \theta'\|_2 \leq \frac{2}{(1 - \gamma\lambda)}\|\theta - \theta'\|_2.
\end{aligned}$$

As $\bar{x}(\theta) = \mathbb{E}\left[\delta_t(\theta) z_{-\infty:t}\right]$, this also implies $\|\bar{x}(\theta) - \bar{x}(\theta')\|_2 \leq \frac{2}{(1-\gamma\lambda)}\|\theta - \theta'\|_2$ which completes the proof.

Part (c): To show that $|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| \leq B^2(\gamma\lambda)^\tau$ for all $\theta \in \Theta_R$ and $\tau \leq t$, we use that $\|\theta - \theta^*\|_2 \leq 2R \leq B$.

$$
\begin{aligned}
|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| &= \left|(\delta_t(\theta) z_{0:t} - \delta_t(\theta) z_{t-\tau:t})^\top (\theta - \theta^*)\right| \\
&\leq |\delta_t(\theta)| \|z_{0:t} - z_{t-\tau:t}\|_2 \|\theta - \theta^*\|_2 \\
&\leq \left|r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right| \left\|\sum_{k=\tau}^{\infty} (\gamma\lambda)^k \phi(s_{t-k})\right\|_2 B \\
&\leq |r_t + 2\|\theta\|_2| \cdot \frac{(\gamma\lambda)^\tau}{(1-\gamma\lambda)} \cdot B \\
&\leq B\frac{(r_{\max} + 2R)}{(1-\gamma\lambda)}(\gamma\lambda)^\tau \\
&= B^2(\gamma\lambda)^\tau.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| &\leq \left|\delta_t(\theta)(z_{0:t} - z_{-\infty:t})^\top(\theta - \theta^*)\right| \\
&\leq |\delta_t(\theta)| \|z_{0:t} - z_{-\infty:t}\|_2 \|\theta - \theta^*\|_2 \\
&\leq \left|\left(r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\right)\right| \left\|\sum_{k=t}^{\infty} (\gamma\lambda)^k \phi(s_{t-k})\right\|_2 B \\
&\leq B\frac{(r_{\max} + 2R)}{(1-\gamma\lambda)}(\gamma\lambda)^t \\
&\leq B^2(\gamma\lambda)^t.
\end{aligned}
$$

□

A.3 Proofs of Additional Lemmas

In this section, prove some additional lemmas stated and used in other parts of the paper. We first give a proof of Lemma 7 which gives an upper bound on the projection radius, R.

Lemma 7. *$\|\theta^*\|_\Sigma \leq \frac{2r_{\max}}{(1-\gamma)^{3/2}}$ and hence $\|\theta^*\|_2 \leq \frac{2r_{\max}}{\sqrt{\omega}(1-\gamma)^{3/2}}$.*

Proof. Because rewards are uniformly bounded, $|V_\mu(s)| \leq r_{\max}/(1-\gamma)$ for all $s \in \mathcal{S}$. Recall that V_μ denotes the true value function of the Markov reward process. This implies that

$$\left\|V_\mu\right\|_D \leq \left\|V_\mu\right\|_\infty \leq \frac{r_{\max}}{(1-\gamma)}.$$

Lemma 2 along with simple matrix inequalities enable a simple upper bound on $\|\theta^*\|_2$. We have

$$\left\|V_{\theta^*} - V_\mu\right\|_D \leq \frac{1}{\sqrt{1-\gamma^2}}\left\|V_\mu - \Pi_D V_\mu\right\|_D \leq \frac{1}{\sqrt{1-\gamma^2}}\left\|V_\mu\right\|_D \leq \frac{1}{\sqrt{1-\gamma}}\left\|V_\mu\right\|_D,$$

where the penultimate inequality holds by the Pythagorean theorem. By the reverse triangle inequality we have $\left|\|V_{\theta^*}\|_D - \left\|V_\mu\right\|_D\right| \leq \left\|V_{\theta^*} - V_\mu\right\|_D$. Thus,

$$\|V_{\theta^*}\|_D \leq \left\|V_{\theta^*} - V_\mu\right\|_D + \left\|V_\mu\right\|_D \leq \frac{2}{\sqrt{1-\gamma}}\left\|V_\mu\right\|_D \leq \frac{2}{\sqrt{1-\gamma}}\frac{r_{\max}}{(1-\gamma)}.$$

Recall from Section 2.3 we have, $\|V_{\theta^*}\|_D = \|\theta^*\|_\Sigma$ which establishes first part of the claim. The second claim uses that $\|\theta^*\|_\Sigma \geq \omega\|\theta^*\|_2$ which follows by Lemma 1. □

Next, we give a combined proof of Lemmas 13 and 32 which quantify the progress of the expected updates towards the limit point θ^* for TD(λ) and the Q-function approximation algorithm. These lemmas can be restated more generally as shown below, instead of using the Bellman operators $F(\cdot)$ and $T^{(\lambda)}(\cdot)$.

Lemma 37. *Consider a linear function approximation such that $J_\theta = \Phi\theta$. Let $\Pi_D H(\cdot)$ be a contraction with respect to $\|\cdot\|_D$ with modulus γ and let J_{θ^*} be the unique fixed point of $\Pi_D H(\cdot)$,*

i.e. $J_{\theta^*} = \Pi_D H J_{\theta^*}$. *Define* $\bar{g}(\theta) = \Phi^\top D\,(H\Phi\theta - \Phi\theta)$ *for all* $\theta \in \mathbb{R}^d$ *to be the expected update. Then,*

$$\bar{g}(\theta)^\top(\theta^* - \theta) \geq (1-\gamma)\|J_{\theta^*} - J_\theta\|_D^2.$$

Proof. We have

$$\begin{aligned}
(\theta^* - \theta)^\top \bar{g}(\theta) &= (\theta^* - \theta)^\top \Phi^\top D\,(H\Phi\theta - \Phi\theta) \\
&= \langle \Phi(\theta^* - \theta), (H\Phi\theta - \Phi\theta)\rangle_D \\
&= \langle \Pi_D \Phi(\theta^* - \theta), (H\Phi\theta - \Phi\theta)\rangle_D && \text{(A.19)} \\
&= \langle \Phi(\theta^* - \theta), \Pi_D\,(H\Phi\theta - \Phi\theta)\rangle_D && \text{(A.20)} \\
&= \langle \Phi(\theta^* - \theta), \Pi_D H\Phi\theta - \Phi\theta\rangle_D \\
&= \langle \Phi(\theta^* - \theta), \Pi_D H\Phi\theta - \Phi\theta^* + \Phi\theta^* - \Phi\theta\rangle_D \\
&= \|\Phi(\theta^* - \theta)\|_D^2 - \langle \Phi(\theta^* - \theta), \Phi\theta^* - \Pi_D H\Phi\theta\rangle_D \\
&\geq \|\Phi(\theta^* - \theta)\|_D^2 - \|\Phi(\theta^* - \theta)\|_D \cdot \|\Pi_D H\Phi\theta - \Phi\theta^*\|_D \\
&\geq \|\Phi(\theta^* - \theta)\|_D^2 - \gamma \cdot \|\Phi(\theta^* - \theta)\|_D^2 && \text{(A.21)} \\
&= (1-\gamma) \cdot \|\Phi(\theta^* - \theta)\|_D^2 = (1-\gamma) \cdot \|J_{\theta^*} - J_\theta\|_D^2,
\end{aligned}$$

where in going to Equation (A.19), we used that $\forall\ \mathbf{x} \in \text{Span}(\Phi)$, we have $\Pi_D\,\mathbf{x} = \mathbf{x}$. In Equation (A.20), we used that the projection matrix Π_D is symmetric. In going to Equation (A.21), we used that that $\Pi_D H(\cdot)$ is a contraction operator with modulus γ with $\Phi\theta^*$ as its fixed point, which implies that $\|\Pi_D H\Phi\theta - \Phi\theta^*\|_D = \|\Pi_D H\Phi\theta - \Pi_D H\Phi\theta^*\|_D \leq \gamma\|\Phi\theta - \Phi\theta^*\|_D$. □

Finally, we restate and prove Lemmas 14 and 15 used in Section 2.11 to analyze Q-learning for optimal stopping problems under a linear approximation model.

Lemma 14. *For any fixed* $\theta \in \mathbb{R}^d$, $\mathbb{E}\left[\|g_t(\theta)\|_2^2\right] \leq 2\sigma^2 + 8\|Q_\theta - Q_{\theta^*}\|_D^2$ *where* $\sigma^2 = \mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right]$.

Proof. We use notation and proof strategy mirroring the proof of Lemma 5. Set $\phi = \phi(s_t)$, $\phi' = \phi(s_t')$ and $U' = U(s')$. Define $\xi = (\theta^* - \theta)^\top\phi$ and $\xi' = (\theta^* - \theta)^\top\phi'$. By stationarity ξ and ξ' have

the same marginal distribution and $\mathbb{E}[\xi^2] = \|Q_{\theta^*} - Q_\theta\|_D^2$. Using the formula for $g_t(\theta)$ in Equation (2.33), we have

$$\begin{aligned}
\mathbb{E}\left[\|g_t(\theta)\|_2^2\right] &\leq 2\mathbb{E}\left[\|g_t(\theta^*)\|_2^2\right] + 2\mathbb{E}\left[\|g_t(\theta) - g_t(\theta^*)\|_2^2\right] \\
&= 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\phi^\top(\theta^* - \theta) - \gamma\left[\max\left(U', \phi'^\top\theta^*\right) - \max\left(U', \phi'^\top\theta\right)\right]\right)\right\|_2^2\right] \\
&\leq 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\left|\phi^\top(\theta^* - \theta)\right| + \gamma\left|\max\left(U', \phi'^\top\theta^*\right) - \max\left(U', \phi'^\top\theta\right)\right|\right)\right\|_2^2\right] \\
&\leq 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\left|\phi^\top(\theta^* - \theta)\right| + \gamma\left|\phi'^\top(\theta^* - \theta)\right|\right)\right\|_2^2\right] \qquad \text{(A.22)} \\
&\leq 2\sigma^2 + 2\mathbb{E}[|\xi + \gamma\xi'|^2] \\
&\leq 2\sigma^2 + 4\left(\mathbb{E}\left[|\xi|^2\right] + \gamma^2\mathbb{E}\left[|\xi'|^2\right]\right) \\
&= 2\sigma^2 + 4(1+\gamma^2)\|Q_\theta - Q_{\theta^*}\|_D^2 \leq 2\sigma^2 + 8\|Q_\theta - Q_{\theta^*}\|_D^2,
\end{aligned}$$

where we used the assumption that features are normalized so that $\|\phi\|_2^2 \leq 1$ almost surely. Additionally, in going to Equation (A.22), we used that $|\max(c_1, c_3) - \max(c_2, c_3)| \leq |c_1 - c_2|$ for any scalars c_1, c_2 and c_3. □

