Global Optimality Guarantees For Policy Gradient Methods

Jalaj Bhandari and Daniel Russo

Columbia University

Abstract

Policy gradients methods are perhaps the most widely used class of reinforcement learning algorithms. These methods apply to complex, poorly understood, control problems by performing stochastic gradient descent over a parameterized class of polices. 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 local minima. This work identifies structural properties – shared by finite MDPs and several classic control problems – which guarantee that policy gradient objective function has no suboptimal local minima despite being non-convex.

Warning. This is a rough draft that we need to finish polishing. Some sections were written hastily. We are sharing it now to provide some details with those who have attended one of our talks.

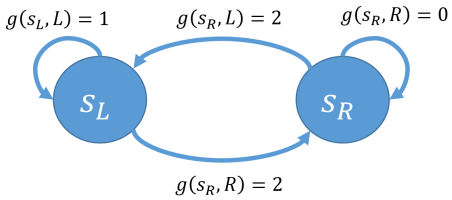
1 Introduction

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 $\{\pi_\theta\}_{\theta \in \mathbb{R}^d}$ by perform stochastic gradient descent on $\ell(\cdot)$, following the iteration

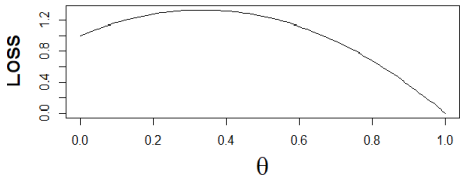
$$\theta_{k+1} = \theta_k - \alpha_k \left(\boldsymbol{\nabla}_\theta \ell(\pi_{\theta_k}) + \text{noise} \right).$$

This approach has several attractive features that have driven their popularity. It is end-to-end, directly optimizing the true decision objective rather than searching for approximate models or value functions that minimize prediction error. It appears to offer advantages when the designer has an inductive bias about the form of policy that might be effective, rather than the form of an approximate model or value function. Finally, it makes only small changes to the policy in each iteration, so some believe it is more stable than approximate policy iteration schemes.

Unfortunately, while policy gradient methods can be applied to a very broad class of problems, it is not clear whether they adequately address even for simple control problems solvable by classical methods. The challenge is that total cost ℓ is a non-convex function of the chosen policy. Typical of results concerning the black-box optimization of non-convex functions, policy gradient methods are widely understood to converge asymptotically to a stationary point or a local minimum. Important theory guarantees this under technical conditions [Baxter and Bartlett, 2001, Marbach and Tsitsiklis, 2001, Sutton et al., 2000] and it is widely repeated in textbooks and surveys [Grondman et al., 2012, Peters and Schaal, 2006, Sutton and Barto, 2018]. But the reinforcement learning literature seems to provide almost no guarantees into the *quality* of the points to which policy gradient methods



(a) Two state MDP



(b) Loss $\ell(\pi_\theta)$ is nonconvex with two local minima.

Figure 1: Policy gradient fails with the simple policy class $\pi_\theta(R|S_L) = \pi_\theta(R|S_L) = \theta \in [0, 1]$.

converge. Worse yet, Example 1 shows that policy gradient methods could get stuck in bad local minima even for very simple examples where the policy class contains the optimal policy.

Example 1 (Failure of policy gradient with a correct policy class)**.** *Consider the MDP depicted in Figure 1. There are two states – left (S_L) and right (S_R) – and two 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 costless ($g(S_R, R) = 0$). Moving between states incurs a per-period cost of 2. As long as the discount factor exceeds $1/2$, the optimal policy plays the action R in either state. This behavior can be encoded by a simple parameterized policy π_θ which plays the action R with probability $\theta \in [0, 1]$ regardless of the current state. Unfortunately, the total discounted cost incurred is a nonconvex function of θ. This is depicted in Figure 1. When initialized with small value of θ, cost is locally increasing as a function of θ, and so a gradient method moves the policy toward a bad local minimum at $\theta = 0$. At this local minimum, the algorithm always plays left, and any chance of moving to the right only increases expected costs because the algorithm is likely to move back to the left immediately thereafter. It is worth noting here that policy gradient methods also face challenges due to unsophisticated exploration or non-linear policy parameterization. This example instead highlights the challenges presented by the non-convexity of ℓ.*

In marked contrast to the example above, important recent work of Fazel et al. [2018] showed that policy gradient on the space of linear policies for deterministic linear quadratic control problem converges to the global optimum, despite the non-convexity of the objective. The authors provided an intricate analysis in this case, leveraging a variety of closed form expressions available for linear-quadratic problems. Separate from the RL literature, Kunnumkal and Topaloglu [2008] propose a stochastic approximation method for setting base-stock levels in inventory control. Surprisingly, despite non-convexity of the objective, an intricate analysis quite different that from Fazel et al. [2018] establishes convergence to the global optimum. How do we reconcile these success stories with the simple counterexample given in Example 1?

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[1] of the policy gradient cost function is a global optimum. The examples we treat include:

1. Finite state and action MDPs combined with the set of all possible stochastic policies.
2. Linear quadratic control problems combined with the set of linear policies.
3. Optimal stopping problems combined with the set of threshold policies.
4. Finite horizon inventory control problems combined with the set of base-stock policies.

Our work aims to understand this phenomenon. Why does gradient descent on a non-convex function reach the global minimum? Why in these examples but not in Example 1?

These examples share 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 a nasty non-convex function of the policy, the policy iteration update involves just solving a quadratic minimization problem. Given this insight, strikingly simple proofs show that any stationary point of the cost function $\ell(\pi_\theta)$ is a global minimum. These same arguments extend beyond linear quadratic control. In fact, for 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 is easily seen to have no suboptimal stationary points. (The fourth example involves a class of non-stationary policies and it is shown in Theorem 2 that slightly weaker conditions are needed in that case.)

We generalize these results to the case where the policy is closed under approximate policy improvement — meaning that a certain weighted policy iteration problem can be solved to within a given tolerance by some policy in the policy class. In that case, our results bounds the optimality gap of any stationary point of ℓ. We can interpret this closure assumption as a requirement that the policy class is sufficiently rich. Crucially, however, this assumption is much weaker than the requirement that the policy class contains (nearly) all possible policies. Instead, the closure property captures examples like those above, where a strong inductive bias allows us to choose policy classes well suited to the objective, even if the policy class has limited expressivity. However, the closure property is stronger than a requirement that the policy class contains (near) optimal policies. Indeed Example 1 shows that is necessary. In that case the policy class contains the optimal policy, but it is not closed under policy improvement and we saw that policy gradient methods could get stuck in bad local minima.

Beyond studying the quality of stationary points, we also study stronger properties of the loss function's landscape that lead to converge rates. In particular, we study an inequality known either as gradient dominance or as a Polyak-lojasiewicz condition [Karimi et al., 2016, Nesterov and Polyak, 2006, Polyak, 1963], which is a relaxation of convexity or strong convexity that still guarantees fast convergence rates for many first-order optimization algorithms. We show that when the the policy class is closed under policy improvement and the a weighted policy iteration problem satisfies the gradient dominance condition, then the policy gradient loss function also satisfies this gradient dominance condition.

[1] In unconstrained optimization, stationary points of a function f are those satisfying $\nabla f(x) = 0$. More generally, in constrained optimization over a set $\mathcal{X}$ they are the points x satisfying the first order necessary conditions for optimality $\nabla f(x)^\top (x' - x) \geq 0 \ \forall x'$.

Finally, in Section 9, we give a specialized analysis convergence rates in finite-state and action MDPs where the policy class contains all stochastic policies. This establishes a geometric convergence rate for several first-order optimization algorithms. We remark that our results also apply to the case of finite horizon problems with non-stationary policy classes by simply using the fact that the policy class contains the optimal policy.

Scope of this work. There are many reasons why practitioners may find simple policy gradient methods, like the classic REINFORCE algorithm Williams [1992], 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 [Amari, 1998, Kakade, 2002], which perform steepest descent in a different metric. The issue can also be alleviated with regularized policy gradient algorithms [Schulman et al., 2015a, 2017].

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. Kakade and Langford [2002] provide a striking example. Combining efficient exploration methods with policy gradients algorithms is challenging, but is an active area of research [see e.g. Nachum et al., 2017, Plappert et al., 2017].

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 [Konda and Tsitsiklis, 2000, Marbach and Tsitsiklis, 2001, Sutton et al., 2000] and appropriate baselines [Mnih et al., 2016, Schulman et al., 2015b].

We emphasize that this paper is focused on the first challenge and on understanding the risks posed by spurious local minima. Such an investigation is relevant to many strategies for searching locally over the policy space, including policy gradient methods, natural gradient methods Kakade [2002] , finite difference methods Riedmiller et al. [2007], random search Mania et al. [2018], and evolutionary strategies Salimans et al. [2017]. For concreteness, one can mostly have in mind the idealized policy gradient iteration $\theta_{k+1} = \theta_k - \alpha_k \nabla \ell(\theta_k)$. We imagine applying policy gradient algorithms in simulation, where an appropriate restart distribution ρ provides sufficient exploration.

2 Further Related Literature

Beyond RL, 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 Bertsekas and Tsitsiklis [1996, 2000]. The ubiquity of non-convex optimization problems in machine learning and especially deep learning has sparked a slew of recent work [Agarwal et al., 2017, Carmon et al., 2018, Jin et al., 2017, Lee et al., 2016] 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, Bhojanapalli et al. [2016], Ge et al. [2015, 2016], Kawaguchi [2016], Sun et al. [2017]. 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 Kunnumkal and Topaloglu [2008] and Fazel et al. [2018]. For tabular MDPs with softmax policy parameterization, Thomas [2014] gives a simple argument that the gradient of the policy gradient cost function is never exactly equal to zero. Work on conservative policy iteration by Kakade and Langford [2002] laid some intellectual groundwork for studying policy gradient methods. An under-appreciated paper by Scherrer [2014] 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 the bottom of Subsection 5.3).

Concurrently with this work, Agarwal et al. [2019] 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 finite action settings. This unified analysis, and especially the closure condition we highlight, seems offer insight into when and why policy gradient methods can succeed despite non-convexity.

3 Problem formulation

Consider a Markov decision process (MDP), which is a six-tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, g, P, \gamma, \rho)$, 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 it is without loss of generality to index the states as $\mathcal{S} = \{1, \cdots, n\}$ where n is possibly infinite. For each state $s \in \mathcal{S}$, $\mathcal{A}_s \subset \mathbb{R}^k$ is 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 $\sup_{s\in\mathcal{S}, a\in\mathcal{A}_s} |g(s, a)| < \infty$. The assumptions that state spaces are at most countably infinite and per-period costs are bounded are standard and allow for a fully rigorous treatment without excessive technicality. We comment further in Remark 1.

Cost-to-go-functions and Bellman operators. A stationary policy is a function $\pi : \mathcal{S} \to \mathcal{A}$ 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. This is simply $\mathbb{R}^n$ when the state spaces is finite and is the set of bounded infinite sequences, typically denoted ℓ_∞, otherwise. 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 $s \in \mathcal{S}$. Similarly, let $P_\pi : \mathcal{J} \to \mathcal{J}$ be the transition operator defined by $(P_\pi J)(s) = \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s))J(s')$. When the state space is finite, this is simply the Markov transition matrix $P_\pi \in \mathbb{R}^{n \times n}$ under π whose (i, j)'th entry is equal to $P(j|i, \pi(i))$.

Define the Bellman operator $T_\pi : \mathcal{J} \to \mathcal{J}$ under the policy π as $T_\pi J := g_\pi + \gamma P_\pi J$. The cost-to-go function $J_\pi \in \mathcal{J}$ under policy π is the unique solution to the Bellman equation $J_\pi = T_\pi J_\pi$ and can be written out as

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

A helpful "variational form" the Bellman equation[2] [Bertsekas, 1995] is

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

for any bounded function $J \in \mathcal{J}$. The Bellman optimality operator is denoted by $T : \mathcal{J} \to \mathcal{J}$ and defined by $TJ = \min_{\pi \in \Pi} T_\pi J$, where the minimization is performed element-wise. 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 T and T_π are monotone and are contraction operators with respect to the maximum norm $\|\cdot\|_\infty$. Additional background is given in Appendix A.

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

$$Q_\pi(s, a) = g(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) J_\pi(s'), \tag{2}$$

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 (2) 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

$$Q_\pi(s, \pi(s)) = J_\pi(s), \qquad Q_\pi(s, \pi'(s)) = (T_{\pi'} J_\pi)(s), \qquad \min_{a \in \mathcal{A}_s} Q_\pi(s, a) = (TJ_\pi)(s). \tag{3}$$

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

$$\eta_\pi = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \rho P_\pi^t = (1 - \gamma)\rho(I - \gamma P_\pi)^{-1}. \tag{4}$$

[2] A closely related expression is called the performance difference lemma in the reinforcement literature, after Kakade and Langford [2002].