Lemma 15. *Define $G = (r_{\max} + 2R)$. With probability 1, $\|g_t(\theta)\|_2 \leq G$ for all $\theta \in \Theta_R$.*

Proof. We start with the mathematical expression for the semi-gradient,

$$g_t(\theta) = \left(u(s_t) + \gamma \max\left\{U(s_t'), \phi(s_t')^\top\theta\right\} - \phi(s_t)^\top\theta\right)\phi(s_t).$$

As $r_{\max} \leq R$, we have: $\max\left\{U(s_t'), \phi(s_t')^\top\theta\right\} \leq \max\left\{U(s_t'), \left\|\phi(s_t')\right\|_2\|\theta\|_2\right\} \leq R$. Then,

$$\begin{aligned}
\|g_t(\theta)\|_2^2 &= \left(u(s_t) + \gamma \max\left\{U(s_t'), \phi(s_t')^\top\theta\right\} - \phi(s_t)^\top\theta\right)^2 \|\phi(s_t)\|^2 \\
&\leq \left(r_{\max} + \gamma R - \phi(s_t)^\top\theta\right)^2 \\
&\leq \left(r_{\max} + \gamma R + \|\phi(s_t)\|_2\|\theta\|_2\right)^2 \leq (r_{\max} + 2R)^2 = G^2.
\end{aligned}$$

We used here that the basis vectors are normalized, $\|\phi(s_t)\|_2 \leq 1$ for all t. □

Appendix B: Additional details for Chapter 3

B.1 Background on Bellman operators and policy iteration

We make repeated use of the basic element-wise inequalities

$$TJ \preceq T_\pi J \quad \text{and} \quad TJ_\pi \preceq J_\pi \tag{B.1}$$

which hold for any policy π. The first inequality follows since T minimizes over actions, $TJ(s) = \min_{a \in \mathcal{A}} Q(s,a)$ (see Equation (3.2)). The second inequality follows by the first since $J_\pi = T_\pi J_\pi$. An important property that we repeatedly make use of is that for bounded cost-to-go functions, J, J', both the Bellman optimality operator T and the Bellman operator T_π for a stationary policy π are contraction operators with respect to the maximum norm with modulus γ. Precisely,

$$\|J - J'\|_\infty \leq \gamma \|J - J'\|_\infty \qquad \|T_\pi J - T_\pi J'\|_\infty \leq \gamma \|J - J'\|_\infty. \tag{B.2}$$

These operators are also monotone, meaning the element-wise inequality $J \preceq J'$ implies $TJ \preceq TJ'$ and $T_\pi J \preceq T_\pi J'$. A simple argument using contractivity of T and T_π together with the triangle inequality shows that for any bounded cost function J_π,

$$\|J_\pi - J^*\|_\infty \leq \frac{1}{1-\gamma} \|J_\pi - TJ_\pi\|_\infty \tag{B.3}$$

where J^* is the optial cost-to-go function. Equation (B.3) can be shown using the definitions: $J_\pi = T_\pi J_\pi$ and $J^* = TJ^*$,

$$\begin{aligned}\|J_\pi - J^*\|_\infty = \|T_\pi J_\pi - TJ_\pi + TJ_\pi - J^*\|_\infty &\leq \|T_\pi J_\pi - TJ_\pi\|_\infty + \|TJ_\pi - TJ^*\|_\infty \\ &\leq \|T_\pi J_\pi - TJ_\pi\|_\infty + \gamma\|J_\pi - J^*\|_\infty\end{aligned}$$

The result in (B.3) is very useful for our analysis as it indicates that near-solutions to the the Bellman equation $J^* = TJ^*$ must themselves be close to the optimal cost-to-go function J^*. The reinforcement learning literature widely uses versions of this inequality that are sensitive to the state distribution. For each bounded function J,

$$J - J_\pi = J - T_\pi J + T_\pi J - T_\pi J_\pi = J - T_\pi J + \gamma P_\pi (J - J_\pi) = \cdots = \sum_{t=0}^{\infty} \gamma^t P_\pi^t \cdot (J - T_\pi J)\,. \quad \text{(B.4)}$$

This expresses the difference of J from J_π in terms of the gap in Bellman's equations at the states visited by the policy π. An especially useful case of this result arises when $\pi = \pi^*$ is an optimal policy in which case,

$$J - J^* \preceq \sum_{t=0}^{\infty} \gamma^t P_{\pi^*}^t \cdot (J - TJ)\,, \quad \text{(B.5)}$$

where the inequality uses (B.1) to conclude $TJ \preceq T_{\pi*}J$. Related inequalities are sometimes called the *Performance difference lemma* [78] in the reinforcement learning literature.

The classic policy iteration algorithm due to [154] can be expressed compactly in terms of Bellman operators. Starting with an initial policy π, the algorithm first evaluates the corresponding cost to go function J_π, and then finds the policy π^+ that attains the minimum in the Bellman update, $\pi^+ = \arg\min_{\tilde{\pi}} T_{\tilde{\pi}} J_\pi$. Equivalently, this can be written as $T_{\pi^+} J_\pi = TJ_\pi$. This implies,

$$J_\pi \geq T_{\pi^+} J_\pi \geq T_{\pi^+}^2 J_\pi \geq \cdots \geq J_{\pi^+}$$

where the first inequality applies (B.1) and the rest follow by inductively applying T_{π^+} to each side and using the monotonicity property of the Bellman operator. The first inequality is strict unless

$J_\pi = T_{\pi^+} J_\pi = TJ_\pi$, in which case $J_\pi = J^*$ and π is an optimal policy. From Equation (B.4) or its more refined variant (B.5), we can see that each step of policy iteration leads to a substantial cost reduction unless the policy is near optimal. We conclude with the proof of a basic extension of Bellman's equation used in our analysis, which we restate here. Recall, η_π to be the discounted state-occupancy measure under policy π (see (3.6)).

Lemma 22 (On average Bellman equation)**.** *For any $\pi \in \Pi$ and $S \sim \eta_\pi$,*

$$J_\pi = J^* \iff \mathbb{E}[J_\pi(S)] = \mathbb{E}[TJ_\pi(S)]$$

Proof. First note that standard results in dynamic programming imply $J_\pi \geq TJ_\pi$ and $J_\pi \geq J^*$ (these in fact hold for any arbitrary policy π and not just for $\pi \in \Pi_\Theta$).

Let $J : \mathcal{S} \to \mathbb{R}$ be an arbitrary cost-to-go function such that $J \geq 0$. Then, we have $\mathbb{E}[J(S)] = 0 \iff J = 0$. To see this, note that the non-negativity of J implies we must have $J(S) = 0$ almost surely. Since $S \sim \eta_\pi \geq (1-\gamma)\rho$ and by assumption, the initial distribution ρ is supported over $\mathcal{S}$, $J(S) = 0$ almost surely if and only if $J(s) = 0$ for all $s \in \mathcal{S}$. Applying this with choice of $J = J_\pi - J^*$ or $J = J_\pi - TJ_\pi$ shows the average Bellman equation above is equivalent to the standard result,

$$J_\pi = J^* \iff J_\pi = TJ_\pi.$$

□

B.2 Background: First order methods

To start, let us define some standard notions from first order optimization. For a convex set $\mathcal{X} \subset \mathbb{R}^d$, we say a function $f : \mathcal{X} \to \mathbb{R}$ is k-Lipshitz if $\|f(x) - f(y)\|_2 \leq k\|x - y\|_2$ for every $x, y \in \mathcal{X}$. We say a function is L-smooth if f is differentiable throughout $\mathcal{X}$ and ∇f is L-Lipschitz. A consequence of smoothness that will be useful throughout our proofs is often called the *descent lemma*. It implies a quadratic upper bound on function values. The proof follows by Taylor expansion and the mean-value theorem [108].

Lemma 38 (Descent Lemma)**.** *If the function $f : \mathcal{D} \to \mathbb{R}$ is L-smooth over a set $\mathcal{X} \subseteq \mathcal{D}$, then for any $(x, y) \in \mathcal{X}$:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2.$$

The following interpretation of projected gradient updates will be very useful for our proof. Recall the notation for orthogonal projection. $\text{Proj}_{\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \|y - x\|_2^2$. The projected gradient descent iteration can be written as

$$x_{t+1} = \text{Proj}_{\mathcal{X}}\left(x_t - \alpha_t \nabla f(x_t)\right) = \arg\min_{x \in \mathcal{X}} \left[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\alpha_t}\|x - x_t\|_2^2 \right]. \tag{B.6}$$

The first equality is the usual definition of projected gradient descent. The second gives a "proximal" interpretation of projection as minimizing a local quadratic approximation. See [eg. 118] for a simple proof.

B.2.1 Asymptotic convergence to stationary points: proof of Lemma 16

For convenience, we first restate the claim.