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, and so 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. *We assume throughout that ρ is supported on the entire state space, meaning that $\rho(s) > 0$ for all $s \in \mathcal{S}$.* Absent very strong assumption on the transition kernel, such an assumption is critical for ensuring the global convergence of policy gradient methods. Since ρ has full support, 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 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 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 next assume appropriate smoothness conditions on the policy class and on the cost and transition functions to ensure that $g(s,\pi_\theta(s))$ and $P(s'|s,\pi_\theta(s))$ are differentiable functions of θ.

To deal with infinite state spaces, we need also an assumption that gradient norms are uniformly bounded. The final regularity condition, involving the sum of the derivatives, comes from considering forward expectations of the form $\sum_{s'\in\mathcal{S}} P(s'|s,\pi_\theta(s))J(s')$. Assuming the summation and derivative can be exchanged, the partial derivative is $\sum_{s'\in\mathcal{S}} \frac{\partial}{\partial\theta_i} P(s'|s,\pi_\theta(s))J(s')$. We would like this derivate be finite for each $J \in \mathcal{J}$, which requires $\sum_{s'\in\mathcal{S}} \left|\frac{\partial}{\partial\theta_i} P(s'|s,\pi_\theta(s))\right| < \infty$. An alternate view is that this is precisely the condition that the partial derivative is integrable under the counting measure, which is the condition needed to apply the Leibniz rule justifying the interchange of the summation and derivative.

Assumption 1. *For all $s, s' \in \mathcal{S}$, $a \mapsto g(s,a)$ and $a \mapsto P(s'|s,a)$ are continuously differentiable functions on an open set containing $\mathcal{A}_s$ and $\theta \mapsto \pi_\theta(s)$ is continuously differentiable on an open set containing Θ. In addition, $\sup_{s\in\mathcal{S},a\in\mathcal{A}_s} \|\nabla_a g(s,a)\| < \infty$, $\sup_{s\in\mathcal{S},\theta\in\Theta} \|\nabla_\theta \pi_\theta(s)\| < \infty$, and for any $i \in \{1,\cdots,d\}$, $\sup_{s\in\mathcal{S},\theta\in\Theta} \sum_{s'\in\mathcal{S}} \left|\frac{\partial}{\partial\theta_i} P(s'|s,\pi_\theta(s))\right| < \infty$.*

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 $\|J\|_\infty$.

Interpretation as an average cost problem with restarts. Our discounted problem can be interpreted as an un-discounted problem, where, in every period, there is a constant probability $1-\gamma$ that the problem "restarts" in a random state drawn from the initial distribution ρ. In this formulation, $\ell(\pi)$ indicates the long-run average cost incurred by policy π and $\eta_\pi(s)$ indicates the fraction of time the policy spends in state s. This construction allows one to directly apply policy gradient results derived for average cost problems and gives a conceptually useful interpretation of ρ as an exploratory restart distribution. We formalize the claim in the following lemma, but this construction is standard in dynamic programming.

Lemma 1. *Suppose $\mathcal{S}$ is a finite and fix $\pi \in \Pi$. Consider a Markov chain $(s_0, s_1, \cdots)$ with transition probabilities*

$$\mathbb{P}\left(s_t = s | s_{t-1}, \cdots, s_0\right) = (1-\gamma)\rho(s) + \gamma P(s|s_{t-1}, \pi(s_{t-1})) \qquad s \in \mathcal{S}.$$

Then, at $T \to \infty$,

$$\frac{1}{T}\sum_{t=0}^{T-1} g_\pi(s_t) \xrightarrow{a.s.} \ell(\pi) \qquad \text{and} \qquad \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{1}(s_t = s) \xrightarrow{a.s.} \eta_\pi(s) \ \forall s \in \mathcal{S}.$$

Remark 1. *Our problem formulation restricts to problems with discrete state spaces and uniformly bounded cost functions, mirroring the main presentation in textbooks [Bertsekas, 1995, Puterman, 2014]. This choice allows us to clearly formalize the paper's main insights without excessive technical burden. It is likely that one can extend our results to problems with general state space, but there are severe measure-theoretic complications in the foundations of general state space DP and all result require careful technical qualifications [see Bertsekas and Shreve, 2004]. We will treat one problem, linear quadratic control, which falls outside our technical scope. Our approach there is to separately verify our general results and proofs can be essentially repeated for linear quadratic control.*

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 will 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.

Let us briefly review convergence to stationary points. A much more complete treatment can be found in nonlinear optimization textbooks [see e.g Bertsekas, 1997]. As made formal by the following definition, a point is said to be stationary if it 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 conditions. Note that, in unconstrained problems, where $\mathcal{X} = \mathbb{R}^d$, a point x is stationary if $\nabla f(x) = 0$.

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 called a*

stationary point[3] *if* $\langle x' - x, \nabla f(x) \rangle \geq 0$ *for all* $x' \in \mathcal{X}$ *. We say* $f(\cdot)$ *has no suboptimal stationary points if any stationary point* x *satisfies* $f(x) = \inf_{x' \in \mathcal{X}} f(x')$.

Under appropriate smoothness and regularity conditions, many optimization algorithms are guaranteed to converge to first-order stationary points. To make this concrete, consider the projected gradient descent algorithm. Applied with step-size sequence $\{\alpha_k\}_{k \in \mathbb{N}}$ and some initial point $x_0 \in \mathcal{X}$ the algorithm produces sequence of iterates

$$x_{k+1} = \text{Proj}_{\mathcal{X}}\left(x_k - \alpha_t \nabla f(x_k)\right) = \arg\min_{x \in \mathcal{X}} \left[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\alpha_t} \|x - x_t\|_2^2 \right].$$

Here $\text{Proj}_{\mathcal{X}}(y) = \arg\min_{x \in \mathcal{D}} \|y - x\|_2^2$ denotes the euclidean projection onto the closed convex set $\mathcal{X}$. The second equality is standard [See e.g. Bertsekas, 1997, chapter 3.3] and gives a "proximal" interpretation of projected gradient descent as minimizing a first order Taylor expansion of $f(\cdot)$ around x_t plus a regularizer that penalizes movement away from x_t. If x_t is a non-stationary point, then $\min_{x \in \mathcal{X}} \langle \nabla f(x_t), x - x_t \rangle < 0$ and there is some feasible direction of descent. For a sufficiently small stepsize choice, this ensures $f(x_{t+1}) < f(x_t)$. At a stationary point, we find $\min_{x \in \mathcal{X}} \langle \nabla f(x_t), x - x_t \rangle = 0$ and projected gradient descent would return $x_{t+1} = x_t$. With appropriate stepsize choices, many first order algorithms would continue making progress in this manner until eventually getting stuck at a stationary point, where progress stalls.

As one might expect from this discussion, under appropriate regularity conditions, projected gradient descent finds a global minimizer of $f(\cdot)$ when it has no suboptimal stationary points. For completeness, we formalize this in the following lemma. This result is meant to be illustrative and 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 8. Recall that a function f is said to be coercive if $f(x) \to \infty$ 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.

Lemma 2. *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, differentiable on an open set containing* $\mathcal{X}$ *and its gradient* ∇f *is Lipschitz continuous on* $\mathcal{X}$ *with Lipschitz constant* L*. Assume as well that either (i)* $\mathcal{X}$ *is compact or (ii)* $f(\cdot)$ *is coercive. Consider the sequence* $x_{k+1} = \text{Proj}_{\mathcal{X}}\left(x_k - \alpha \nabla f(x_k)\right)$ *where* $\alpha \in (0, 1/L]$*. If* $f(\cdot)$ *has no suboptimal stationary points then* $f(x_k) \to \min_{x \in \mathcal{X}} f(x)$ *as* $k \to \infty$.

Here we do not treat stochastic noise in the gradient evaluations. However, such extensions are possible. A rich literature on stochastic approximation shows that, under regularity conditions and decaying step-sizes, most noisy iterative algorithms converge to the same limit as their deterministic counterparts. See Borkar [2009] for a very general treatment and Bertsekas and Tsitsiklis [1996] for a very readable introduction. A rapidly growing literature studies convergence of stochastic first order methods in non-convex optimization [see e.g. Davis and Grimmer, 2019, Davis et al., 2020, Defazio et al., 2014, Ghadimi and Lan, 2013, 2016, Reddi et al., 2016a,b,c, Xiao and Zhang, 2014].

[3] 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.

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 in some sense 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 much more broadly. Beyond linear quadratic control, we instantiate this theory on finite state and action MDPs and on stopping problems with threshold policies.

5.1 Motivation from linear quadratic control

Leveraging many of the closed form expressions available for linear quadratic control, recent work of Fazel et al. [2018] showed that policy gradient methods 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 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. First, as highlighted in Equation (6) below, the class of linear policies is *closed under policy improvement steps*. Second, the policy iteration problem in (8) could be solved to optimality by a gradient method, since it is convex and hence has no suboptimal stationary points.

In this sense, policy iteration objective in (8) – which considers the impact of changing the policy for only one time step – has a very nice structure. 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. To simplify the presentation, we consider only deterministic linear quadratic control[4]. 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:*

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

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 finite per-step costs, $\|R\|_2, \|C\|_2 < \infty$. 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 Bertsekas [1995, 2011], Evans [2005]. 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 Fazel et al. [2018] for a simple example), making it unclear whether or why gradient descent on $\ell(\theta)$ would reach the global minimum. Precisely, if a policy policy π_θ is applied

[4]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.

from a state s_0 then from the linear dynamics we have $s_t = (A + B\theta)s_{t-1} = \cdots = (A + B\theta)^t s_0$. From this we have

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

A linear policy π_θ, is said to be stable if $\lambda_{\max}(A + B\theta) < 1$. Formally, let us define the set of stable policies as

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

For a stable linear policy, K_θ is finite and positive definite ensuirng the cost-to-go is finite. We assume the system, (A, B), is controllable so there exists at least one stable policy.

Even though the total cost function $\ell(\theta)$ is non-convex, the classical dynamic programming theory applies to the policy iteration algorithm for LQ control. The study of policy iteration in linear quadratic control dates back at least to Kleinman [1968], who showed that even in the undiscounted case, beginning with a stable linear policy it produces a sequence of stable linear policies with strictly improving cost to go until that converges toward an optimal policy. Essentially, the complexity of $\ell(\theta)$ is due the multi-period nature of the problem where 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.

$$\begin{aligned} TJ_{\pi_\theta}(s) &= \min_{a \in \mathbb{R}^k} \left[a^\top R a + s^\top C s + \gamma J_{\pi_\theta}(As + Ba) \right] \\ &= \min_{a \in \mathbb{R}^k} \left[a^\top R a + s^\top C s + \gamma (As + Ba)^\top K_\theta (As + Ba) \right] \end{aligned} \tag{5}$$

For an arbitrary 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 $\pi_{\overline{\theta}}$ is a policy iteration update if we choose $\overline{\theta} = -\gamma(R + \gamma B^\top K_\theta B)^{-1} B^\top K_\theta A$ In terms of Bellman operators, this means

$$T_{\pi_{\overline{\theta}}} J_{\pi_\theta} = T J_{\pi_\theta} \tag{6}$$

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

$$(T_{\pi_{\overline{\theta}}} J)(s) := (\overline{\theta} s)^\top R(\overline{\theta} s) + s^\top C s + \gamma J(As + B\overline{\theta} s). \tag{7}$$

This argument above shows how 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 begining of this section. For a given s, this optimization problem depends on $\overline{\theta}$ only though $\overline{\theta} s$ and hence an entire subspace of optimal solutions. The solution becomes unique either by requiring optimality at a set of states that span R^n or by solving the weighted policy iteration problem

$$\min_{\overline{\theta}} \mathbb{E}_{s \sim \eta} \left[Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) \right], \tag{8}$$

where the covariance of s under η has full rank. Note that following (3), we can equivalently write (6) in terms of Q-functions as: $Q_{\pi_\theta}(s, \overline{\theta}s) = \min_a Q_{\pi_\theta}(s, a)$ for all s, where

$$Q_{\pi_\theta}(s, a) = \left[a^\top Ra + s^\top Cs + \gamma(As + Ba)^\top K_\theta(As + Ba)\right]$$

Since $Q_{\pi_\theta}(s, a)$ is a quadratic function of a, $Q_{\pi_\theta}(s, \overline{\theta}s)$ is 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 (8) is convex (strongly convex in this case) and hence can be solved efficiently by a gradient method which is the second key property we identified.

Notice that the LQ control example does not fit within our general problem formulation, which assumed single period costs to be uniformly bounded. For LQ control, $\|J_\pi\|_\infty = \infty$ for any policy π, and so we cannot study convergence of algorithms in this norm. As is standard in dynamic programming, we instead study convergence in the weighted maximum norm $\|J_\pi\|_{\infty,w} = \sup_s |J_\pi(s)|w(s)$. Given the cost-to-go functions are quadratic, a natural choice is to take the weighting to be $w(s) = 1/\|s\|_2^2$. For any $K \in \mathbb{R}^{n\times n}$ such that $K \succ 0$, let $\mathcal{J}_q$ be the set of quadratic cost-to-go functions

$$\mathcal{J}_q = \left\{J : J(s) = s^\top Ks \ \ \forall\, s \in \mathbb{R}^n\right\}$$

With our choice of weighting, for any $J \in \mathcal{J}_q$, we have $\|J(s)\|_{\infty,w} = \|K\|_2$ which is the spectral norm of the matrix K. The follwoing lemma shows that the Bellman operator T_{π_θ}, corresponding to any stable linear policy $\theta \in \Theta_S$, and the Bellman optimality operator T are monotone contraction operators with respect to this weighted maximum norm. See Appendix C for a complete proof.