Lemma 16. *Consider the optimization problem $\min_{x \in \mathcal{X}} f(x)$ where $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set. Assume f is bounded below and its α–sub-level sets $\{x \in \mathcal{X} : f(x) \leq \alpha\}$ are bounded for each $\alpha \in \mathbb{R}$. Consider the sequence $x_{k+1} = \text{Proj}_{\mathcal{X}}\left(x_k - \alpha \nabla f(x_k)\right)$*

1. *(Amir Beck, [119, 118]) Assume f is differentiable on an open set containing $\mathcal{X}$ and its gradient ∇f is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant L. If $\alpha \in (0, 1/L]$, the sequence $\{x_k\}$ has at least one limit point and any limit point x_∞ is a stationary point of $f(\cdot)$ on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.*

2. *Suppose $f(\cdot)$ is continuously twice differentiable on an open set containing the sub-level set $\{x \in \mathcal{X} : f(x) \leq f(x_0)\}$ and its gradient ∇f is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant L. Then, for a sufficiently small $\alpha > 0$, the sequence $\{x_k\}$ has at least one limit point and any limit point x_∞ is a stationary point of $f(\cdot)$ on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.*

Proof. We refer the readers to the simple proofs in [119, 118] for part 1. To show the claim in part 2, note that the sub-level set $S := \{x \in \mathcal{X} : f(x) \leq f(x_0)\}$ is compact (continuity of $f(\cdot)$ implies its closed and we assumed it to be bounded). Also, for a sufficiently small ϵ, $f(\cdot)$ is twice continuously differentiable over the compact set,

$$S_\epsilon := \{x + y : x \in S_1, \|y\| \leq \epsilon\}.$$

which follows by our assumption that $f(\cdot)$ is twice continuously differentiable on an open set containing S. We denote $G = \max_{x \in S} \|\nabla f(x)\|_2$ and $L = \max_{x \in S_\epsilon} \|\nabla^2 f(x)\|_2$. Note that $G, L < \infty$ as ∇f is continuous by assumption and S, S_ϵ are compact sets. For any $x \in S_1$, define $x^+ = \text{Proj}_{\mathcal{X}}\left(x - \alpha \nabla f(x)\right)$. For the step-size $\alpha = \min\{\epsilon/G, 1/L\}$, it is easy to see that $x^+ \in S_\epsilon$

$$\|x^+ - x\|_2 = \|\text{Proj}_{\mathcal{X}}\left(x - \alpha \nabla f(x)\right) - \text{Proj}_{\mathcal{X}}(x)\|_2 \leq \|\alpha \nabla f(x)\|_2 \leq \epsilon,$$

which follows as projection operators are non-expansive. By using a standard property of projection operators [See Theorem A.1.3 in 119], we get

$$\langle x - \alpha \nabla f(x) - x^+, x - x^+ \rangle \leq 0 \;\; \Rightarrow \;\; \|x - x^+\|^2 - \alpha \langle \nabla f(x), x - x^+ \rangle \leq 0.$$

As $x^+ \in S_\epsilon$, using smoothness of $f(\cdot)$ over S_ϵ implies,

$$\begin{aligned} f(x^+) &\leq f(x) + \langle \nabla f(x)\,,\, x^+ - x \rangle + \frac{L}{2}\|x^+ - x\|^2 \\ &\leq f(x) + \left(\frac{L}{2} - \frac{1}{\alpha}\right)\|x^+ - x\|^2 \end{aligned}$$

which implies $f(x^+) \leq f(x)$ as we choose[1] $\alpha \leq \frac{1}{L}$. Therefore, for a small enough step-size, the projected gradient update reduces cost monotonically, $f(x^+) \leq f(x)$ and therefore stays in the sub-level set, $x^+ \in S$. Repeating this argument inductively shows that $\{x_k\} \subset S$ and $f(x_{k+1}) \leq$

[1] It is easy to see that when x is not a stationary point, i.e. $x^+ \neq x$, the inequality is strict.

$f(x_k)\ \forall\, k$.

Since $\{x_k\}$ is contained in a compact set S, it has a convergent sub-sequence, $\{x_{k_i}\}$ with some limit point x_∞. We have

$$\lim_{k\to\infty} f(x_k) = \lim_{i\to\infty} f(x_{k_i}) = f(x_\infty),$$

where the final inequality uses continuity of $f(\cdot)$. Note, the limit of $f(x_t)$ is assured to exist because the sequence is monotone decreasing and bounded below. The proof to show that any limit point, x_∞ is a stationary point, follows exactly the same argument as shown in [119]. We omit this for brevity. □

B.2.2 Convergence rates under gradient dominance: Proof of Lemma 27.

We first restate the claim.

Lemma 27 (Convergence rates for gradient dominated smooth functions)**.** *Consider the problem,* $\min_{x\in\mathcal{X}} f(x)$ *where* $\mathcal{X} \subseteq \mathbb{R}^d$. *Assume* f *be* L*–smooth on* $\mathcal{X}$ *and a non-empty solution set. Denote* $f(x^*) = \min_{x'\in\mathcal{X}} f(x')$. *Consider the sequence* $x_{t+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_t - \alpha\nabla f(x_t)\right)$.

1. *Let* $\mathcal{X} \subset \mathbb{R}^d$ *such that* $\|x - x'\|_2 \le R < \infty$ *for all* $x, x' \in \mathcal{X}$. *Assume* f *is* k*-lipschitz and* $(c, 0)$*–gradient-dominated and where* $\alpha \le \min\{\frac{1}{k}, \frac{1}{L}\}$. *Then,*

$$\min_{t\le T}\{f(x_t) - f(x^*)\} \le \sqrt{\frac{2R^2c\,(f(x_0) - f(x^*))}{\alpha T}}$$

2. *Karimi et. al. [74], Polyak [72] Assume* $\mathcal{X} = \mathbb{R}^d$ *and* $\alpha = 1/L$. *If* f *is* (c, μ)*–gradient-dominated for* $\mu > 0$*, then,*

$$f(x_t) - f(x^*) \le \left(1 - \frac{\mu}{c^2 L}\right)^t (f(x_0) - f(x^*))$$

Proof of Lemma 27. Recall, by Definition 3 that a function f is defined to be (c, μ)*–gradient dom-*

inated over $\mathcal{X}$ if there exists a constant $c > 0$ and $\mu \geq 0$ such that

$$f(x^*) \geq f(x) + \min_{y \in \mathcal{X}} \left[c \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \right] \quad \forall x \in \mathcal{X}.$$

Proof of Part (a): We assume $\mu = 0$ in which case for any $x \in \mathcal{X}$, we have

$$\min_{y \in \mathcal{X}} \left[c \langle \nabla f(x), y - x \rangle \right] \leq f(x^*) - f(x) \tag{B.7}$$

Therefore, for any $x \neq x^*$, we have $\min_{y \in \mathcal{D}} \langle \nabla f(x_t), y - x \rangle < 0$. Let $\{x_t\}$ be the iterates produced by projected gradient descent. At iterate x_t, let $\bar{y} = \arg\min_{y \in \mathcal{X}} \langle \nabla f(x_t), y - x_t \rangle$ and denote $\delta_t = \min_{y \in \mathcal{X}} \langle \nabla f(x_t), y - x_t \rangle$. Note that $\delta_t \leq 0$ and $|\delta_t| \leq \|\nabla f(x_t)\| \|y - x_t\| \leq kR$ as f is assumed to be k-Lipschitz. We take a constant stepsize, $\alpha_t = \alpha \leq \min\{\frac{1}{k}, \frac{1}{L}\}$. Then,

$$\begin{aligned}
f(x_{t+1}) - f(x_t) &\overset{(a)}{\leq} \min_{y \in \mathcal{D}} \left[\langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2\alpha} \|y - x_t\|_2^2 \right] \\
&\overset{(b)}{=} \min_{\beta \in [0,1]} \left[\langle \nabla f(x_t), x_t + \beta(\bar{y} - x_t) - x_t \rangle + \frac{1}{2\alpha} \|x_t + \beta(\bar{y} - x_t) - x_t\|_2^2 \right] \\
&= \min_{\beta \in [0,1]} \left[\beta \langle \nabla f(x_t), (\bar{y} - x_t) \rangle + \frac{\beta^2}{2\alpha} \|\bar{y} - x_t\|_2^2 \right] \\
&\leq \min_{\beta \in [0,1]} \left[\beta \delta_t + \frac{\beta^2 R^2}{2\alpha} \right] = \frac{-\alpha \delta_t^2}{2R^2}
\end{aligned} \tag{B.8}$$

where the minimizer $\beta^* = -\delta_t \alpha / R^2 \leq k\alpha / R \leq 1$ as $\alpha \leq \min\{\frac{1}{k}, \frac{1}{L}\}$ (we assume $R > 1$ without loss of generality as we can take any upper bound while minimizing in (B.8)). Here (a) follows by using the equivalence shown in (B.6) and the quadratic upper bound on the function values implied by the descent lemma. Equality (b) uses the fact that right hand side of (a) can be optimized by searching over the steepest descent direction $x_t \to y$. Using (B.7), we get

$$f(x_{t+1}) - f(x_t) \leq \frac{-\alpha}{2R^2 c^2} \left(f(x^*) - f(x_t) \right)^2$$

Rearraning, we get our desired result

$$\begin{aligned}\min_{t\leq T}\left(f(x_t)-f(x^*)\right)^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\left(f(x_t)-f(x^*)\right)^2 &\leq \frac{2R^2c^2}{\alpha T}\sum_{t=0}^{T-1} f(x_t)-f(x_{t+1})\\ &\leq \frac{2R^2c^2}{\alpha T}\left(f(x_0)-f(x_T)\right)\\ &\leq \frac{2R^2c^2}{\alpha T}\left(f(x_0)-f(x^*)\right)\end{aligned}$$

Therefore,

$$\min_{t\leq T}\left\{f(x_t)-f(x^*)\right\} \leq \sqrt{\frac{2R^2c^2\left(f(x_0)-f(x^*)\right)}{\alpha T}}$$

Proof of Part (b): Please see the proof in [74] which can be dated back to the work of [72]. □

B.3 On the necessity of an exploratory initial distribution

The following example is sometimes called a "chain" MDP [155, 78] or the "river swim" problem [156, 157]. Many other examples in the reinforcement learning literature, like the "combination lock" problem [158] and the "grid world" problem [159] highlight the same issue. While these examples are typically used to highlight a statistical challenge, here we focus on the optimization landscape. This example is partly inspired by one in [78]. A similar discussion appears also in [99]. We include this discussion to keep the paper self contained. In addition, it does not seem that past work has shown clearly that $\ell(\pi)$ may have suboptimal local minima in the absense of an exploratory initial distribution, instead showing the existence of suboptimal polices with small but nonzero gradient norm.