Lemma 3 (Bellman operators for LQ control)**.** *Consider the linear quadratic control problem formulated in Example 2. Fix the state-weighting $w(s) = 1/\|s\|_2^2$. 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. *(Contraction) $\|TJ - T\bar{J}\|_{\infty,w} \le \gamma\|J - \bar{J}\|_{\infty,w}$ and $\|T_{\pi_\theta}J - T_{\pi_\theta}\bar{J}\|_{\infty,w} \le \gamma\|J - \bar{J}\|_{\infty,w}$.*

A simple proof shows that for LQ control, the policy gradient loss function has no suboptimal stationary points depsite being non-convex. Starting from a suboptimal stable linear policy π_θ, the policy iteration step produces a new stable linear policy $\pi_{\overline{\theta}}$ with reduced cost. Using the convexity of the policy iteration objective[5] along with standard dynamic programming arguments, we show that $(\overline{\theta} - \theta)$ forms a descent direction for the policy gradient 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 4. *For the linear quadratic control problem formulated in Example 2, if ρ has a strictly positive density on $\mathbb{R}^n$, any stable linear policy θ satisfies $\nabla\ell(\theta) = 0$ if and only if $J_{\pi_\theta} = J^*$.*

[5] We showed above that $a \to Q_{\pi_\theta}(s, a)$ is convex quadratic in a.

Proof. Fix a stable linear policy π_θ and let $\pi_{\overline{\theta}}$ be a policy iteration update to π_θ. In other words, $\overline{\theta}$ is a solution to (8) and satisfies $T_{\pi_{\overline{\theta}}} J_{\pi_\theta} = T J_{\pi_\theta}$. Set $\theta^\alpha = (1-\alpha)\theta + \alpha\overline{\theta}$ which implies $\pi_{\theta^\alpha}(s) = (1-\alpha)\theta s + \alpha\overline{\theta} s$ as 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 before. Repeatedly applying the Bellman operator and using the monotonicity property, $J_{\pi_\theta} \preceq T J_{\pi_\theta}$, shown above in Lemma 3 gives:

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

From this, we have

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

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. But since $E(s)$ is continuous (it is the difference in quadratic functions), there is an open subset of states on which $E(\cdot)$ is strictly negative. Since ρ has a positive density on all of $\mathbb{R}^n$, the expectation on the right hand side is strictly negative showing $\overline{\theta} - \theta$ to be a descent direction. Other side of the claim follows using the policy gradient theorem as shown in Lemma 5 (see Section 5.4). □

5.2 General results

Let us generalize (8) by introducing the weighted policy iteration, or "Bellman" objective,

$$\mathcal{B}(\pi' \mid \eta, J_\pi) = \mathbb{E}_{s\sim\eta} \left[(T_{\pi'} J_\pi(s)) \right], \tag{9}$$

and overload notation to write $\mathcal{B}(\theta \mid \eta, J_\pi) = \mathcal{B}(\pi_\theta \mid \eta, J_\pi)$. Recall that $(T_{\pi'} J_\pi)(s) \equiv Q_\pi(s, \pi'(s))$. Clearly, when η is supported on all states, minimizing (9) over all policies π' is equivalent to classical policy iteration. Policy gradient methods are more closely related to the following weighted policy iteration scheme:

$$\theta_{t+1} = \arg\min_{\theta\in\Theta} \mathcal{B}(\theta \mid \eta, J_{\pi_{\theta_t}}), \tag{10}$$

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 $\pi \in \Pi_\Theta$ and distribution η supported over the entire $\mathcal{S}$, there is a $\pi^+ \in \Pi_\Theta$ such that $T_{\pi^+} J_\pi = T J_\pi$. Equivalently, $\mathcal{B}(\pi^+|\eta, J_\pi) = \min_{\pi' \in \Pi} \mathcal{B}(\pi'|\eta, J_\pi)$.*

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 have considered so far. This is due to the multi-period nature of the decision problem, in which changes to the policy can have a compounding effect[6] over time. Since it considers only a single period decision problem, the weighted policy iteration objective $\bar{\theta} \mapsto \mathcal{B}(\bar{\theta}|\eta_{\pi_\theta}, J_{\pi_\theta})$ is typically much simpler.

Condition 2 (Stationary points of the policy improvement objective)**.** *For each $\pi \in \Pi_\Theta$ and distribution η supported over $\mathcal{S}$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta, J_\pi)$ has no sub-optimal stationary points.*

Recall, we already proved this condition for optimal stopping problem in Example 4. For finite MDPs, we showed $\theta \mapsto \mathcal{B}(\theta \mid \eta, J_\pi)$ is linear while for LQ control, it is quadratic. The next theorem offers a broad generalization of the result for LQ control shown in Lemma 4. After developing some supporting results in the next subsection, we prove Theorem 1 in Section 5.5.

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

5.3 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 vector of probabilities where each component a_i denotes the probability of taking the ith action. Cost and transition functions can be naturally extended to functions on the probability simplex by defining:*

$$g(s, a) = \sum_{i=1}^{k} g(s, e_i)\, a_i \qquad\qquad P(s'|s, a) = \sum_{i=1}^{k} P(s'|s, e_i)\, a_i. \tag{11}$$

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

For this tabular setting, a natural paramerization 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

[6] In terms of the distribution of states and actions visited over a trajectory under the updated policy vs the old 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 $\Pi = \{\pi \in \mathbb{R}_+^{n\times k} : \sum_{i=1}^k \pi_{s,i} = 1 \; \forall s \in \{1, \cdots, n\}\}$ which obviously implies that Π is 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,

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

is convex (linear) in π' and can be solved efficiently by projected gradient method, for example.

Remark 2. *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 $\pi_\theta(s) \in \Delta^{k-1}$ is a probability distribution whose components $\pi_\theta(s) \equiv (\pi_\theta(1|s), \cdots, \pi_\theta(k|s))$ satisfy*

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

We simplify the discussion by assuming $\theta_{s,1} = 1$ is fixed. This means that each θ defines a unique policy. Otherwise, the policy class is over parameterized. It is important to note that our result about stationary points in Theorem 1 does not apply in a meaningful way to softmax policies. In non-degenerate cases, any policy π_θ corresponding to a given θ is suboptimal, so our result suggests there are is no θ that is a stationary point of $\ell(\theta)$. Convergence 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 2.

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. 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 $U_i = 1/k$ for each i. We have chosen to regularize the single stage cost functions, rather than $\ell(\theta)$ directly, because this form lies within the scope of our problem formulation. For such a regularizer, $R(\pi_\theta(s)) \to \infty$ if $\|\theta_s\| \to \infty$, 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 general results can be used to show $\ell(\theta)$ has no suboptimal stationary points, so Lemma 2 shows 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 where in each round the agent observes a contextual information, $x_t \in \mathcal{X}$ which evolves according to an uncontrolled Markov chain with transition kernel from x to x' given by $p(x'|x)$. Conditioned on context x_t, the agent receives an offer $y_t \in \mathcal{Y}$ drawn i.i.d from some distribution $q_{x_t}(\cdot)$. 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 costless. The agent's objective is to maximize ravenue[7].We assume the context set, $\mathcal{X}$ to be finite and the offer set to $\mathcal{Y}$ to be a bounded subset of $\mathbb{R}$.*

[7] One can imagine costs to be the negative reward to be consistent with our formulation.

The problem can be formalized as a Markov decision process with the state-space $\mathcal{S} = \mathcal{S}_{\mathrm{C}} \cup \{T\}$, consisting of a fnite set of continuing states, $\mathcal{S}_{\mathrm{C}} = (\mathcal{X} \times \mathcal{Y})$ and a terminal state T that is costless, $g(T, a) = 0$ and absorbing, $P(T|T, a) = 1$. We assume that the initial distribution $\rho(s)$ has a positive density over $\mathcal{S}_{\mathrm{C}}$. There are two actions, $\mathcal{A} = \{0, 1\}$, action $a = 0$ corresponds to accepting the offer and terminating while $a = 1$ continues the game by transitioning to a new state:

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

We consider the class of threshold policies, $\Pi_\theta := \{\theta \in \mathbb{R}^{|\mathcal{X}|} : \theta_x \in \mathcal{Y}\}$, parametrized by one threshold per context x such that the policy, $\pi_\theta(s = (x, y)) = \mathbb{1}(y < \theta_x)$, rejects all offers below the threshold. It is easy to verify that the class of threshold policies is closed under policy improvement. For any $\pi \in \Pi_\theta$, consider the policy iteration update for any state $s = (x, y) \in \mathcal{S}_{\mathrm{C}}$:

$$\pi^+(x, y) = \arg\max_{a \in \{0,1\}} Q_\pi((x, y), a) = ay + (1 - a)\gamma \sum_{(x', y') \in \mathcal{S}} p(x'|x) q_x(y') J_\pi((x', y'))$$

Clearly, $\pi^+(x, y) = 1$ iff y exceeds the continuation value, $c_\pi(x) := \gamma \sum_{(x', y') \in \mathcal{S}} p(x'|x) q_x(y') J_\pi((x', y'))$. Thus, π^+ is itself a threshold policy. We can also 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 $\pi_\theta, \pi \in \Pi_\theta$ and η with positive density over $\mathcal{S}_{\mathrm{C}}$,

$$\frac{\partial}{\partial \theta_x} \mathcal{B}(\theta|\eta, J_\pi) = \frac{\partial}{\partial \theta_x} \int_{y \in \mathcal{Y}} \eta(x, y) \cdot Q_\pi\left((x, y), \pi_\theta(x, y)\right) dy = \left(c_\pi(x) - \theta_x\right) \eta(x, \theta_x)$$

Thus, $\frac{\partial}{\partial \theta_x} \mathcal{B}(\theta|\eta, J_\pi) = 0 \iff \theta_x = c_\pi(x)$ which corresponds to the optimal policy[8].

Descent direction for Examples 3 and 4: Astute readers will observe that the proof of lemma 4 was highly specialized to the setup with linear policy classes and can be extended to convex policy classes in the sense that policy iteration update still provides a feasible descent direction. So while that proof does apply to the tabular MDP example, it breaks for the class of threshold policies which is not convex[9]. In the next subsection, instead of constructing a descent direction, we develop a more general approach to argue about the stationary points of $\ell(\cdot)$ using Conditions 1 and 2 along with the policy gradient lemma.

5.4 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

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \mathcal{B}(\theta \mid \eta, J_{\pi_{\theta_t}}), \tag{12}$$

[8] For an optimal policy, π_θ it must be true that $\theta_x = c_{\pi_\theta}(x)$. That is, the threshold should equal the continuation value – if that does not hold, then we can always improve revenue.

[9] It is easy to see that for any two threshold policies, $\pi_\theta, \pi_{\theta'}$, their convex combination :$\alpha\pi_\theta + (1 - \alpha)\pi_{\theta'}$ is not a threshold policy. The class of threshold policies is therefore not convex.

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 iteraion objective.

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

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 (9) 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.

The policy gradient theorem below which enables this connection can essentially be derived from the result of Silver et al. [2014]. While they call this result a deterministic policy gradient theorem, it actually generalizes the stochastic policy gradient theorem of Sutton et al. [2000] when $\pi_\theta(s) \in \Delta^{k-1}$ specifies a probability distribution over k base actions. We provide a short proof sketch here, since the presentation of the result and the proof differ from Silver et al. [2014] and Sutton et al. [2000].