Example 6. *Consider the MDP depicted below. There are N states and from each state the agent can move either left or right from each state. The decision-maker (DM) always begins in the leftmost state (i.e.* $\rho(s_1) = 1$*). The DM incurs a cost of 2 per-period when in any state other than the leftmost or rightmost state, a cost* $g(s_1) = 1$ *from the leftmost state and a cost of* $g(s_N) = 0$ *per period in when in the rightmost state. A stationary policy* $\pi \in [0,1]^N$ *is a vector where* $\pi(s)$

specifies the probability of choosing the action R in state S. When the horizon is sufficiently long, the optimal policy moves right in each period. From Lemma 20, one can calculate the policy gradient as

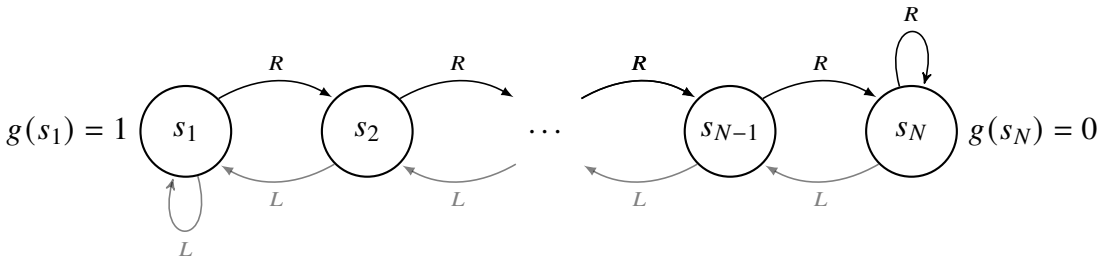
$$\frac{\partial \ell(\pi)}{\partial \pi(s)} = \eta_\pi(s) \left(Q_\pi(s, R) - Q_\pi(s, L)\right).$$

A suboptimal policy π that always moves left is a local minimum of $\ell(\cdot)$. To see this, first note that the agent will always start and stay in the leftmost state, so $\eta_\pi(s_i) = 0$ when $i \geq 2$. The only possible nonzero component of $\nabla \ell(\pi)$ is the first term corresponding to state s_1. Therefore, for any policy $\pi' \in [0, 1]^N$,

$$\langle \nabla \ell(\pi), \pi' - \pi \rangle = \eta_\pi(s_1) \left(Q_\pi(s_1, R) - Q_\pi(s_1, L)\right) \left(\pi'(s_1) - \pi(s_1)\right) \geq 0,$$

which follows as $Q_\pi(s_1, R) > Q_\pi(s_1, L)$, given that moving to s_2 for a single period is more costly than staying in s_1 and the fact that $\pi(s_1) = 0$, so $\pi'(s_1) - \pi(s_1) \geq 0$ for any feasible policy π'.

Similar issues arise under a (non-degenerate) stochastic policy. The main idea is that policies that are more likely to move left from every *are expected to require exponentially (in N) many periods to reach the rightmost state. An explicitly bound confirming that the policy gradient can be exponentially small in N is given in [99].*



B.4 Details for LQ control

Throughout this section, we consider the linear quadratic control problem as described in Example 2, continuing with the notation and assumptions introduced there.

Smoothness. We first show that the policy gradient objective for LQ control is smooth over sub-level sets.

Lemma 18. *Consider the linear quadratic control problem formulated in Example 2. The set Θ_S is open and ℓ is twice continuously differentiable on Θ_S. For any $\alpha \in \mathbb{R}$, the sublevel set $C_\alpha := \left\{\theta \in \mathbb{R}^{n\times k} : \ell(\theta) \leq \alpha\right\}$ is a compact subset of Θ_S and $\sup_{\theta\in C_\alpha} \|\nabla^2 \ell(\theta)\| < \infty$.*

Proof. We first show that any sub-level set only contains stable policies, $C_\alpha \subset \Theta_S$. Recall, we assumed $\Sigma := \mathbb{E}_{s\sim\rho}[ss^\top] \succ 0$. As shown in Section 3.5.1, we can write the total cost function corresponding to a linear policy, π_θ, as:

$$\ell(\theta) = \mathbb{E}_{s\sim\rho}\left[J_{\pi_\theta}(s)\right] = \mathbb{E}_{s\sim\rho}\left[s^\top K_\theta s\right] \geq \|K_\theta\|_2 \, \lambda_{\min}(\Sigma)$$

where $K_\theta \in \mathbb{R}^{n\times n}$ is defined in Example 2. Therefore, $\ell(\theta) < \infty$ implies that $\|K_\theta\|_2 < \infty$. Then,

$$\begin{aligned} n\|K_\theta\|_2 \geq \mathrm{Tr}(K_\theta) &= \mathrm{Tr}\left(\sum_{t=0}^{\infty} \gamma^t \left[(A+B\theta)^t\right]^\top (\theta^\top R\theta + C)\left[(A+B\theta)^t\right]\right) \\ &\geq \lambda_{\min}(\theta^\top R\theta + C) \sum_{t=0}^{\infty} \gamma^t \mathrm{Tr}\left(\left[(A+B\theta)^t\right]^\top \left[(A+B\theta)^t\right]\right) \\ &\geq \lambda_{\min}(\theta^\top R\theta + C) \sum_{t=0}^{\infty} \gamma^t \|A+B\theta\|_2^{2t} \end{aligned}$$

where we use standard properties of the trace operator. Clearly, $\sum_{t=0}^{\infty} \gamma^t \|A+B\theta\|_2^{2t} \to \infty$ for any $\theta \notin \Theta_S$. As $R, C \succ 0$, this implies that $\|K_\theta\|_2 \to \infty$ for any $\theta \notin \Theta_S$.

With a little bit of algebra ([125]), it is easy to see that $\ell(\theta)$ is twice continuously differentiable for any $\theta \in \Theta_S$ and hence over sublevel sets as $C_\alpha \subset \Theta_S$. We show that sub-level sets are compact by showing that they are closed and bounded. As ℓ is continuous, by definition its sub-level sets are closed. For the class of linear policies, $\pi_\theta(s) = \theta s$, we can show $\ell(\theta)$ is a coercive function, that is $\lim_{\|\theta\|_2\to\infty} \ell(\theta) = +\infty$. To see this, consider

$$\ell(\theta) = \mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty} \gamma^t s_t^\top (\theta^\top R\theta + C) s_t\right]$$

where s_t evolves according to linear dynamics, $s_t = (A + B\theta)\, s_{t-1}$. Define $\Sigma_\theta := \mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t s_t s_t^\top\right]$. Clearly, $\Sigma_\theta \succ \Sigma = \mathbb{E}_{s_0\sim\rho}\left[s_0 s_0^\top\right] \succ 0$. Therefore,

$$\ell(\theta) = \text{Trace}\left((\theta^\top R\theta + C)\Sigma_\theta\right) \geq \lambda_{\min}(\Sigma_\theta)\text{Trace}\left(\theta^\top R\theta + C\right) \geq \lambda_{\min}(\Sigma_\theta)\|\theta^\top R\theta + C\|_2$$
$$\Rightarrow \ \|\theta^\top R\theta\|_2 \leq \frac{\ell(\theta)}{\lambda_{\min}(\Sigma_\theta)} + \|C\|_2. \tag{B.9}$$

As we assume $\|R\|_2, \|C\|_2 < \infty$ and $\lambda_{\min}(\Sigma_\theta)$ is uniformly lower bounded ($\lambda_{\min}(\Sigma_\theta) > \lambda_{\min}(\Sigma) > 0$), it is clear by (B.9) that $\lim_{\|\theta\|_2\to\infty}\ell(\theta) = +\infty$. By definition, the sub-level sets of a coercive function are bounded (see [160] for example) which completes our argument.

Now, it is easy to show that $\ell(\cdot)$ is smooth over sub-level sets. By definition, any twice differentiable function $f : \mathcal{X} \to \mathbb{R}$ is smooth on a subset $D \subseteq \mathcal{X}$ if $\nabla^2 f(x) \preceq LI$ for some constant $L < \infty$. As $\ell(\cdot)$ is twice continuously differentiable and sub-level sets of $\ell(\cdot)$ are compact, by Extreme Value Theorem, we get that $\|\nabla^2\ell(\theta)\|$ is bounded on sub-level sets, i.e. $\sup_{\theta\in C_\alpha}\|\nabla^2\ell(\theta)\| < \infty$. □

Concentrability coefficient.

Lemma 26 (Concentrability in LQ control)**.** *Consider the linear quadratic control problem in Example 2. Suppose* $\Sigma_\rho := \mathbb{E}_{s_0\sim\rho}[s_0 s_0^\top] \succ 0$ *and let* $\theta^* \in \mathbb{R}^{n\times k}$ *denote the parameter of an optimal policy. Then,*

$$\kappa_\rho \leq \frac{(1-\gamma)}{1-\gamma\|A + B\theta^*\|_2^2} \cdot \frac{\lambda_{\max}\left(\Sigma_\rho\right)}{\lambda_{\min}\left(\Sigma_\rho\right)}.$$

Proof. First observe that by monotonicity of the Bellman operator as shown in Lemma 17, we have $J_{\pi_\theta} \succ TJ_{\pi_\theta}$. Following Lemma 17, $TJ_{\pi_\theta} \in \mathcal{J}_q$ and therefore $J_{\pi_\theta} - TJ_{\pi_\theta} \in \mathcal{J}_q$. Then,

$$\|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\rho} = \mathbb{E}_{s\sim\rho}\left[J_{\pi_\theta}(s) - TJ_{\pi_\theta}(s)\right] = \mathbb{E}_{s\sim\rho}\left[s^\top K s\right] = \text{Trace}(K\Sigma_\rho). \tag{B.10}$$

for some $K \in \mathbb{R}^{n\times n}, K \succ 0$ which simplifies the right hand side in the definition of κ_ρ in (3.20). To

simplify the left hand side, we use the variational Bellman equation in (3.5) to show

$$(J_{\pi_\theta} - J_{\pi_{\theta^*}})(s_0) = \sum_{t=0}^{\infty} \gamma^t \left(J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_\theta}\right)(s_t^*)$$

where $s_t^* = (A + B\theta^*)^t s_0$ is the state at time t if π_{θ^*} is applied from initial state s_0. To see this, recall that by definition,

$$(T_{\pi_\theta} J)(s) = (\theta s)^\top R(\theta s) + s^\top C s + \gamma J(As + B\theta s). \tag{B.11}$$

From this, and the Bellman equation, we have

$$J_{\pi_\theta} - J_{\pi_{\theta^*}} = T_{\pi_\theta} J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_{\theta^*}} = \left(T_{\pi_\theta} J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_\theta}\right) + \left(T_{\pi_{\theta^*}} J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_{\theta^*}}\right)$$

Applying this from a particular state s_0, and using (B.11) gives

$$\begin{aligned} J_{\pi_\theta}(s_0) - J_{\pi_{\theta^*}}(s_0) &= \left(T_{\pi_\theta} J_{\pi_\theta}(s_0) - T_{\pi_{\theta^*}} J_{\pi_\theta}(s_0)\right) + \gamma \left(J_{\pi_\theta}((A + B\theta^*)s_0) - J_{\pi_{\theta^*}}((A + B\theta^*)s_0)\right) \\ &= \left(T_{\pi_\theta} J_{\pi_\theta}(s_0) - T_{\pi_{\theta^*}} J_{\pi_\theta}(s_0)\right) + \gamma \left(J_{\pi_\theta}(s_1^*) - J_{\pi_{\theta^*}}(s_1^*)\right). \end{aligned}$$