Lemma 5 (Policy gradient theorem)**.** *$\ell(\theta)$ is differentiable and*

$$\nabla \ell(\theta) = \nabla_{\overline{\theta}} \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta}) \Big|_{\overline{\theta}=\theta} = \sum_{s \in \mathcal{S}} \eta_{\pi_\theta}(s) \left[\nabla_{\overline{\theta}} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) \Big|_{\overline{\theta}=\theta} \right].$$

Proof sketch. Put $J_\theta \equiv J_{\pi_\theta}$, $Q_\theta \equiv Q_{\pi_\theta}$, $P_\theta \equiv P_{\pi_\theta}$ and $T_\theta \equiv T_{\pi_\theta}$ and $\eta_\theta \equiv \eta_{\pi_\theta}$. We leverage the variational Bellman equation (1) (take $\pi = \pi_{\overline{\theta}}$ an $J = J_\theta$ in that expression). This gives,

$$\begin{aligned}
\ell(\overline{\theta}) - \ell(\theta) = (1-\gamma)\rho\left(J_{\overline{\theta}} - J_\theta\right) &= (1-\gamma)\rho\left((I - \gamma P_{\overline{\theta}})^{-1}(T_{\overline{\theta}} J_\theta - J_\theta)\right) \\
&= \eta_{\overline{\theta}}\left(T_{\overline{\theta}} J_\theta - J_\theta\right) \\
&= \left(T_{\overline{\theta}} J_\theta - J_\theta\right) + \underbrace{\left[(\eta_{\overline{\theta}} - \eta_\theta)\left(T_{\overline{\theta}} J_\theta - T_\theta J_\theta\right)\right]}_{O(\|\overline{\theta}-\theta\|_2^2)}.
\end{aligned}$$

Note that $\eta_\theta \left(T_{\overline{\theta}} J_\theta\right) = \mathcal{B}(\overline{\theta} \mid \eta_\theta, J_{\pi_\theta})$. Evaluating the derivative with respect to $\overline{\theta}$ at $\overline{\theta} = \theta$ yields the first result. The second equality follows by rewriting $\mathcal{B}(\overline{\theta} \mid \eta_\theta, J_{\pi_\theta}) = \mathbb{E}_{s \sim \eta_\theta}\left[Q_\theta(s, \pi_{\overline{\theta}}(s))\right]$ and interchanging the expectation and derivative. To make this proof fully rigorous, we need to justify this interchange and also to show formally that $\|\eta_{\overline{\theta}} - \eta_\theta\|_\infty = O(\|\overline{\theta} - \theta\|_2)$ and $\|T_{\overline{\theta}} J_\theta - T_\theta J_\theta\|_\infty = O(\|\overline{\theta} - \theta\|_2)$. See Appendix E.1 for a detailed proof. □

5.5 Proof of Theorem 1

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 6. *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 5, 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 (9) 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 7, 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 optimaility, i.e. $J_\pi = J^* \iff J_\pi = T J_\pi$. For completeness, we give a proof in Appendix A.

Lemma 7 (On average Bellman equation)**.** *For any $\pi \in \Pi_\Theta$ and $S \sim \eta_\pi$,*

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

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

Completing the proof of Theorem 1. 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 6 and the second equality uses Condition 1. Finally, Lemma 7 shows that satisfying the average Bellman equation implies optimiality. □

6 Beyond closed policy classes: the case of nonstationary policy classes

For finite horizon problems with non-stationary policy classes, we can guarantee that there are no spurious local minima for policy gradient 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[10]. 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 the result relies critically on the use of a non-stationary policy class. In particular, 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. Kunnumkal and Topaloglu [2008] have previously showed through a somewhat intricate analysis that a stochastic approximation algorithm converges to the optimal policy, despite non-convexity of the objective.

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

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 $s_t \in \mathbb{R}$ 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 $w_t \in [0, w_{\max}]$ to be the random demand (assumed to be i.i.d for simplicity). For a problem with horizon H, the seller's objective is to minimize total expected cost*

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

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. We assume that $p > c$. Otherwise, the optimal policy never orders inventory.

It is well known that a base-stock policy is optimal for this setting [Bertsekas, 1995]. 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 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. Osband et al. [2017]). Essentially, one can imagine that the state space factorizes into $H+1$ components, 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 encoding 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 costless 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. For any fixed $s \in \mathcal{S}_h$, $\pi_\theta(s)$ depends only on θ_h.*

We now state the main result of this subsection, which applies under conditions much weaker than those for Theorem 1. 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 objectives corresponding to the optimal cost-to-go function, J^*.

Condition 4. *For any distribution η with full supported on $\mathcal{S} = \cup_{h=1}^{H+1}$, the problem $\min_{\theta \in \Theta} \mathcal{B}(\theta|\eta, J^*)$ has no suboptimal stationary points.*

Recall that Condition 2 considers stationary points of single-period problems, $\theta \to \mathcal{B}(\theta|\eta, J_\pi)$, induced by any suboptimal policy $\pi \in \Pi_\theta$. The full proof is given in Appendix E.2 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 act optimally from any state in $\mathcal{S}_h$ for $h < H$.

Theorem 2. *Suppose Conditions 3 and 4 hold. Further assume that ρ is supported over each $\mathcal{S}_h$ for $h \leq H$. If the parameterized policy class Π_Θ contains an optimal policy π^*, then any stationary point θ of $\ell : \Theta \to \mathbb{R}$ satisfies $J_{\pi_\theta} = J^*$.*

For the inventory control problem in Example 5, a simple argument shows that $\mathcal{B}(\theta|\eta, J^*)$ has no suboptimal stationary points for any η supported over $\mathcal{S}$. The result, shown in Appendix E.2, essentially follows by using that the optimal Q-function, $Q^*(s, a)$ is convex in a (see Chapter 3 in Bertsekas [1995] for details).

7 The initial distribution and concentrability coefficients

7.1 The role of the initial distribution

Our analysis relies critically on an exploratory initial distribution ρ. This is not an artifact of the proof technique, as indeed, policy gradient methods will in general have poor convergence properties otherwise. One counterexample was provided by Kakade and Langford [2002]. Examples of this form can be assumed away, effectively by imposing certain uniform ergodicity assumptions on the underlying MDP. Instead, we choose to make transparent that such a ρ will be generally be needed for robust results. In many real-world learning scenarios, restarting in very different states would be costly if not impossible. Perhaps for this reason, introductory materials on policy gradient methods often omit an discussion of an exploratory initial distribution. However, nearly all implementations of policy gradient methods of which we are aware involve training in simulated environments or laboratory envirnments. It is common to initial present robots with diverse scenarios or diverse tasks at training time, with the engineering playing the role of an (adaptive) choice of ρ.

7.2 Concentrability

Our results in Section 5 focussed on characterzing the quality of stationary points by arguing that at any stationary point, the cost-to-go functions satisfy an average Bellman equation which, crucially, implies optimality (see Lemma 7). An important concept used to extend our analysis grapples with connecting errors in the Bellman equation to the optimality gap in terms of cost-to-go functions. As an example, the Bellman optimality operator T is a contraction in a norm $\|\cdot\|$ then

$$\|J - J^*\| \leq \frac{1}{(1-\gamma)}\|J - TJ\| \quad \forall\, J \in \mathcal{J}. \tag{13}$$

See Bertsekas [1995] or (23) in Appendix A. Here it was critical that T is a contraction in the same norm used to measure the distance from the optimal cost-to-go function. We define a constant, κ_ρ, which enables the same inequality as in (13) but instead with the weighted norm, $\|\cdot\|_{1,\rho}$. Intuitively, κ_ρ captures how errors in the cost-to-go functions manifest in Bellman errors that are detectable by 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 11 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*

$$\|J - J^*\|_{1,\rho} \leq \frac{\kappa_\rho}{(1-\gamma)} \|J - TJ\|_{1,\rho} \qquad \forall J \in \mathcal{J}_\Theta. \tag{14}$$

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 (14) modulo a constant factor. Second, due to the policy gradient formula in Lemma 5, our results crucially on errors Bellman equation weighted under the state occupancy measure η_π. See Lemma 7, 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 (Munos [2003, 2007]) that play a key role in the analysis of approximate value and policy iteration algorithms [Farahmand et al., 2010, Geist et al., 2017, Kakade and Langford, 2002, Munos, 2003, 2007, Munos and Szepesvári, 2008, Scherrer and Geist, 2014]. See Scherrer [2014] for a detailed comparison on different notions of the concentrability coefficient. Note, instead of stating a more general regularity assumption on the MDP, our definition of κ_ρ in (14) is precisely the quantitiy 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 $\kappa_\rho < \min_s 1/\rho(s)$ in any finite state problem, but it could also be finite in some infinite state problems.

Lemma 8. *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 Kakade and Langford [2002] and can be derived using the variational form of the Bellman equation (see 25 in Appendix A) or using the performance difference lemma Kakade and Langford [2002]. For completeness, we give a short proof in Appendix E.3. Such *distributional mismatch* terms also appears in the works of Agarwal et al. [2019], Scherrer and Geist [2014].

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

$$c\|J\| \leq \|J\|_{1,\rho} \leq C\|J\| \qquad \forall\, J \in \mathcal{J}, \tag{15}$$

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 (23) in Appendix A). 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 9 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 10 (Concentrability in optimal stopping)**.** *Suppose $\mathcal{S} = \mathcal{S}_C \cup \{T\}$ consists of a finite set of continuing states $\mathcal{S}_C$ and terminal state T that is absorbing ($P(T|T,a) = 1$) and costless ($g(T,a) = 0$). There are two actions $\mathcal{A} = \{0,1\}$, stop ($a = 0$) and continue ($a = 1$). Consider the policy 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 Tsitsiklis and Van Roy [2001] shows T is a contraction in $\|\cdot\|_{2,\mu}$. Similarly, one can show T is a contraction with modulus γ in $\|\cdot\|_{1,\mu}$. The result then follows immediately from Lemma 9. See Appendix E.3 for details. □

The definition of κ_ρ only requires that 14 allows for bounds that depend on regularity properties in the cost-to-go functions of interest. These regularity properties are not captured by the bound 8, but can sometimes be

is also useful in cases where the bound stated in is potentially pessimistic as it does not capture any regularity properties in the cost-to-go functions of interest. For example, in linear quadratic control, cost-to-go functions induced by the class of linear policies are quadratic. As a result, we need only the initial distribution to explore basis of the state space sufficiently, rather than requiring it to mimic the steady state distribution of the (unknown) optimal policy. For the LQ control problem formulated in Example 2, the following lemma shows that k_ρ is bounded by the dimension of the basis of $\mathcal{S}$ (which is n as $\mathcal{S} = \mathbb{R}^n$) as well as the condition number of the (unnormalized) initial state correlation matrix, $\Sigma = \mathbb{E}_{s\sim\rho}[ss^\top]$, reflecting asymmetry in exploring the different directions.

Lemma 11 (Concentrability in LQ control)**.** *Consider the linear quadratic control problem in Example 2. Suppose $\Sigma = \mathbb{E}_{s\sim\rho}[ss^\top] \succ 0$. Then, $\kappa_\rho \leq n \cdot \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$.*

8 Convergence rates for policy gradient methods

Our result in Theorem 1 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, $\theta \mapsto \mathcal{B}(\theta \,|\, \eta, J_\pi)$ 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 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, $\theta \mapsto \mathcal{B}(\theta \,|\, \eta, J_\pi)$ is gradient dominated as well as closure of the policy class (condition 1). We

use the policy gradient theorem in Lemma 5 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 Nesterov, 2018]. Such gradient dominance conditions also underly the proof of Fazel et al. [2018] for linear quadratic control.

We remark that assuming the policy improvement objective to be convex is not entirely impractical. As noted before in Section 5, $\theta \mapsto \mathcal{B}(\theta \,|\, \eta, J_\pi)$, is linear for finite MDPs and quadratic for LQ control problem and therefore our results in this section immediately apply to these examples.

8.1 Background on Gradient Dominance

Throughout this section, 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 differentiable on an open set containing $\mathcal{X}$. Recall 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

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

When $\|x - x'\| \leq R$ for all $x, x' \in \mathcal{X}$, one can deduce from this the condition 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

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

When $\mathcal{X}$ is unconstrained, 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 called gradient dominance. This definition, essentially, assumes a critical implication of convexity or strong convexity rather than assuming the 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. The definition below is somewhat non-standard because we discuss constrained optimization. Most authors discuss unconstrained optimization, in which case the definition reduces to the usual gradient dominance or so-called Polyak condition, $\min_{x'\in\mathcal{X}} f(x) \geq f(x) - \frac{\mu}{2c}\|\nabla f(x)\|_2^2$.

Definition 3. *For a closed convex set $\mathcal{X} \subset \mathbb{R}^d$ and function f that is differentiable on an open set containing $\mathcal{X}$, we say f is (c, μ)–gradient dominated over $\mathcal{X}$ if there exists a constant $c > 0$ and $\mu \geq 0$ such that*

$$\min_{x'\in\mathcal{X}} f(x') \geq f(x) + \min_{x'\in\mathcal{X}} \left[c\,\langle \nabla f(x), x' - x\rangle + \frac{\mu}{2}\left\|x - x'\right\|_2^2\right] \qquad \forall\, x \in \mathcal{X}. \tag{16}$$

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

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. For concreteness, we will focus on projected gradient descent, strengthening Lemma 2 by providing a convergence rate. The first result is obtained by using a well known fact that projected gradient descent reaches an approximate stationary point rapidly (approximately at $\mathcal{O}(1/\sqrt{T})$ rate), and then using the definition of gradient dominance to relate approximate stationarity to the optimality gap. Analysis in the second case is essentially identical to the typical analysis of projected gradient descent for strongly convex objectives and shows a geometric convergence rate. 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 12 (Convergence rates for gradient dominated smooth functions)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set and f be L–smooth on $\mathcal{X}$. Consider the sequence $x_{t+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_t - \alpha \nabla f(x_t)\right)$ where $\alpha \leq 1/L$. Set $f(x^*) = \min_{x' \in \mathcal{X}} f(x')$. Then,*

1. *If f is $(c, 0)$–gradient-dominated and $\|x - x'\|_2 \leq R < \infty$ for all $x, x' \in \mathcal{X}$, then*

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

2. *If f is (c, μ)–gradient-dominated for $\mu > 0$, then,*

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

Proof. See Appendix B.1 for a detailed proof. □

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 the stronger property that it is gradient dominated. This condition holds for the examples in Section 5 and hold whenever the single period objective solved by policy iteration is convex.

Condition 5 (Gradient dominance of the weighted policy iteration objective)**.** *For any $\pi \in \Pi_\Theta$ and probability distribution η supported over $\mathcal{S}$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta, J_\pi)$ 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 3 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 14. The Corollary 1 provides a more interpretable statement, which follows because any convex $\mathcal{B}(\theta|\eta, J_\pi)$ is $(1, 0)$–gradient-dominated and any $\mathcal{B}(\theta|\eta, J_\pi)$ and any μ–strongly–convex $\mathcal{B}(\theta|\eta, J_\pi)$ is $(1, \mu)$–gradient-dominated.

Theorem 3. *If conditions 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 1 holds. If, for every $\pi \in \Pi_\Theta$ and probability distribution η supported over $\mathcal{S}$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta, J_\pi)$ is convex, then $\ell(\theta)$ is gradient dominated of degree one. If $\theta \mapsto \mathcal{B}(\theta \mid \eta, J_\pi)$ 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 of the policy class (condition 1) 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 sufficient exploration, $\rho(s) > 0\ \forall\, s \in \mathcal{S}$. It is important to note how *optimality* here crucially depends on our assumption that the start state distribution ρ is supported over the entire state space as it ensures that $\eta_\pi(s) > 0\ \forall\, s \in \mathcal{S}$ for any $\pi \in \Pi$. 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 5.

Proof of Theorem 3. We first derive a consequence of the closure condition. Let S denote a random draw from η_{π_θ}. We have,

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

Here (a) uses that $J_{\pi_\theta} \succeq T J_{\pi_\theta}$, (b) directly applies the definition of κ_ρ in Definition 14, (c) directly applies the policy closure condition in Condition 1, and (d) uses that $\eta_{\pi_\theta} \geq (1-\gamma)\rho$ (by definition, see (4)).

As we assume $\theta \mapsto \mathcal{B}(\theta \mid \eta, J_\pi)$ is (c, μ)–gradient dominated for each $\pi \in \Pi_\Theta$ and η supported over $\mathcal{S}$, we have that $\forall\, \theta \in \Theta$

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

which gives our desired result. The second inequality above uses the policy gradient theorem in Lemma 5. □

9 Geometric convergence in finite action problems

While most of this paper was completed simultaneously with Agarwal et al. [2019], this section is inspired directly by their work. In particular, one contribution of that work is to provide rates of convergence for policy gradient methods – and in particular, natural policy gradient methods – with exact gradient evaluations in finite state and action problems with a policy class consisting of all stochastic policies. Depending on the precise algorithm, 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 by posing these as smooth nonlinear optimization problems and suggesting small stepsizes to control for the error due to local linearization. In this work, we highlight that many first order methods applied to $\ell(\cdot)$, including natural policy gradient methods, can work with extremely large stepsizes and attain a *linear* rate of convergence, or equivalently, they require only $O(\log(1/\epsilon))$ iterations to reach an ϵ–optimal policy.

Throughout this paper, we have reduced the analysis of policy gradient methods to that of first order methods on the weighted policy iteration cost function, $\mathcal{B}(\bar{\pi}|\eta, J_\pi)$. The special feature of the finite action case is that the policy iteration step reduces to a linear optimization problem over the probability simplex, a trivial problem which simply selects the best element among a finite set. First order methods can potentially solve such problems in just a single iteration using a large stepsize. From this, we are able to deduce a geometric rate of convergence for policy gradient methods, akin to making exact policy iteration updates.

9.1 Setup

Recall the problem setup for finite state action MDPs as shown in Example 3. There are n state and k deterministic actions to choose from. We take the set of feasible actions, $\mathcal{A} = \Delta^{k-1}$, to be the set of all probability distributions over these k actions. We consider the following two commonly used policy parameterizations:

Natural parameterization. Here, $\pi \in \mathbb{R}^{n\times k}$ denotes a stochastic policy, viewed as a matrix where each row is a probability distribution over actions for a given state. The set $\Pi = \{\pi \in \mathbb{R}_+^{n\times k} : \sum_{i=1}^k \pi_{s,i} = 1 \ \forall s \in \{1, \cdots, n\}\}$ deontes the set of all stationary randomized policies.

Softmax parameterization. For unconstrained $\theta \in \mathbb{R}^{n\times k}$, the softmax policy specifies an action-distribution $\pi_\theta(s) \in \Delta^{k-1}$ for each $s \in \{1, \cdots, n\}$. The vector $\pi_\theta(s) \equiv (\pi_\theta(1|s), \cdots, \pi_\theta(k|s))$ has components

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

Policy gradients. Gradient calculations take on a particular transparent form in tabular problems. Let e_i be the i-th standard basis vector, representing one of the k possible deterministic actions. We can rewrite the policy iteration cost function as

$$\mathcal{B}(\bar{\pi}|\eta_\pi, J_\pi) = \sum_{s=1}^n \eta_\pi(s) \left(\sum_{i=1}^k \bar{\pi}_{s,i} \cdot Q_\pi(s, e_i) \right) = \langle \bar{\pi} \, , \, Q_\pi \rangle_{\eta_\pi \times 1} \tag{17}$$

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). The policy

gradient theorem (Lemma 5) shows $\nabla \ell(\pi) = \nabla \mathcal{B}(\bar{\pi}|\eta_\pi, J_\pi)\big|_{\bar{\pi}=\pi}$. Expressed in terms of partial derivatives, this gives $\frac{\partial}{\partial \pi_{s,i}} \ell(\pi) = \frac{1}{1-\gamma} \eta_\pi(s) Q_\pi(s, e_i)$. First order methods use the local linear approximation to ℓ,

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

It is important that $\eta_\pi(s) > 0$ for all s, since $\rho(s) > 0$.

9.2 Algorithms

We now specialize several first-order algorithms to this setting. Recall the set of stochastic policies $\Pi = \{\pi \in \mathbb{R}_+^{n \times k} : \sum_{i=1}^k \pi_{s,i} = 1 \ \forall s \in \{1, \cdots, n\}\}$, and note that $\Pi = \Delta^{k-1}(1) \times \cdots \times \Delta^{k-1}(n)$ 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

$$\pi' = \arg\min_{\bar{\pi} \in \Pi} \langle \nabla \ell(\pi), \bar{\pi} \rangle = \arg\min_{\bar{\pi} \in \Pi} \langle Q_\pi, \bar{\pi} \rangle_{\eta_\pi \times 1} \tag{18}$$

and then updates the policy to $\pi^+ = (1-\alpha)\pi + \alpha\pi'$. In this case, π' is exactly a policy iteration update to π so *Frank-Wolfe mimics a soft-policy iteration step*, akin to the update in Kakade and Langford [2002]. Note, the minimization problem in (18) decouples across states to optimize a linear objective over the probability simplex, so $\pi^+(s) \in \arg\min_{d \in \Delta^{k-1}} \eta_\pi(s) d^\top Q_\pi(s, \cdot)$ 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 quadratically 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 (18), 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: $\pi_s^+ = \mathrm{Proj}_{2, \Delta^{k-1}}(\pi_s - \alpha Q_\pi(s, \cdot))$, 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 ($\mathcal{O}(k \log k)$) soft-thresholding operation [Duchi et al., 2008].

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^+:

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

where KL divergence is defined as $D_{\mathrm{KL}}(p||q) = \sum_{i=1}^{k} p_i \log(p_i/q_i)$. Equivalently, π^+ can be written as the exponentiated gradient update,

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

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 Kakade [2002] applied to the softmax parameterization described above. This is closely related to the widely used TRPO algorithm of Schulman et al. [2015a]. NPG with softmax policies is actually an instance of mirror descent with a specific regualrizer. In particular, beginning with some policy $\pi \in \Pi$, an iteration of NPG updates to π^+:

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

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.

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. But it is known to be equivalent to mirror descent under some conditions [Raskutti and Mukherjee, 2015]. In this case, readers can check that the exponentiated update above matches the explicit formula for the NPG update given in Kakade [2002] and Agarwal et al. [2019].

The choice of stepsizes is an important issue for most first order methods. Each a of the algorithms above can be applied with a sequence of stepsizes $\{\alpha_t\}$ to produce a sequence of policies $\{\pi_t\}$. At iteration t, the rules above actually specify a one dimensional family of updated policies π^α_{t+1} that depends on the choice of stepsize α. An idealized stepsize rule is *exact line search*, which directly optimizes over this choice of stepize at each iteration:

$$\alpha_t = \arg\min_{\alpha \geq 0} \ell(\pi^\alpha_t) \tag{19}$$

If this minimum is not attained, then all results apply by choosing an ϵ–optimal solution in (19), for some arbitrarily small ϵ.

9.3 Geometric convergence

So far, we have described different vaariants of implementing policy gradients for tabular settings. Essentially, all of these algorithms make policy iteration updates for sufficiently large stepsizes.

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 4 below.

Our first result confirms that all of the algorithms we presented in the previous section will converge 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 atleast 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 $\min_{s\in\mathcal{S}} \rho(s)$ term 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 constant stepsizes. The simplest result applies to the Frank-Wolfe algorithm, which we already showed to be exactly equivalent to a soft policy iteration update. However, the NPG result is likely more important since variants of this algorithm are widely used in practice. We show NPG with exact gradient updates will reach an ϵ optimal policy in $O(\log(1/\epsilon))$ iterations with sufficiently large stepsizes. The constant error term is inversely related to the stepsize α_t and reflects the fact that NPG updates with finite stepsizes only approximately resemble the policy iteration updates. As $\alpha_t \to \infty$, we recover the exact result as one would expect for policy iteration.

Many caveats apply to these results. The literature claims to effectively approximate natural policy gradient updates with complex deep neural networks [Schulman et al., 2015a], 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 we believe it is valuable for researchers to have a clear understanding of rates of convergence in this idealized case.

Theorem 4 (Geometric convergence)**.** *Suppose one of the first-order algorithms in subsection 9.2 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 step-sizes chosen by exact line search as in (19), then*

$$\|J_{\pi^t} - J^*\|_\infty \leq (1 - \min_{s\in\mathcal{S}} \rho(s)\,(1-\gamma)^t\,\|J_{\pi^0} - J^*\|_\infty.$$

(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) ***Constant stepsize natural policy gradient.*** *Fix any $\epsilon > 0$. Under NPG with stepsize sequence $\alpha_t \geq \frac{2\log(2)}{(1-\gamma)\epsilon}$,*

$$\|J_{\pi^t} - J^*\|_\infty \leq \left(\frac{1+\gamma}{2}\right)^t \|J_{\pi^0} - J^*\|_\infty + \epsilon.$$

10 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 were 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. Karaesmen and Van Ryzin, 2004, L'Ecuyer and Glynn, 1994, Van Ryzin and Vulcano, 2008], 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 Rajeswaran et al. [2017], or using state aggregation Bertsekas [2019], Ferns et al. [2004], Singh et al. [1995] are preferred. Here we conclude 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 F.

Given an expressive policy class Π_Θ,

$$\inf_{\pi \in \Pi_\Theta} \|T_\pi J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\eta_{\pi_\theta}} \tag{20}$$

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 (20) 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 5 bounds the optimiality gap at a stationary point by the approximation error in (20). Our result in Theorem 5 is reminiscent of results in the study of approximate policy iteration methods, pioneered by Antos et al. [2008], Bertsekas [2011], Bertsekas and Tsitsiklis [1996], Munos [2003], Munos and Szepesvári [2008], 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 Kakade and Langford [2002] 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 wieghted under the initial distribution ρ.

Theorem 5. *Suppose Condition 2 holds. Then, if θ is a stationary point of $\ell(\cdot)$,*

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

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 6 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 $T_\pi J_{\pi_\theta} \succeq T J_{\pi_\theta}$ 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 (14) and the second inequality uses that $\eta_{\pi_\theta} \geq (1-\gamma)\rho$ (by definition, see (4)).

□

11 Notation

Table 1: Table of Notation

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 (9)
κ_ρ	$\triangleq$	Effective concentrability coefficient described in Section 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

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM, 2017.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

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

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Dimitir P Bertsekas and Steven Shreve. *Stochastic optimal control: the discrete-time case*. 2004.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Dimitri P Bertsekas. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335, 2011.

Dimitri P Bertsekas. Feature-based aggregation and deep reinforcement learning: A survey and some new implementations. *IEEE/CAA Journal of Automatica Sinica*, 6(1):1–31, 2019.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific Belmont, MA, 1996.

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.