The variation Bellman formula follows by iterating over this recursion. Note that as $TJ_{\pi_\theta} = \inf_\pi T_\pi J_{\pi_\theta} \preceq T_{\pi_{\theta^*}} J_{\pi_\theta}$, we have

$$(J_{\pi_\theta} - J_{\pi_{\theta^*}})(s_0) \leq \sum_{t=0}^{\infty} \gamma^t \left(J_{\pi_\theta} - TJ_{\pi_\theta}\right)(s_t^*) = \sum_{t=0}^{\infty} \gamma^t (s_t^*)^\top K s_t^* = \text{Trace}\left(K \sum_{t=0}^{\infty} \gamma^t s_t^* (s_t^*)^\top\right)$$

Taking expectations gives

$$\|J_{\pi_\theta} - J_{\pi_{\theta^*}}\|_{1,\rho} = \mathbb{E}_{s\sim\rho}\left[(J_{\pi_\theta} - J_{\pi_{\theta^*}})(s_0)\right] \leq \frac{1}{1-\gamma} \cdot \text{Trace}\left(K\Sigma_{\eta_{\theta^*}}\right), \tag{B.12}$$

where we define

$$\Sigma_{\eta_{\theta^*}} = (1-\gamma)\mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t s_t^*\left(s_t^*\right)^\top\right] = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\left[(A+B\theta^*)^t\right]\Sigma_\rho\left[(A+B\theta^*)^t\right]^\top.$$

Combining (B.10) and (B.12) gives

$$\|J_{\pi_\theta} - J_{\pi_{\theta^*}}\|_{1,\rho} \le \|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\rho}\cdot\frac{\text{Trace}\left(K\Sigma_{\eta_{\theta^*}}\right)}{(1-\gamma)\text{Trace}\left(K\Sigma_\rho\right)} \le \|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\rho}\cdot\frac{\lambda_{\max}\left(\Sigma_{\eta_{\theta^*}}\right)}{(1-\gamma)\lambda_{\min}\left(\Sigma_\rho\right)}$$

The last inequality uses a standard property of the trace operator that for any two positive semidefinite symmetric matrices A, B: $\lambda_{\min}(A)\text{Trace}(B) \le \text{Trace}(AB) \le \lambda_{\max}(A)\text{Trace}(B)$, where $\lambda_{\min}(A), \lambda_{\max}(A)$ are the minimum and maximum eigenvalues of A respectively. See [161] for example. Comparing to (3.20) gives us our desired result

$$\kappa_\rho \le \frac{\lambda_{\max}\left(\Sigma_{\eta_{\theta^*}}\right)}{\lambda_{\min}\left(\Sigma_\rho\right)} \le \frac{(1-\gamma)}{1-\gamma\|A+B\theta^*\|_2^2}\cdot\frac{\lambda_{\max}\left(\Sigma_\rho\right)}{\lambda_{\min}\left(\Sigma_\rho\right)}.$$

where the final inequality uses that

$$\|\Sigma_{\eta_{\theta^*}}\|_2 \le (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\|A+B\theta^*\|_2^{2t}\|\Sigma_\rho\|_2 = \frac{(1-\gamma)}{1-\gamma\|A+B\theta^*\|_2^2}\cdot\|\Sigma_\rho\|_2$$

□

B.5 Details for Optimal Stopping

In this section, we first verify Condition 0 for the policy gradient lemma to hold for the optimal stopping problem as described in Example 4. Recall, we assume the context set $\mathcal{X}$ to be finite and the offer set to be bounded, $\mathcal{Y} = [y_{\min}, y_{\max}]$ and we take $y_{\min} > 0$ without loss of generality. We consider the class of threshold policies, $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta = [y_{\min}, y_{\max}]^{|\mathcal{X}|}\}$ and assume that the offer distribution has a density $q_x(\cdot)$ supported over $\mathcal{Y}$.

Notation. For any policy $\pi_\theta \in \Pi_\Theta$, we simplify notation to write $\eta_\theta := \eta_{\pi_\theta}$, $T_\theta := T_{\pi_\theta}$, $J_\theta := J_{\pi_\theta}$ and $P_\theta := P_{\pi_\theta}$. Clearly, for any $\theta \in \Theta$, the discounted state occupancy measure η_θ factorizes as $\eta_\theta(x, y) = \eta'_\theta(x)q_x(y)$, where η'_θ denotes the marginal distribution over $\mathcal{X} \cup \{T\}$. We find it more convenient to directly work with $\eta'_\theta(x)$ and $q_x(y)$. We denote $P'_\theta \in \mathbb{R}^{(|\mathcal{X}|+1)\times(|\mathcal{X}|+1)}$ to be the transition matrix over $\mathcal{X} \cup \{T\}$ under π_θ, defined as

$$P'_\theta(x'|x) = p(x'|x) \int_{\mathcal{Y}} \mathbb{1}(y < \theta_x) q_x(y) dy, \quad P'_\theta(T|x) = 1 - \sum_{x' \in \mathcal{X}} P'_\theta(x'|x), \quad P'_\theta(T|T) = 1 \tag{B.13}$$

for all $x', x \in \mathcal{X}$.

Twice Continuous Differentiability. We verify the differentiability properties in Condition 0. for the policy gradient lemma to hold. We actually verify stronger twice differentiability results that will be helpful later. For any $\theta, \overline{\theta} \in \Theta$, we first show that $\mathcal{B}(\overline{\theta}|\eta_\theta, J_\theta) = \mathbb{E}_{s\sim\eta_\theta}\left[\left(T_{\overline{\theta}} J_\theta - J_\theta\right)(s)\right]$ is a twice continuously differentiable function of $\overline{\theta}$. Let $c_\pi(x)$ denote the continuation value of policy π as $c_\pi(x) = \gamma \sum_{(x',y')\in\mathcal{S}} p(x'|x) q_x(y') J_\pi(x', y')$. Then, in (3.14) we showed,

$$\frac{\partial}{\partial \overline{\theta}_x} \mathcal{B}(\overline{\theta}|\eta_\theta, J_\theta) = \eta'_\theta(x) q_x(\overline{\theta}_x) \left(c_{\pi_\theta}(x) - \overline{\theta}_x\right) \tag{B.14}$$

Note that (B.14) is itself continuously differentiable because of the assumption that $q_x(\cdot)$ is continuously differentiable.

Next, we show $\mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta)$ to be twice continuously differentiable in $\overline{\theta}$. Using that the continuation value from the terminal state T is zero, we can write

$$\mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta) = \sum_{x\in\mathcal{X}} \eta'_{\overline{\theta}}(x) \int J_\theta(x, y) q_x(y) dy$$

We show $\eta'_{\overline{\theta}} \in \mathbb{R}^{|\mathcal{X}|+1}$ is twice differentiable, which establishes the result. Note that for any $x \in \mathcal{X} \cup \{T\}$,

$$\eta'_{\overline{\theta}}(x) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \rho (P'_{\overline{\theta}})^t = (1-\gamma)\rho(I - \gamma P'_{\overline{\theta}})^{-1}$$

By our definition of $P'_{\overline{\theta}}$ in (B.13), its easy to see that $P'_{\overline{\theta}}$ is twice differentiable in $\overline{\theta}$ (using our assumption on $q_x(\cdot)$). Also note that $(I - \gamma P'_\theta)$ is invertible as $P'_{\overline{\theta}}$ is a stochastic matrix, i.e. $\|P'_{\overline{\theta}}\|_\infty \leq 1$. Therefore, the inverse function theorem implies that $\eta'_{\overline{\theta}}(\cdot)$ is twice continuously differentiable which establishes our claim for $\mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta)$.

Gradient Dominance. Next, we show that the total reward function, $\ell(\theta)$, for the optimal stopping problem is gradient dominated We first recall the claim.

Lemma 28 (Gradient dominance for optimal stopping)**.** *Denote κ_ρ to be the concentrability coefficient of the initial distribution $\rho(x, y) = \nu(x)q_x(y)$. Define $\psi := \min_{(x,y)\in\mathcal{S}_C} \nu(x)\, q_x(y)$ and $\beta := \max_{(x,y)\in\mathcal{S}_C} \nu(x)\, q_x(y)$. Then, $\ell(\cdot)$ is $(\frac{\beta}{\kappa_\rho \psi}, 0)$ gradient dominated.*

Proof. Recall that we formulate the optimal stopping problem in Example 4 as a maximization problem. Therefore, to show that $\ell(\cdot)$ is $(c, 0)$-gradient dominated for some $c > 0$, we need to show:

$$\ell(\theta^*) - \ell(\theta) \leq c \cdot \max_{\theta' \in \Theta} \ell(\theta)^\top (\theta' - \theta) \tag{B.15}$$

where θ^* denotes the optimal threshold and π_{θ^*} is the corresponding policy. The initial distribution factors as $\rho(x, y) = \nu(x)q_x(y)$ where $\nu(x) > 0 \; \forall x \in \mathcal{X}$. Then, the total cost function can be written as:

$$\ell(\theta) = (1 - \gamma) \sum_{x\in\mathcal{X}} \nu(x) \left[\int_{y\in\mathcal{Y}} J_\theta(x, y) q_x(y) dy \right]$$

for any $\pi_\theta \in \Pi_\Theta$. Similarly, the cost-to-go functions from any continuation state $(x, y) \in \mathcal{S}_C$ can be written as:

$$J^*(x, y) = \mathbb{1}\left(y > \theta^*_x\right) \cdot y + \mathbb{1}\left(y < \theta^*_x\right) \cdot c_{\pi^*}(x)$$
$$J_\theta(x, y) = \mathbb{1}\left(y > \theta_x\right) \cdot y + \mathbb{1}\left(y < \theta_x\right) \cdot c_{\pi_\theta}(x)$$

where $c_\pi(x) = \gamma \sum_{x'\in\mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J_\pi(x', y') q_x(y') dy'$ denotes the continuation value for any policy

$\pi \in \Pi_\Theta$.