Lawrence C Evans. An introduction to mathematical optimal control theory. *Lecture Notes, University of California, Department of Mathematics, Berkeley*, 2005.

Yuguang Fang, Kenneth A Loparo, and Xiangbo Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pages 568–576, 2010.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 162–169. AUAI Press, 2004.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Matthieu Geist, Bilal Piot, and Olivier Pietquin. Is the bellman residual a bad proxy? In *Advances in Neural Information Processing Systems*, pages 3205–3214, 2017.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine Learning Proceedings 1995*, pages 261–268. Elsevier, 1995.

Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert 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)*, 42(6):1291–1307, 2012.

Ronald A Howard. Dynamic programming and markov processes. 1960.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

Itir Karaesmen and Garrett Van Ryzin. Overbooking with substitutable inventory classes. *Operations Research*, 52(1):83–104, 2004.

Hamed Karimi, Julie Nutini, and Mark 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*, pages 795–811. Springer, 2016.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

D. Kleinman. On an iterative technique for riccati equation computations. *IEEE Transactions on Automatic Control*, 13:114 – 115, 1968.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

Sumit Kunnumkal and Huseyin Topaloglu. Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems. *Operations Research*, 56(3):646–664, 2008.

Pierre L'Ecuyer and Peter W Glynn. Stochastic optimization by simulation: Convergence proofs for the gi/g/1 queue in steady-state. *Management Science*, 40(11):1562–1578, 1994.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1800–1809, 2018.

Peter Marbach and John N Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567, 2003.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

Ofir Nachum, Mohammad Norouzi, and Dale Schuurmans. Improving policy gradient by exploring under-appreciated rewards. *CoRR*, abs/1611.09321, 2017.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2012.

Ian Osband, Benjamin Van Roy, Daniel Russo, and Zheng Wen. Deep exploration via randomized value functions. *arXiv preprint arXiv:1703.07608*, 2017.

Anthony L Peressini, Francis E Sullivan, and J Jerry Uhl. *The mathematics of nonlinear programming*. Springer-Verlag New York, 1988.

Jan Peters and Stefan Schaal. Policy gradient methods for robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2219–2225. IEEE, 2006.

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Aravind Rajeswaran, Kendall Lowrey, Emanuel V Todorov, and Sham M Kakade. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems*, pages 6550–6561, 2017.

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016a.

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251. IEEE, 2016b.

Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153, 2016c.

Martin Riedmiller, Jan Peters, and Stefan 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*, pages 254–261. IEEE, 2007.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

John Rust. Using randomization to break the curse of dimensionality. *Econometrica: Journal of the Econometric Society*, pages 487–516, 1997.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Bruno Scherrer. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pages 1314–1322, 2014.

Bruno Scherrer and Matthieu 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*, pages 35–50. Springer, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *ICML*, 2014.

Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, pages 361–368, 1995.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, Feb 2017. ISSN 0018-9448. doi: 10.1109/TIT.2016.2632162.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Philip Thomas. Bias in natural actor-critic algorithms. In *International conference on machine learning*, pages 441–448, 2014.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1-3):59–94, 1996.

John N Tsitsiklis and Benjamin Van Roy. Regression methods for pricing complex american-style options. *IEEE Transactions on Neural Networks*, 12(4):694–703, 2001.

Benjamin Van Roy. Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2):234–244, 2006.

Garrett Van Ryzin and Gustavo Vulcano. Simulation-based optimization of virtual nesting controls for network revenue management. *Operations Research*, 56(4):865–880, 2008.

Ward Whitt. Approximations of dynamic programs, i. *Mathematics of Operations Research*, 3(3): 231–243, 1978.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

A Background on Bellman operators and policy iteration

We make repeated use of the basic element-wise inequalities

$$TJ \preceq T_\pi J \quad \text{and} \quad TJ_\pi \preceq J_\pi \tag{21}$$

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)). 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, $J, J' : \mathcal{S} \to \mathbb{R}$, both 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,

$$\|J - J'\|_\infty \leq \gamma \|J - J'\|_\infty \qquad\qquad \|T_\pi J - T_\pi J'\|_\infty \leq \gamma \|J - J'\|_\infty. \tag{22}$$

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,

$$\|J_\pi - J^*\|_\infty \leq \frac{1}{1-\alpha} \|J_\pi - TJ_\pi\|_\infty \tag{23}$$

Using the definitions: $J_\pi = T_\pi J_\pi$ and $J^* = TJ^*$,

$$\begin{aligned} \|J_\pi - J^*\|_\infty = \|T_\pi J_\pi - TJ_\pi + TJ_\pi - J^*\|_\infty &\leq \|T_\pi J_\pi - TJ_\pi\|_\infty + \|TJ_\pi - TJ^*\|_\infty \\ &\leq \|T_\pi J_\pi - TJ_\pi\|_\infty + \gamma \|J_\pi - J^*\|_\infty \end{aligned}$$

Equation (23) 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,

$$J - J_\pi = J - T_\pi J + T_\pi J - T_\pi J_\pi = J - T_\pi J + \alpha P_\pi (J - J_\pi) = \cdots = \sum_{t=0}^{\infty} \alpha^t P_\pi^t (J - T_\pi J). \tag{24}$$

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 usesful case of this result arises when $\pi = \pi^*$ is an optimal policy in which case,

$$J - J^* \preceq \sum_{t=0}^{\infty} \alpha^t P_{\pi^*}^t (J - TJ), \tag{25}$$

where the inequality uses (21) to conclude $TJ \preceq T_{\pi *} J$. Related inequalities are sometimes called the *Performance difference lemma* in the reinforcement learning literature, after crucial lemma of Kakade and Langford [2002].

The classic policy iteration algorithm due to Howard [1960] can be expressed compactly in terms of Bellman operators. Starting with an initial policy π, the algorithm first evalues 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 $T_{\pi^+} J_\pi = TJ_\pi$. One finds

$$J_\pi \succeq T_{\pi^+} J_\pi \succeq T_{\pi^+}^2 J_\pi \succeq \cdots \succeq J_{\pi^+}$$

where the first inequality applies (21) 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 $J_\pi = T_{\pi^+} J_\pi = T J_\pi$, in which case $J_\pi = J^*$ and π is an optimal policy. From Equation (24) or its more refined variant (25), we can see that each step of policy iteration leads to a substantial cost reduction unless the policy is near optimal. We conclude with a proof of a basic extension of Bellman's equation used in our analysis, which we restate here. Recall, η_π to be the discunted state-occupancy measure under policy π (see (4)).

Lemma 7 (On average Bellman equation)**.** *For any $\pi \in \Pi_\Theta$ and $S \sim \eta_\pi$,*

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

Proof. First note that standard results in dynamic programming imply $J_\pi \succeq T J_\pi$ and $J_\pi \succeq J^*$ (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 arbitray cost-to-go function such that $J \succeq 0$. 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 $S \sim \eta_\pi \succeq (1-\gamma)\rho$ and by assumption, $\rho(s) > 0$ for all $s \in \mathcal{S}$, $J(S) = 0$ 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 $J = J_\pi - T J_\pi$ shows the average Bellman equation above is equivalent to the standard result,

$$J_\pi = J* \iff J_\pi = T J_\pi.$$

□

B Background: First order methods.

As stated in Section 8, we consider the optimization problem of the form $\min_{x \in \mathcal{X}} f(x)$ where $\mathcal{X} \subset \mathbb{R}^d$ is a non-empty, closed convex set. In addition, we will assume $\mathcal{X}$ to be compact and the function f to be L-smooth over $\mathcal{X}$.

Definition 4 (L-smoothness)**.** *A function $f : \mathcal{D} \to \mathbb{R}$ is L-smooth over a set $\mathcal{X} \subseteq \mathcal{D}$ if it is differentiable over $\mathcal{X}$ and satisfies,*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \quad \forall\, (x, y) \in \mathcal{X}$$

where $L > 0$ is called the smoothness parameter.

A consequence of smoothness that will be useful throughout our proofs is the following *descent lemma* which implies a quadratic upper bound on function values.

Lemma 13 (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 $(x, y) \in \mathcal{X}$:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2.$$

B.1 Proof of Lemma 12

The following interpretation of projected gradient updates will be very useful for our proof. Recall that the projected gradient updates take the form:

$$x_{t+1} = \Pi_{\mathcal{X}}\left(x_t - \alpha_t \nabla f(x_t)\right) \tag{26}$$

where $\Pi_{\mathcal{X}}(\cdot)$ is the projection operator defined as $\Pi_{\mathcal{X}}(x) = \arg\min_{y\in\mathcal{X}} \|y - x\|_2^2$. Crucially, the update step (26) can be shown to be equivalent to the minimizer of a local quadratic approximation:

$$x_{t+1} = \arg\min_{x\in\mathcal{X}} \left[f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2\alpha_t}\|x - x_t\|_2^2 \right] \tag{27}$$

For convenience, we first restate Lemma 12 below.

Lemma 12 (Convergence rates for gradient dominated smooth functions)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set and f be L–smooth on $\mathcal{X}$. Consider the sequence $x_{t+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_t - \alpha\nabla f(x_t)\right)$ where $\alpha \le 1/L$. Set $f(x^*) = \min_{x'\in\mathcal{X}} f(x')$. Then,*

1. *If f is $(c, 0)$–gradient-dominated and $\|x - x'\|_2 \le R < \infty$ for all $x, x' \in \mathcal{X}$, then*

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

2. *If f is (c, μ)–gradient-dominated for $\mu > 0$, then,*

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

Proof of Lemma 12. Throughout, we assume a constant stepsize, $\alpha_t = \alpha \le \frac{1}{L}$. Recall, by Definition 3 that a function f is defined to be (c, μ)–*gradient dominated* over $\mathcal{X}$ if there exists a constant $c > 0$ and $\mu \ge 0$ such that

$$f(x^*) \ge f(x) + \min_{y\in\mathcal{X}} \left[c\,\langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2 \right] \qquad \forall\, x \in \mathcal{X}.$$

Proof of Part (a): We first assume $\mu = 0$ in which case for any $x \in \mathcal{X}$, we have

$$\min_{y\in\mathcal{X}} \left[c\,\langle \nabla f(x), y - x\rangle\right] \le f(x^*) - f(x) \tag{28}$$

Therefore, for any $x \ne x^*$, 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 $\delta_t \le 0$). Then,

$$\begin{aligned}
f(x_{t+1}) - f(x_t) &\overset{(a)}{\le} \min_{y\in\mathcal{D}} \left[\langle \nabla f(x_t), y - x_t\rangle + \frac{1}{2\alpha}\|y - x_t\|_2^2 \right] \\
&\overset{(b)}{=} \min_{\beta\in[0,1]} \left[\langle \nabla f(x_t), x_t + \beta(\bar{y} - x_t) - x_t\rangle + \frac{1}{2\alpha}\|x_t + \beta(\bar{y} - x_t) - x_t\|_2^2 \right] \\
&= \min_{\beta\in[0,1]} \left[\beta\langle \nabla f(x_t), (\bar{y} - x_t)\rangle + \frac{\beta^2}{2\alpha}\|\bar{y} - x_t\|_2^2 \right] \\
&\le \min_{\beta\in[0,1]} \left[\beta\delta_t + \frac{\beta^2 R^2}{2\alpha} \right] = \frac{-\alpha\delta_t^2}{2R^2}
\end{aligned}$$

where (a) follows by using the equivalence shown in (27) 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 (28), we get

$$f(x_{t+1}) - f(x_t) \leq \frac{-\alpha}{2R^2 c} \left(f(x^*) - f(x_t)\right)^2$$

Rearraning, we get our desired result

$$\min_{t \leq T} \left(f(x_t) - f(x^*)\right)^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \left(f(x_t) - f(x^*)\right)^2 \leq \frac{2R^2 c}{\alpha T} \sum_{t=0}^{T-1} f(x_t) - f(x_{t+1}) \leq \frac{2R^2 c}{\alpha T} \left(f(x_0) - f(x_T)\right)$$
$$\leq \frac{2R^2 c}{\alpha T} \left(f(x_0) - f(x^*)\right)$$

Therefore,

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

Proof of Part (b): Assuming $f(\cdot)$ is (c, μ)–gradient dominated for $c, \mu > 0$, by definition we have that for any $x \neq x^*$

$$\min_{y \in \mathcal{D}} \left[c \left\langle \nabla f(x_t), y - x \right\rangle + \frac{\mu}{2} \|y - x\|_2^2\right] \leq f(x^*) - f(x) < 0$$

Therefore, $\min_{y \in \mathcal{D}} \left\langle \nabla f(x_t), y - x \right\rangle < 0$. The following simple argument proves our desired result.

$$\begin{aligned} f(x^*) - f(x_t) &\geq \min_{y \in \mathcal{D}} \left[c \left\langle \nabla f(x_t), y - x_t \right\rangle + \frac{\mu}{2} \|y - x_t\|_2^2\right] \\ &\overset{(a)}{=} \left(\frac{c^2}{\mu \alpha}\right) \cdot \min_{y \in \mathcal{D}} \left[\left\langle \nabla f(x), y - x_t \right\rangle + \frac{1}{2\alpha} \|y - x_t\|_2^2\right] \\ &\overset{(b)}{\geq} \left(\frac{c^2}{\mu \alpha}\right) \left(f(x_{t+1}) - f(x_t)\right) \end{aligned}$$