We use two important facts for optimal stopping. First, for any context $x \in \mathcal{X}$, we have $c_{\pi_\theta}(x) < c_{\pi^*}(x)$ for all $\theta \neq \theta^*$ as $J_\theta(x,y) < J^*(x,y)\ \forall\,(x,y) \in \mathcal{S}_C$. Second, by definition, for the optimal policy π_{θ^*}, we have the threshold equal to the continuation value[2]: $\theta^*_x = c_{\pi^*}(x)$. To show gradient dominance, we start with the left hand side in (B.15),

$$\begin{aligned}
\ell(\theta^*) - \ell(\theta) &= (1-\gamma) \sum_{x \in \mathcal{X}} \nu(x) \left[\int_{y \in \mathcal{Y}} \left(J_{\pi^*}(x,y) - J_{\pi_\theta}(x,y) \right) q_x(y) dy \right] \\
&= (1-\gamma) \| J_{\pi^*} - J_{\pi_\theta} \|_{1,\rho} \\
&\leq \kappa_\rho \| T J_{\pi_\theta} - J_{\pi_\theta} \|_{1,\rho} \qquad \text{(B.16)}
\end{aligned}$$

where the last inequality follows from the definition of the concentrability coefficient. Recall the policy iteration step for optimal stopping as shown in Example 4. For any $\pi_\theta \in \Pi_\Theta$, the improved policy accepts offers above the continuation value $c_{\pi_\theta}(x)$

$$\pi^+_\theta(x,y) = \begin{cases} 1 & \text{if } y < c_{\pi_\theta}(x), \\ 0 & \text{if } y \geq c_{\pi_\theta}(x) \end{cases}$$

Therefore,

$$\begin{aligned}
J_\theta(x,y) &= \mathbb{1}\left(y \geq \theta_x\right) \cdot y + \mathbb{1}\left(y < \theta_x\right) \cdot c_{\pi_\theta}(x) \\
T J_\theta(x,y) &= \mathbb{1}\left(y \geq c_{\pi_\theta}(x)\right) \cdot y + \mathbb{1}\left(y < c_{\pi_\theta}(x)\right) \cdot c_{\pi_\theta}(x)
\end{aligned}$$

which implies,

$$\int_{\mathcal{Y}} \left(T J_\theta(x,y) - J_\theta(x,y) \right) q_x(y) dy = \int_{c_{\pi_\theta}(x)}^{\theta_x} y \cdot q_x(y) dy + \int_{\theta_x}^{c_{\pi_\theta}(x)} c_{\pi_\theta}(x) \cdot q_x(y) dy \qquad \text{(B.17)}$$

[2]It is easy to see that for our problem setting with positive offers, i.e. $\mathcal{Y} = [y_{\min}, y_{\max}]$ with $y_{\min} > 0$, setting $\theta^* = y_{\min}$ is never optimal as the offer distribution, $q_x(y)$ is assumed to be a density supported over $\mathcal{Y}$. An optimal threshold must correspond to a stationary policy. Hence, by our arguments in Example 4, it must hold that $\theta^*_x = c_{\pi^*}(x)$.

Case (1): We first assume that $\theta_x < \theta_x^*$. As we can improve the total cost by moving towards θ^* (i.e. by increasing the value of θ_x), it must be true that $\frac{\partial}{\partial \theta_x}\ell(\theta) > 0$. Using the gradient expression in (3.14) along with the policy gradient lemma implies that $c_{\pi_\theta}(x) > \theta_x$. Then, we get

$$\begin{aligned}\int_{\mathcal{Y}} \left(TJ_\theta(x,y) - J_\theta(x,y)\right) q_x(y)dy &= \int_{\theta_x}^{c_{\pi_\theta}(x)} c_{\pi_\theta}(x) \cdot q_x(y)dy - \int_{\theta_x}^{c_{\pi_\theta}(x)} y \cdot q_x(y)dy \\ &\leq \int_{\theta_x}^{c_{\pi_\theta}(x)} \left(c_{\pi_\theta}(x) - \theta_x\right) \cdot q_x(y)dy \\ &\leq m(x) \left(c_{\pi_\theta}(x) - \theta_x\right)^2\end{aligned}$$

where $m(x) = \max_{y\in\mathcal{Y}} q_x(y) < \infty$, as $q_x(y)$ is assumed to be a density over the bounded set $\mathcal{Y}$. Therefore,

$$\begin{aligned}\|TJ_\theta - J_\theta\|_{1,\rho} = \sum_{x\in\mathcal{X}} \nu(x) \int_{\mathcal{Y}} \left(TJ_\theta(x,y) - J_\theta(x,y)\right) q_x(y)dy &\leq \beta \sum_{x\in\mathcal{X}} \left(c_{\pi_\theta}(x) - \theta_x\right)^2 \\ &\leq \beta \sum_{x\in\mathcal{X}} \left(c_{\pi_\theta}(x) - \theta_x\right)\left(c_{\pi^*} - \theta_x\right)\end{aligned} \tag{B.18}$$

where we define $\beta = \max_{(x,y)\in\mathcal{S}_C} \nu(x)\, q_x(y)$ and the final inequality uses that $c_{\pi_\theta} < c_{\pi^*}$ as argued above. Policy gradient lemma and (3.14) also imply

$$\begin{aligned}\max_{\theta'\in\Theta} \nabla\ell(\theta)^\top(\theta' - \theta) &= \sum_{x\in\mathcal{X}} \left[\max_{\theta'_x\in\mathcal{Y}} \left\{\left(\theta'_x - \theta_x\right) \cdot \left(c_{\pi_\theta}(x) - \theta_x\right) \eta'_\theta(x) q_x(\theta_x)\right\}\right] \\ &\geq \sum_{x\in\mathcal{X}} \left(c_{\pi_\theta}(x) - \theta_x\right)\left(c_{\pi^*} - \theta_x\right) \eta'_\theta(x) q_x(\theta_x) \\ &\geq \min_{x\in\mathcal{X}} \left[\eta'_\theta(x) q_x(\theta_x)\right] \cdot \sum_{x\in\mathcal{X}} \left(c_{\pi_\theta}(x) - \theta_x\right)\left(c_{\pi^*} - \theta_x\right) \\ &\geq \underbrace{\min_{(x,y)\in\mathcal{S}_C} \nu(x)\, q_x(y)}_{=\psi} \sum_{x\in\mathcal{X}} \left(c_{\pi_\theta}(x) - \theta_x\right)\left(c_{\pi^*} - \theta_x\right)\end{aligned} \tag{B.19}$$

Combining equations (B.15), (B.16), (B.18) and (B.19), we conclude that $\ell(\cdot)$ is $(c, 0)$-gradient dominated for $c = \frac{\beta}{\kappa_\rho \psi}$.

Case (2): For $\theta_x < \theta_x^*$, a similar argument as given for case (1) shows that $c_{\pi_\theta}(x) < \theta_x$ as we can improve the total cost by moving towards θ^*, i.e. by decreasing the value of θ_x. Therefore, we get

$$\begin{aligned}\int_{\mathcal{Y}} \left(TJ_\theta(x,y) - J_\theta(x,y)\right) q_x(y)dy &= \int_{c_{\pi_\theta}(x)}^{\theta_x} y \cdot q_x(y)dy - \int_{c_{\pi_\theta}(x)}^{\theta_x} c_{\pi_\theta}(x) \cdot q_x(y)dy \\ &\leq \int_{c_{\pi_\theta}(x)}^{\theta_x} \left(\theta_x - c_{\pi_\theta}(x)\right) \cdot q_x(y)dy \\ &\leq m(x) \left(c_{\pi_\theta}(x) - \theta_x\right)^2\end{aligned}$$

as before. Following exactly the same steps as for case (1) confirms our result. □

Smoothness. Next, we show that the policy gradient objective $\ell(\cdot)$ for the optimal stopping problem is smooth over the parameter space, $\Theta = [y_{\min}, y_{\max}]^{|\mathcal{X}|}$ by showing that $\max_{\theta\in\Theta} \|\nabla^2 \ell(\theta)\| < \infty$.

Lemma 29 (Cost function for optimal stopping)**.** *In the optimal stopping problem in Example 4,* $\max_{\theta\in\Theta} \|\nabla^2 \ell(\theta)\| < \infty$.

Proof. Using the policy gradient theorem as shown in Lemma 20 and the derivative calculations above, we have

$$\frac{\partial}{\partial \theta_x} \ell(\theta) = \frac{\partial}{\partial \overline{\theta}_x} \mathcal{B}(\overline{\theta} | \eta_\theta, J_\theta)\Big|_{\overline{\theta}=\theta} = \left(c_{\pi_\theta}(x) - \theta_x\right) \eta'_\theta(x) q_x(\theta_x).$$

We showed above that η'_θ is continuously differentiable, and we have that $q_x(\theta_x)$ is continuously differentiable in θ_x by assumption. Therefore, the gradient $\nabla\ell(\theta)$ has a continuous derivative if the continuation value $c_{\pi_\theta}(x)$ is continuously differentiable in θ. To show this, recall that by

definition[3],

$$c_{\pi_\theta}(x) = \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int J_\theta(x', y') q_x(y') dy = \gamma \sum_{x' \in \mathcal{X}} p(x'|x) J'_\theta(x) \tag{B.20}$$

where we define $J'_\theta(x) := \int J_\theta(x', y') q_x(y') dy$ to be the expected cost-to-go function from context x. Let us also define $g'_\theta(x) := \int_{\mathcal{Y}} \mathbb{1}(y \geq \theta_x)\, y q_x(y) dy$ as the expected reward earned from context x. Clearly, g'_θ is continuously differentiable. Then, by definition, for all $x \in \mathcal{X}$

$$J'_\theta(x) = \eta'_\theta(x) g'_\theta(x).$$

As both η'_θ and g'_θ are continuously differentiable, so is J'_θ and hence using (B.20) we find that c_{π_θ} is continuously differentiable in θ. Therefore, $\nabla^2 \ell(\theta)$ exists and is continuous. Since Θ is compact, the extreme value theorem ensures $\max_{\theta \in \Theta} \|\nabla^2 \ell(\theta)\|$ exists and is finite. □

Concentrability coefficient.

Lemma 25 (Concentrability in optimal stopping)**.** *For the optimal stopping problem in Example 4 consider the policy π_C that never stops ($\pi_C(s) = 1$ for each $s \in \mathcal{S}_C$) and suppose the induced Markov process has stationary distribution $\mu = \mu P_{\pi_C}$. Then, for the choice $\rho = \mu$, $\kappa_\rho \leq 1$.*

Proof. We show that the Bellman operator T is a contraction with modulus γ in $\|\cdot\|_{1,\mu}$. The proof then follows immediately using Lemma 24.

As we consider a policy that never stops, the stationary distribution over continuation states, $(x, y) \in \mathcal{S}_{\mathrm{C}}$ factorizes as $\mu(x, y) = \mu'(x) q_x(y)$ where μ' is the marginal stationary distribution over context states $\mathcal{X}$ such that $\mu'(x') = \sum_{x \in \mathcal{X}} \mu'(x) p(x'|x)$. Then,

$$\|TJ - TJ'\|_{1,\mu} = \sum_{x \in \mathcal{X}} \mu'(x) \left| \int_{\mathcal{Y}} \left(TJ(x, y) - TJ'(x, y)\right) q_x(y) dy \right|$$

[3]Clearly, the terminal state $\{T\}$ has zero continuation value, $c_{\pi_\theta}(T) = 0$ for any policy $\pi_\theta \in \Pi_\Theta$.

By definition,

$$TJ(x,y) = \max\{y, \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J(x',y') q_{x'}(y') dy'\}$$

Note that for any scalars (x_1, x_2, y), we have $|\max\{y, x_1\} - \max\{y, x_2\}| \leq |x_1 - x_2|$. Therefore,

$$\begin{aligned} |TJ(x,y) - TJ'(x,y)| &\leq \gamma \left| \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} \left(J(x',y') - J'(x',y')\right) q_{x'}(y') dy' \right| \\ &\leq \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} |J(x',y') - J'(x',y')| \, q_{x'}(y') dy' \end{aligned} \tag{B.21}$$

As the right hand side in (B.21) is independent of y, clearly

$$\int_{\mathcal{Y}} |TJ(x,y) - TJ'(x,y)| \, q_x(y) dy \leq \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} |J(x',y') - J'(x',y')| \, q_{x'}(y') dy'$$

Thus,

$$\begin{aligned} \|TJ - TJ'\|_{1,\mu} &= \sum_{x \in \mathcal{X}} \mu'(x) \int_{\mathcal{Y}} |TJ(x,y) - TJ'(x,y)| \, q_x(y) dy \\ &\leq \gamma \sum_{x \in \mathcal{X}'} \mu'(x) \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} |J(x',y') - J'(x',y')| \, q_{x'}(y') dy' \\ &\overset{(a)}{=} \gamma \sum_{x' \in \mathcal{X}'} \mu'(x') \int_{\mathcal{Y}} |J(x',y') - J'(x',y')| \, q_{x'}(y') dy' \\ &= \gamma \|J - J'\|_{1,\mu} \end{aligned}$$

where (a) follows as μ' is the stationary distribution over $\mathcal{X}$. For $\rho = \mu$, we have $C, c = 1$ in Lemma 24 implying $\kappa_\rho \leq 1$. □

B.6 Details for finite horizon inventory control

Recall the setting in Example 5 where the inventory level evolves as: $s_{t+1} = s_t + a_t - w_t \; \forall \, t = \{0, \ldots, H-1\}$ for non-negative inventory orders a_t and i.i.d demands $w_t \in [0, w_{\max}]$. We consider

the class of base-stock-policies parameterized as $\Pi_\Theta = \{\theta = (\theta_0, \ldots, \theta_{H-1}) \in \mathbb{R}^H : \theta_t > 0\}$ which orders inventory $\pi_\theta(s_t) = \max\{0, \theta_t - s_t\}$ at time t. We treat finite-horizon time-inhomogenous MDPs as a special case of infinite horizon MDP using the following factorization.

Condition 3. *Suppose the state space factors as $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H \cup \mathcal{S}_{H+1}$, where for a state $s \in \mathcal{S}_h$ with $h \leq H$, $\sum_{s' \in \mathcal{S}_{h+1}} P(s'|s, a) = 1$ for all $a \in \mathcal{A}_s$. The final subset $\mathcal{S}_{H+1} = \{\tau\}$ contains a single cost-less absorbing state, with $P(\tau|\tau, a) = 1$ and $g(\tau, a) = 0$ for any action a. The parameter space is the product set $\Theta = \Theta_1 \times \cdots \times \Theta_H$, where a policy parameter $\theta = (\theta_1, \ldots, \theta_H) \in \Theta$ is the concatenation of H sub-vectors.*

Recall that for inventory control problem in Example 5, $\mathcal{S}_h = [-M, M]$ as well as $\Theta_h = [0, M]$ are bounded intervals of $\mathbb{R}$ for all $h \in \{0, 1, \ldots, H-1\}$.

Differentiability. We first verify Condition 0 required for the policy gradient theorem to hold. Throughout, we will appeal to the Leibniz rule which states conditions for differentiating an integral. Let $\Theta', \mathcal{S}' \subset \mathbb{R}$ be bounded intervals. Consider the function $F : \Theta' \mapsto \mathbb{R}$ given by,

$$F(\theta) = \mathbb{E}[f(\theta, s)] = \int_{\mathcal{S}'} f(\theta, s) P(ds)$$

where $\mathcal{S}$ is the state space, $f : \Theta' \times \mathcal{S}' \mapsto \mathbb{R}$ is a real valued function, P is a probability measure supported over $\mathcal{S}'$ and for each $\theta \in \Theta'$, the function $f(\theta, \cdot)$ is P-integrable, i.e. $\mathbb{E}[|f(\theta, s)|] < \infty$. Let $\mathcal{D}_s(\theta)$ be the set of points $s \in \mathcal{S}'$ such that $f(\cdot, s)$ is non-differentiable at θ. By Leibniz rule, F is differentiable at θ if (i) $\mathcal{D}_s(\theta)$ has zero measure under P, i.e. $P[\mathcal{D}_s(\theta)] = 0$ and (ii) $f'(\theta, \cdot)$ is P-integrable. This is useful in context of the inventory control problem as a threshold policy, $\pi_\theta(s)$, for a fixed θ is differentiable everywhere except at $s = \theta$.

To verify Condition 0, we first argue differentiability of η_{π_θ}. Note that for any $h \leq H - 1$, state $s_h \in \mathcal{S}_h$ and a measurable set $A \in \mathcal{S}_{h+1}$, the probability transition operator under π_θ, $P_{\theta_h}(s_h, A) = \mathbb{E}_{w_h}[\mathbb{1}(s_h + \pi_\theta(s_h) - w_h \in A)]$ is clearly differentiable for any $s_h \neq \theta_h$ and $|\partial / \partial \theta_h \pi_\theta(s_h)| \leq$

$1 \; \forall \, s_h \neq \theta_h$. Therefore, the left linear operator

$$\lambda P_{\theta_h}(A) = \int_{S_h} P_{\theta_h}(s_h, A)\lambda(s_h)ds_h$$

is differentiable for any $\theta_h \in \Theta_h$ for assuming λ is a probability density supported over $\mathcal{S}_h$. A similar argument holds for the t-step counterparts, $\lambda P^t_{\theta_h}(A)$ for all $h \leq H-1$, which are defined as the product of transition kernels. Hence, for any $A \in \mathcal{S}_h$, the state-occupancy measure $\eta_\theta(A) = \sum_{i=0}^{h} \rho P^{h-i}_{\theta_i}(A)$ is differentiable in θ. It is noteworthy to see that η_θ is supported over each $\mathcal{S}_h$ for $h \leq H-1$ as the initial distribution ρ.

Next we argue that $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta}|\eta_{\pi_\theta}, J_\theta)$ is a continuously differentiable function of $\overline{\theta}$. For any $\theta, \overline{\theta} \in \Theta$, consider the Bellman update $T_{\overline{\theta}}J_\theta$. For any $s_h \in \mathcal{S}_h, h \leq H-1$, we use expression for the cost-to-go function as shown in (3.18) to write,

$$T_{\overline{\theta}}J_\theta(s_h) = \left[c\pi_{\overline{\theta}}(s_h) + \sum_{i=h+1}^{H} g(s_i, a_i) \right]$$

where for $i = \{h+1, \ldots, H-1\}$, the per step cost function is denoted by $g(s_i, a_i) := ca_i + \mathbb{E}_{w_i}[r(s_i + a_i - w_i)]$ for inventory orders $a_i = \pi_\theta(s_i)$ and holding/backlogging costs $r : \mathbb{R} \to \mathbb{R}$ defined as $r(x) = b\max(0, x) + p\max(0, -x)$. Note that we do not allows for inventory orders in the final period, i.e. $a_H = 0$. Using the above, we have

$$\frac{\partial}{\partial \overline{\theta}_h} T_{\overline{\theta}}J_\theta(s_h) = \left[c \; \frac{\partial}{\partial \overline{\theta}_h}\pi_{\overline{\theta}}(s_h) + \sum_{i=h+1}^{H} \frac{\partial}{\partial \overline{\theta}_h} g(s_i, a_i) \right]$$

Note that,

$$\frac{\partial}{\partial \overline{\theta}_h} g(s_i, a_i) = \frac{\partial}{\partial \overline{\theta}_h}\mathbb{E}_{w_i}[r(s_i + a_i - w_i)] = \frac{\partial}{\partial y_i}\mathbb{E}_{w_i}[r(y_i - w_i)]\Big|_{y_i = s_i + a_i} \frac{\partial s_i}{\partial \overline{\theta}_h}$$

As $s_i = s_h + a_h - w_h + \sum_{k=h+1}^{i-1}(a_k - w_k)$ with $a_h = \pi_{\overline{\theta}}(s_h)$, we have s_i to be a differential function of $\overline{\theta}_h$ except for $s_h = \overline{\theta}_h$. Similarly, $r(\cdot)$ is a differentiable function of y_i except at $y_i = w_i$ and by

definition, $|\partial/\partial y_i r(y_i - w_i)| \leq \max\{b, p\}$. As we assumed the demand distribution has a density over $[0, w_{\max}]$, we conclude that $T_{\overline{\theta}} J_\theta$ is a differentiable function of $\overline{\theta}$ except at finite number of points, $\mathcal{D}_s(\theta) = \{s_h \in \mathcal{S}_h : s = \theta_h \ \forall\, h \leq H-1\}$. As argued above, η_{π_θ} is supported over each $\mathcal{S}_h$ for $h \leq H-1$ which implies $\mathcal{B}(\overline{\theta}|\eta_{\pi_\theta}, J_\theta) = \int_{\mathcal{S}} T_{\overline{\theta}} J_\theta(s) \eta_{\pi_\theta}(s) ds$ is differentiable for all $\theta \in \Theta$.

Condition 4 for inventory control. The following lemma shows how Condition 4 holds for the finite horizon inventory control problem.