Rearranging, we get:

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

Inequality (b) follows by interpreting the projected gradient update as the minimizer of a quadratic approximation (see Equation (27)). For (a), we use the following argument. At iterate x_t, let $\bar{y} = \arg\min_{y \in \mathcal{D}} \left\langle \nabla f(x_t), y - x_t \right\rangle$ and denote $\delta_t = \left\langle \nabla f(x_t), \bar{y} - x_t \right\rangle$. Then,

$$\begin{aligned} \min_{y \in \mathcal{D}} \left[c \left\langle \nabla f(x_t), y - x_t \right\rangle + \frac{\mu}{2} \|y - x_t\|_2^2\right] &= \min_{\beta \in [0,1]} \left[c \left\langle \nabla f(x_t), x_t + \beta(\bar{y} - x_t) - x_t \right\rangle + \frac{\mu}{2} \|x_t + \beta(\bar{y} - x_t) - x_t\|_2^2\right] \\ &= \min_{\beta \in [0,1]} \left[c\beta \left\langle \nabla f(x_t), \bar{y} - x_t \right\rangle + \frac{\mu \beta^2}{2} \|\bar{y} - x_t\|_2^2\right] \\ &= \min_{\beta \in [0,1]} \left[c\beta \delta_t + \frac{\mu \beta^2}{2} \|\bar{y} - x_t\|_2^2\right] \\ &= \frac{-c^2 \delta_t^2}{2\mu \|\bar{y} - x\|_2^2} \end{aligned}$$

where the first equality again uses the argument that minimizer of the left had side can be obtained by optimizing along the steepest descent direction, $x_t \to \bar{y}$. It also follows that for any $\alpha > 0$,

$$\min_{y \in \mathcal{D}} \left[\langle \nabla f(x), y - x \rangle + \frac{1}{2\alpha} \|y - x\|_2^2 \right] = \frac{-\alpha \delta_t^2}{2\|\bar{y} - x\|_2^2}$$

which proves (a). □

C Further details on LQ control

C.1 Proof of Lemma 3: Properties of the Bellman operator

Recall the statement of Lemma 3.

Lemma 3 (Bellman operators for LQ control)**.** *Consider the linear quadratic control problem formulated in Example 2. Fix the state-weighting $w(s) = 1/\|s\|_2^2$. 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. *(Contraction) $\|TJ - T\bar{J}\|_{\infty,w} \leq \gamma \|J - \bar{J}\|_{\infty,w}$ and $\|T_{\pi_\theta} J - T_{\pi_\theta} \bar{J}\|_{\infty,w} \leq \gamma \|J - \bar{J}\|_{\infty,w}$.*

Proof. These results can be found (or straighforwadly derived from) in any standard textbook on dynamic programming, for example Bertsekas [1995]. Details are only provided to aid the readers who might be unfamilier with some of the relevant material.

We prove each of the three claims in order. Consider any two cost-to-go functions $J, \bar{J} \in \mathcal{J}_q$ as:

$$J(s) = s^\top K s \quad \bar{J}(s) = s^\top \bar{K} s$$

defined for some $K, \bar{K} \succ 0$. Consider a stable linear policy $\pi_\theta(s) = \theta s$ for all $s \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^{k \times n}$. Recall the set-up for LQ control from Section 5.1 where for action $a \in \mathbb{R}^k$, the state evolution follows $s' = As + Ba$ and the per-period cost $g(s, a) = a^\top R a + s^\top C s$ for some cost matrices $R, C \succ 0$.

Part (1) By definition of the Bellman operator for policy π_θ, we have:

$$\begin{aligned}(T_{\pi_\theta} J)(s) &= (\theta s)^\top R(\theta s) + s^\top C s + \gamma J(As + B\theta s) \\ &= s^\top \left(\theta^\top R \theta + C + [A + B\theta]^\top K [A + B\theta] \right) s \end{aligned} \tag{29}$$

which simply follows by taking $a = \pi_\theta(s) = \theta s$. Note that $\theta^\top R \theta + C + [A + B\theta]^\top K [A + B\theta] \succ 0$ as $R, C, K \succ 0$ and so $T_{\pi_\theta} J \in \mathcal{J}_q$. Next, by definition,

$$TJ(s) = \min_{a \in \mathbb{R}^k} \left[a^\top R a + s^\top C s + \gamma (As + Ba)^\top K (As + Ba) \right] \tag{30}$$

Clearly, the minimizing action here is $a^* = -\gamma (R + \gamma B^\top K_\theta B)^{-1} B^\top K_\theta A s$ and therefore $TJ = T_{\pi_{\overline{\theta}}} J$ for $\pi_{\overline{\theta}}(s) = \overline{\theta} s$ with $\overline{\theta} = -\gamma (R + \gamma B^\top K B)^{-1} B^\top K_\theta A$. Using Equation (29), it is easy to see that $TJ \in \mathcal{J}_q$ as well.

Part (2) These inequalities can be checked immediately from the definitions of the Bellman operators. For $J \preceq \bar{J}$ and any s,

$$\begin{aligned} T_{\pi_\theta} J(s) &= (\theta s)^\top R(\theta s) + s^\top C s + \gamma J(As + B\theta s) \\ &\leq (\theta s)^\top R(\theta s) + s^\top C s + \gamma \bar{J}(As + B\theta s) = T_{\pi_\theta} \bar{J}(s) \end{aligned}$$

We can similarly show that $TJ(s) \leq T\bar{J}(s)$ for all s, by minimizing over $a \in \mathbb{R}^k$ on both sides of the following inequality which holds as $J \preceq \bar{J}$.

$$\left[a^\top R a + s^\top C s + \gamma J(As + Ba)\right] \leq \left[a^\top R a + s^\top C s + \gamma \bar{J}(As + Ba)\right]$$

Part (3) Let $\|Q\|_2 = \sup_{s \in \mathbb{R}^n} \frac{\|Qs\|_2}{\|s\|_2^2}$ denote the spectral norm for any matrix $Q \in \mathbb{R}^{n \times n}$. Then,

$$\|J - \bar{J}\|_{\infty,w} = \sup_{s \in \mathbb{R}^n} |J(s) - \bar{J}(s)| w(s) = \sup_{s \in \mathbb{R}^n} \frac{\left|s^\top \left(K - \bar{K}\right) s\right|}{\|s\|_2^2} = \left\|K - \bar{K}\right\|_2 .$$

From definition of the Bellman operator, we have

$$\begin{aligned} \|T_{\pi_\theta} J - T_{\pi_\theta} \bar{J}\|_{\infty,w} &= \sup_{s \in \mathbb{R}^n} \frac{1}{\|s\|_2^2} \cdot \gamma \left[J\left([A + B\theta]s\right) - \bar{J}\left([A + B\theta]s\right)\right] \\ &= \sup_{s \in \mathbb{R}^n} \frac{1}{\|s\|_2^2} \cdot \gamma s^\top [A + B\theta]^\top \left(K - \bar{K}\right) [A + B\theta] s \\ &= \gamma \left\|[A + B\theta] \left(K - \bar{K}\right) [A + B\theta]\right\|_2 \\ &\leq \gamma \|A + B\theta\|_2^2 \left\|K - \bar{K}\right\|_2 \\ &\leq \gamma \left\|K - \bar{K}\right\|_2 . \end{aligned}$$

which uses the fact that indcued operator norms are subultiplicative. The final inequality follows using $\|A + B\theta\|_2 \leq 1$ as we assumed π_θ to be a stable linear policy.

Given the contraction property for $T_{\pi_\theta}(\cdot)$ for any stable policy, $\theta \in \Theta_{\mathrm{S}}$, we show the contraction result for the Bellman optimality operator: $TJ = \min_{\theta \in \Theta_{\mathrm{S}}} T_{\pi_\theta} J \ \ \forall\, J \in \mathcal{J}_q$. Note that it is sufficient to search over the set of stable policies as we assumed the system (A, B) to be controllable. Starting with $J \in \mathcal{J}_q$ such that $\|J\|_{\infty,w}$ is finite, applying the Bellman operator can only decrease costs from every state. This implies,

$$\begin{aligned} \|TJ - TJ'\|_{\infty,w} = \sup_{s \in \mathbb{R}^n} \frac{1}{\|s\|^2} |TJ(s) - T\bar{J}(s)| &= \sup_{s \in \mathbb{R}^n} \frac{1}{\|s\|_2^2} |\min_{\theta \in \Theta_{\mathrm{S}}} T_{\pi_\theta} J(s) - \min_{\theta \in \Theta_{\mathrm{S}}} T_{\pi_\theta} \bar{J}(s)| \\ &\overset{(a)}{\leq} \sup_{s \in \mathbb{R}^n} \frac{1}{\|s\|_2^2} \max_{\theta \in \Theta_{\mathrm{S}}} |T_{\pi_\theta} J(s) - T_{\pi_\theta} \bar{J}(s)| \\ &\leq \max_{\theta \in \Theta_{\mathrm{S}}} \|T_{\pi_\theta} J - T_{\pi_\theta} \bar{J}\|_{\infty,w} \\ &\overset{(b)}{\leq} \gamma \left\|K - \bar{K}\right\|_2 = \|J - \bar{J}\|_{\infty,w}. \end{aligned}$$

where (b) follows from the contraction property of T_{π_θ} for any $\theta \in \Theta_{\mathrm{S}}$. Inequality (a) follows from the following short result. Consider any two functions f, g and a set $\mathbb{Z} \subseteq \mathrm{Dom}(f) \cap \mathrm{Dom}(g)$. Then,

$$\left|\min_{z_1 \in \mathbb{Z}} f(z_1) - \min_{z_2 \in \mathbb{Z}} g(z_2)\right| \leq \max_{z \in \mathbb{Z}} |f(z) - g(z)| .$$

To see this, note

$$\min_{z_1\in\mathbb{Z}} f(z_1) - \min_{z_2\in\mathbb{Z}} g(z_2) = \min_{z_1\in\mathbb{Z}} [f(z_1) + \max_{z_2\in\mathbb{Z}} -g(z_2)] = \min_{z_1\in\mathbb{Z}} \max_{z_2\in\mathbb{Z}} [f(z_1) - g(z_2)] \leq \max_{z\in\mathbb{Z}} [f(z) - g(z)]$$
$$\min_{z_2\in\mathbb{Z}} g(z_2) - \min_{z_1\in\mathbb{Z}} f(z_1) = \min_{z_2\in\mathbb{Z}} [g(z_2) + \max_{z_1\in\mathbb{Z}} -f(z_1)] = \min_{z_2\in\mathbb{Z}} \max_{z_1\in\mathbb{Z}} [g(z_2) - f(z_1)] \leq \max_{z\in\mathbb{Z}} [g(z) - f(z)].$$

□

C.2 Cost function for LQ control is smooth on sublevel sets

Lemma 14 (Cost function for LQ control)**.** *Let $\Sigma = \mathbb{E}_{s\sim\rho}\left[ss^\top\right] \succ 0$. Then, for the LQ control problem in Example 2 and any $\theta \in \Theta_S$,*

$$\sup\left\{\|\nabla^2\ell(\theta')\| : \ell(\theta') \leq \ell(\theta)\right\} < \infty$$

Let $\Sigma = \mathbb{E}_{s\sim\rho}\left[ss^\top\right] \succ 0$. Then, the total cost function for the LQ control problem in Example 2, $\ell(\theta) : \Theta_S \to \mathbb{R}$, is smooth in θ on sublevel sets.

Proof. To show that $\ell(\cdot)$ is smooth on sub-level sets, we argue that the sub-level sets of ℓ are compact[11] and that ℓ is infinitely differentiable over it. Recall that from Section 5.1 that we can wrtite the total cost function $\ell(\theta)$ as:

$$\ell(\theta) = \sum_{t=0}^{\infty}\gamma^t\left(s_t^\top\theta^\top R\theta s_t + s_t^\top C s_t\right) = \mathbb{E}_{s_0\sim\rho}\left\{s_0^\top\left[\sum_{t=0}^{\infty}\gamma^t\underbrace{\left((A+B\theta)^t\right)^\top\left(\theta^\top R\theta + C\right)(A+B\theta)^t}_{:=K_{\theta^t}}\right]s_0\right\}$$
$$= \sum_{t=0}^{\infty}\gamma^t\text{Trace}\left(K_{\theta^t}\Sigma\right)$$

which forms a power series in θ and is hence infinitely differentiable for $\theta \in \Theta_S$ (as the total cost is finite for stable policies).