Lemma 39. *Consider the finite horizon inventory control problem in Example 5. Let J^* be the cost-to-go function corresponding to the optimal policy. Then, for any $\pi, \pi_\theta \in \Pi_\Theta$, the weighted policy iteration objective $\mathcal{B}(\theta|\eta_\pi, J^*)$ has no suboptimal stationary points.*

Proof. Let $Q^*(s,a)$ be the Q-function corresponding to the optimal policy. It can be easily shown that $Q^*(s,a)$ is strictly convex in a which follows as the optimal cost-to go function, $J^*(\cdot)$ is convex and the per step costs (of ordering and holding/backlogging) are strictly convex, see results in chapter 3 of [101]. We want to show that for any stationary point $\theta \in \Theta$, the corresponding base-stock policy π_θ is optimal. Note that $\pi_\theta(s_h) = \theta_h - s_h$ for all $s_h < \theta_h$ and $\pi_\theta(s_h) = 0$ for all $s_h > \theta_h$. By Leibniz rule and the fact that η_π is supported over $\mathcal{S}_h$ for each $h \leq H-1$ (i.e. its not a delta function),

$$\begin{aligned}
\frac{\partial}{\partial \theta_h} \mathcal{B}(\theta|\eta_\pi, J^*) &= \frac{\partial}{\partial \theta_h} \int_{s_h \in \mathcal{S}_h} Q^*(s_h, \pi_\theta(s_h))\, \eta_\pi(s_h)\, ds_h \\
&= \frac{\partial}{\partial \theta_h} \left[\int_{s_h < \theta_h} Q^*(s_h, \theta_h - s_h)\, \eta_\pi(s_h)\, ds_h + \int_{s_h > \theta_h} Q^*(s_h, 0)\, \eta_\pi(s_h)\, ds_h \right] \\
&= \int_{s_h < \theta_h} \frac{\partial}{\partial \theta_h} Q^*(s_h, \theta_h - s_h)\, \eta_\pi(s_h)\, ds_h \\
&= \int_{s_h < \theta_h} \frac{\partial}{\partial a} Q^*(s_h, a)\Big|_{a = \pi_\theta(s_h)} \eta_\pi(s_h)\, ds_h
\end{aligned}$$

Consider any θ such that $\theta_h \neq \theta_h^*$. Let $\theta_h^\alpha = \theta_h + \alpha(\theta_h^* - \theta_h)$. Then, $\frac{d}{d\alpha} \mathcal{B}(\theta_h^\alpha|\eta_\pi, J^*)|_{\alpha=0} < 0$. This follows as $Q^*(s_h, a)$ is strictly convex with a minimum at $a = \pi_{\theta^*}(s_h)$ and therefore we can reduce cost by moving θ_h in the direction of θ_h^* (which is feasible as Θ_h is a bounded interval of

$\mathbb{R}$). Therefore, θ cannot be a stationary point. □

B.7 Miscellaneous Proofs

Proof of Theorem 6. For the reader's convenience, we restate Theorem 6.

Theorem 6. *Suppose Conditions 3 and 4 hold. If the parameterized policy class Π_Θ contains an optimal policy π^*, then any stationary point θ of $\ell : \Theta \to \mathbb{R}$ satisfies $J_{\pi_\theta} = J^*$.*

Proof. For simplicity, assume there is a unique optimal policy, π^*. For any $h \in \{1, \cdots, H\}$ and $\theta_h \in \Theta_h$, let $\pi_{\theta_h} : \mathcal{S}_h \to \mathcal{A}$ denote the policy for period h. Similarly, $\pi_{\theta_h^*}$ denotes the optimal policy for period h. Let θ is a stationary point of $\ell(\cdot)$. The product structure of the policy class, $\Theta = \Theta_1 \times \cdots \Theta_H$ implies that for all $h \leq H$,

$$\left\langle \frac{\partial}{\partial \overline{\theta}_h} \mathcal{B}(\overline{\theta} | \eta_{\pi_\theta}, J_{\pi_\theta}) \Big|_{\overline{\theta}=\theta}, \theta'_h - \theta_h \right\rangle \geq 0 \quad \forall\, \theta'_h \in \Theta_h$$

using the policy gradient theorem in Lemma 20. By definition, $J_{\pi_\theta}(s) = J^*(s) = 0$ for $s \in \mathcal{S}_{H+1}$ (as $\mathcal{S}_{H+1}$ contains a single, cost-less absorbing state). Our argument follows by backward induction.

Base Case: We first show that $J_{\pi_\theta}(s) = J^*(s)$ for $s \in \mathcal{S}_H$. To see this, note that for any $s \in \mathcal{S}_H$ and action a, we have $Q_{\pi_\theta}(s, a) = Q^*(s, a) = g(s, a)$. This is because $J_{\pi_\theta}(\tau) = J^*(\tau) = 0$. Therefore,

$$\left\langle \frac{\partial}{\partial \overline{\theta}_H} \mathcal{B}(\overline{\theta} | \eta_{\pi_\theta}, J^*) \Big|_{\overline{\theta}=\theta}, \theta'_H - \theta_H \right\rangle \geq 0 \quad \forall\, \theta'_H \in \Theta_H$$

Hence, for $S \sim \eta_{\pi_\theta}$ such that $S \in \mathcal{S}_H$

$$\mathbb{E}[J_{\pi_\theta}(S)] = \min_{\overline{\theta}_H \in \Theta_H} \mathbb{E}[Q^*(S, \pi_{\overline{\theta}_H}(S))] = \mathbb{E}[Q^*(S, \pi_{\theta_H^*}(S))] = \mathbb{E}[J^*(S)]$$

where the first equality follows by assumption that $\overline{\theta} \to \mathcal{B}(\overline{\theta} | \eta_{\pi_\theta}, J^*)$ has no suboptimal stationary points and the second equality uses the assumption that policy class Π_Θ contains the optimal policy. As we assumed that ρ is a density supported over $\mathcal{S}_H$, our desired result follows.

Induction step: We now show that if $J_{\pi_\theta}(s) = J^*(s)\ \forall\, s \in \mathcal{S}_{h+1}$ for any $h < H$, then $J_{\pi_\theta}(s) = J^*(s)$ for all $s \in \mathcal{S}_h$. By the definition, for any state $s \in \mathcal{S}_h$ and action a,

$$\begin{aligned} Q_{\pi_\theta}(s,a) &= g(s,a) + \gamma \sum_{s' \in \mathcal{S}_{h+1}} P(s'|s,a) J_{\pi_\theta}(s') \\ &= g(s,a) + \gamma \sum_{s' \in \mathcal{S}_{h+1}} P(s'|s,a) J^*(s') \\ &= Q^*(s,a). \end{aligned}$$

As θ is a stationary point,

$$\begin{aligned} &\left\langle \frac{\partial}{\partial \overline{\theta}_h} \mathcal{B}(\overline{\theta}|\eta_{\pi_\theta}, J_{\pi_\theta})\Big|_{\overline{\theta}=\theta}, \theta'_h - \theta_h \right\rangle \geq 0 \quad \forall\, \theta_h \in \Theta_h \\ \implies &\left\langle \frac{\partial}{\partial \overline{\theta}_h} \mathcal{B}(\overline{\theta}|\eta_{\pi_\theta}, J^*)\Big|_{\overline{\theta}=\theta}, \theta'_h - \theta_h \right\rangle \geq 0 \quad \forall\, \theta_h \in \Theta_h \end{aligned}$$

By exactly the same argument as above,

$$\mathbb{E}[J_{\pi_\theta}(S)] = \min_{\overline{\theta}_h \in \Theta_h} \mathbb{E}[Q^*(S, \pi_{\overline{\theta}_h}(S))] = \mathbb{E}[Q^*(S, \pi_{\theta^*_h}(S))] = \mathbb{E}[J^*(S)]$$

for any $S \sim \eta_{\pi_\theta}$ such that $S \in \mathcal{S}_h$. Our result follows by again noting that ρ is assumed to be a density supported over $\mathcal{S}_h$. □

Concentrability coefficients.

Lemma 23. *Let π^* denote any optimal stationary policy. Then,*

$$\kappa_\rho \leq \sup_{s \in \mathcal{S}} \frac{\eta_{\pi^*}(s)}{\rho(s)}$$

Proof. Fix some $J_\pi \in J_\Theta$, where dependence on π is there to make transparent that this must be a cost-to-go function for some policy $\pi \in \Pi_\Theta$. Denotes $J^* = J_{\pi^*}$, Then, the variational form of

Bellman's inequality in (3.5) gives

$$J_\pi - J^* = (I - \gamma P_{\pi^*})^{-1} (J_\pi - T_{\pi^*} J_\pi) \leq (I - \gamma P_{\pi^*})^{-1} (J_\pi - T J_\pi)$$

Left multiplying by ρ, using the definition $\eta_{\pi^*} = (1-\gamma)\rho(I - \gamma P_{\pi^*})^{-1}$ (see Equation (3.6) in Section 3.3) and that $J_\pi - TJ_\pi \geq 0$ and $J_\pi \geq J^*$ gives us the desired result.

$$\begin{aligned} \|J_\pi - J^*\|_{1,\rho} = \rho (J_\pi - J^*) \leq \frac{1}{(1-\gamma)} \eta_{\pi^*} (J_\pi - TJ_\pi) &\leq \frac{\left(\sup_{s \in \mathcal{S}} \frac{\eta_{\pi^*}(s)}{\rho(s)}\right)}{(1-\gamma)} \rho (J_\pi - TJ_\pi) \\ &= \frac{\left(\sup_{s \in \mathcal{S}} \frac{\eta_{\pi^*}(s)}{\rho(s)}\right)}{(1-\gamma)} \|J_\pi - TJ_\pi\|_{1,\rho}. \end{aligned}$$

□

B.8 An example of state aggregation

State aggregation is the simplest form of value function approximation employed in reinforcement learning and comes with strong stability properties [162, 163, 164]. It is common across several academic communities [e.g 165, 166]. Numerous theoretical papers carefully construct classes of MDPs with sufficient smooth dynamics, and upper bound the error from planning on a discretized state space [e.g 167]. The following example considers a continuous state, finite action problem which reduces to the tabular MDP case (in Example 3) with state aggregation. It is not unreasonable to expect that an appropriate partitioning of the state space results in the policy class (class of stochastic policies over finite aggregated states) being approximately closed under policy improvement.

Example 7 (State aggregation)**.** *We consider a problem with finite number of deterministic actions k and take $\mathcal{A} = \Delta^{k-1}$ to be the set of probability distributions over actions. Let the state space, $\mathcal{S} \subset \mathbb{R}^n$, be a bounded convex subset of euclidean space where the dimension n is thought to be small. We consider a partition of the state space into m disjoint subsets, $\mathcal{S} = \cup_{i=1}^m \mathcal{S}_i$ and the set of*

stochastic policies over these subsets $\Pi = \{\pi \in \mathbb{R}_+^{m \times k} : \sum_{i=1}^{k} \pi(\mathcal{S}_j, i) = 1 \ \forall j = \{1, \cdots, m\}\}$ *such that* $\pi(s, i) = \pi(\mathcal{S}_j, i) \ \forall \ s \in \mathcal{S}_j$. *Our result applies by assuming the partition is effective such that the approximation error in the policy iteration update,* $\inf_{\pi' \in \Pi} \|T_{\pi'} J_\pi - T J_\pi\|_{1, \eta_\pi}$, *is small for any policy* $\pi \in \Pi$.