Next, it is easy to show that sub-level sets of $\ell(\cdot)$ are compact by showing that they are closed and bounded. As ℓ is continuous (we argued above it is infinitely differentiable), by definition it's sub-level sets are closed. Also note that 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 $\Sigma_\theta := \mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t s_t s_t^\top\right]$ which implies, $\Sigma_\theta \succ \Sigma = \mathbb{E}_{s_0\sim\rho}\left[s_0 s_0^\top\right] \succ 0$. Therefore,

$$\ell(\theta) = \text{Trace}\left((\theta^\top R\theta + C)\Sigma_\theta\right) \geq \lambda_{\min}(\Sigma_\theta)\text{Trace}\left(\theta^\top R\theta + C\right) \geq \lambda_{\min}(\Sigma_\theta)\|\theta^\top R\theta + C\|_2$$
$$\Rightarrow\ \|\theta^\top R\theta\|_2 \leq \frac{\ell(\theta)}{\lambda_{\min}(\Sigma_\theta)} + \|C\|_2. \tag{31}$$

[11] By definition, sub-level sets of $\ell(\cdot)$ correspond to stable policies, $\theta \in \Theta_S$.

As $\|R\|_2, \|C\|_2$ is assumed to be bounded and $\lambda_{\min}(\Sigma_\theta)$ is uniformly lower bounded ($\lambda_{\min}(\Sigma_\theta) > \lambda_{\min}(\Sigma) > 0$), it is clear by (31) that $\lim_{\|\theta\|_2 \to \infty} \ell(\theta) = +\infty$. By definition, the sub-level sets of a coercive function are bounded (see Peressini et al. [1988] for example). This completes the argument to show compactness of sub-level sets of $\ell(\cdot)$.

To argue for smoothness, recall that 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 finite constant L. It is well known by Extreme Value Theorem that continuous real-valued functions on compact sets are bounded (see Rudin et al. [1964] for example). This completes our argument: as $\ell(\cdot)$ is infinitely differentiable and sub-level sets of $\ell(\cdot)$ are compact, $\|\nabla^2 \ell(\theta)\|$ is bounded on sub-level sets. □

D Proof of Theorem 4

We first restate the geometric convergence rate result in Theorem 4.

Theorem 4 (Geometric convergence)**.** *Suppose one of the first-order algorithms in subsection 9.2 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 step-sizes chosen by exact line search as in (19), then*

$$\|J_{\pi^t} - J^*\|_\infty \leq (1 - \min_{s \in \mathcal{S}} \rho(s)\,(1-\gamma)^t \,\|J_{\pi^0} - J^*\|_\infty.$$

(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) ***Constant stepsize natural policy gradient.*** *Fix any $\epsilon > 0$. Under NPG with stepsize sequence $\alpha_t \geq \frac{2\log(2)}{(1-\gamma)\epsilon}$,*

$$\|J_{\pi^t} - J^*\|_\infty \leq \left(\frac{1+\gamma}{2}\right)^t \|J_{\pi^0} - J^*\|_\infty + \epsilon.$$

Throughout the proof, we use some standard properties of the Bellman operator. For example, $J^* = TJ^*$ and that T, T_π are a contraction in $\|\cdot\|$ with modulus γ for any $\pi \in \Pi$. Refer to Section A for details. We denote π^t_+ to be the policy iteration update to any policy $\pi^t \in \Pi$.

Part (a): Proof for exact line-search: First, let us revisit the exact line search method applied for optimizing stepsizes as descibed in Section 9. At any iterate, π^t, let δ^t be a descent direction. Then exact linesearch solves for

$$\alpha^* = \arg\min_{\alpha \geq 0} \ell(\pi^t_\alpha) := \ell(\pi^t + \alpha\delta^t)$$

and updates to $\pi^{t+1} = \pi^t_{\alpha^*}$. The first-order algorithms we described in Section 9 differ in the choice of descent directions. For example, Frank-Wolfe method chooses $\delta_t = \left(\pi^t_+ - \pi^t\right)$ as the

descent directon while the projected gradient and natural policy gradient methods use $\nabla \ell(\pi^t)$ and its preconditioned (with the Fisher information matrix) counterpart respectively. We give a unified proof using a simple observation that updates for all the different first-order algorithms exactly equal the policy iteration update for some value of α (see Section 9). Therefore, π_+^t is a feasible update for the exact line search method.

Proof. Noting that $\pi_{\alpha^*}^t = \pi^{t+1}$, we have

$$\ell(\pi^{t+1}) = \|J_{\pi^{t+1}}\|_{1,\rho} \leq \|T_{\pi^{t+1}} J_{\pi^t}\|_{1,\rho} \leq \|T J_{\pi^t}\|_{1,\rho} \tag{32}$$

where the first equality follows by definition. To see the second inequality, observe that

$$T_{\pi^{t+1}} J_{\pi^t} \preceq J_{\pi^t}$$

Essentially, we cannot increase costs as $\pi^{t+1} = \pi^t$, corresponding to $\alpha = 0$, is a feasible update. Then, monotonicity of $T_{\pi^{t+1}}$ implies: $J_{\pi^t}(s) \geq T_{\pi^{t+1}} J_{\pi^t}(s) \geq T_{\pi^{t+1}}^2 J_{\pi^t}(s) \geq \ldots \geq J_{\pi^{t+1}}(s)$ using the definition $J_{\pi^{t+1}}(s) = \lim_{n\to\infty} T_{\pi^{t+1}}^n J_{\pi^t}(s)$. The final inequality follows as π_+^t is feasible update for all the four policy gradent methods with linesearch as explained above. Equation (32) gives us a lower bound on the progress made with line search. Denote $\rho_{\min} := \min_{s\in\mathcal{S}} \rho(s)$. Then,

$$\begin{aligned}
\ell(\pi^t) - \ell(\pi^{t+1}) \geq \|J_{\pi^t} - T J_{\pi^t}\|_{1,\rho} &\geq \rho_{\min} \|J_{\pi^t} - T J_{\pi^t}\|_\infty \\
&= \rho_{\min} \|J_{\pi^t} - J^* - (T J_{\pi^t} - J^*)\|_\infty \\
&\geq \rho_{\min} \left\{ \|J_{\pi^t} - J^*\|_\infty - \|T J_{\pi^t} - T J^*\|_\infty \right\} \\
&\geq \rho_{\min} (1-\gamma) \|J_{\pi^t} - J^*\|_\infty
\end{aligned} \tag{33}$$

where we use that $\|J\|_{1,\rho} \geq \min_{s\in\mathcal{S}} \rho(s) \|J\|_\infty$ along with $J^* = T J^*$ and contraction of T. It is easy to see,

$$\|J_{\pi^t} - J^*\|_\infty + \|J^* - J_{\pi^{t+1}}\|_\infty \geq \|J_{\pi^t} - J_{\pi^{t+1}}\|_\infty \geq \ell(\pi^t) - \ell(\pi^{t+1}) \tag{34}$$

We get our desired result combining (33) and (34) as well as using the fact that $J^* \preceq J_{\pi^{t+1}}$.

$$J_{\pi^{t+1}} - J^*\|_\infty \leq (1 - \rho_{\min}(1-\gamma)) \|J_{\pi^t} - J^*\|_\infty \ldots \leq (1 - \rho_{\min}(1-\gamma))^{t+1} \|J_{\pi^0} - J^*\|_\infty$$

□

Part (b): Proof for constant stepsize Frank-Wolfe: Recall from Section 9, Frank-Wolfe exactly equals soft-policy iteration:

$$\pi^{t+1}(s) = (1-\alpha)\pi^t(s) + \alpha \pi_+^t(s)$$

where we denote π_+^t to be the policy iteration update to π^t. Note that starting from a feasbile policy $\pi^0 \in \Pi$, we always maintain feasibility for $\alpha \in (0, 1]$. Hence, for any state s

$$\begin{aligned}
T_{\pi^{t+1}} J_{\pi^t}(s) - T J_{\pi^t}(s) &= \langle \pi^{t+1}(s), Q^t(s,\cdot)\rangle - \langle \pi_+^t(s), Q^t(s,\cdot)\rangle \\
&= (1-\alpha)\langle \pi^t(s), Q^t(s,\cdot)\rangle - (1-\alpha)\langle \pi_+^t(s), Q^t(s,\cdot)\rangle \\
&= (1-\alpha)\left[J_{\pi^t}(s) - T J_{\pi^t}(s)\right]
\end{aligned}$$

Using $TJ_{\pi^t} \preceq J_{\pi^t}$, we also get

$$T_{\pi^{t+1}} J_{\pi^t}(s) = (1-\alpha) J_{\pi^t}(s) + \alpha T J_{\pi^t}(s) \leq J_{\pi^t}(s)$$

Monotonicty of $T_{\pi^{t+1}}$ implies: $J_{\pi^t}(s) \geq T_{\pi^{t+1}} J_{\pi^t}(s) \geq T^2_{\pi^{t+1}} J_{\pi^t}(s) \geq \ldots \geq J_{\pi^{t+1}}(s)$ using the definition $J_{\pi^{t+1}}(s) = \lim_{n\to\infty} T^n_{\pi^{t+1}} J_{\pi^t}(s)$. Therefore,

$$J_{\pi^{t+1}}(s) - T J_{\pi^t}(s) \leq T_{\pi^{t+1}} J_{\pi^t}(s) - T J_{\pi^t}(s) = (1-\alpha)\left[J_{\pi^t}(s) - T J_{\pi^t}(s)\right]$$

Subtracting $J^*(s)$ from both sides and rearranging terms,

$$J_{\pi^{t+1}}(s) - J^*(s) \leq (1-\alpha)\left(J_{\pi^t}(s) - J^*(s)\right) + \alpha\left(T J_{\pi^t}(s) - J^*(s)\right)$$

The above inequality holds for any s, therefore

$$\|J_{\pi^{t+1}} - J^*\|_\infty \leq (1-\alpha)\|J_{\pi^t} - J^*\|_\infty + \alpha\|TJ_{\pi^t} - J^*\|_\infty \leq \left[(1-\alpha) + \gamma\alpha\right]\|J_{\pi^t} - J^*\|_\infty$$

where we used that $J^* = TJ^*$ and $\|TJ_{\pi^t} - TJ^*\|_\infty \leq \gamma\|J_{\pi^t} - J^*\|_\infty$ (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 constant stepsize natural policy gradient: Recall that for $\theta \in \mathbb{R}^{n\times k}$, the softmax policy parameterization takes action i in state s with probability $\pi_\theta(i|s)$:

$$\pi_\theta(i|s) = \frac{e^{\theta_{s,i}}}{\sum_{j=1}^k e^{\theta_{s,j}}} \qquad i = 1, \ldots, k.$$

As shown in Section 9, the natural policy gradient (NPG) updates with a constant stepsize α take the simple form:

$$\pi^{t+1}(i|s) = \frac{\pi^t(i|s)\cdot e^{-\alpha Q^t(i,s)}}{\sum_{j=1}^k \pi^t(j|s)\cdot e^{-\alpha Q^t(j,s)}}, \tag{35}$$

where we use the shorthand notation $\pi^t(\cdot)$ to denote $\pi_{\theta^t}(\cdot)$ and $Q^t(s,i)$ to denote $Q_{\pi_{\theta^t}}(s,i)$.

Our proof strategy follows essentially shows that for any state $s \in \mathcal{S}$, the NPG update decrease 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[12] 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

[12]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.

implies that action 1 is optimal in state s under policy π^t. For any $c \in (0,1)$, define $O_t^-(s)$ and $O_t^+(s)$ as:

$$O_t^-(s) := \left\{ i \mid Q^t(s,i) - Q^t(s,1) \geq \frac{\log(1/c)}{\alpha} \right\}$$
$$O_t^+(s) := \left\{ i \mid Q^t(s,i) - Q^t(s,1) \leq \frac{\log(1/c)}{\alpha} \right\}$$

For simplicity, we will take $c = \frac{1}{2}$ although all our results hold for any $c \in (0,1)$. Denote $\epsilon := \frac{\log 2}{\alpha}$. 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 the policy π^t. The following lemma shows that the NPG updates decrease the probability of playing sub-optimal actions by a multiplicative factor.

Lemma 15. *For any state s, $\frac{\pi^{t+1}(i|s)}{\pi^t(i|s)} \leq \frac{1}{2} \;\; \forall\, i \in O_t^-(s)$.*

Proof. The proof follows a simple argument. By definition, for any $i \in O_t^-(s)$:

$$\begin{aligned} & \alpha \left(Q^t(s,i) - Q^t(s,1) \right) \geq \log(2) \\ \Rightarrow \; & \alpha \left(Q^t(s,i) - Q^t(s,1) \right) \geq \log(2) - \log\big(\pi^t(1|s)\big) \end{aligned}$$

which follows as $\pi^t(1|s) \in (0,1)$. Rearranging, we get

$$\log\left(\pi^t(1|s) e^{-\alpha Q^t(s,1)}\right) + \log\left(\frac{1}{2}\right) \geq -\alpha Q^t(s,i)$$

This implies,

$$\log\left(\sum_{j=1}^{k} \pi^t(j|s) e^{-\alpha Q^t(s,j)}\right) + \log\left(\frac{1}{2}\right) \geq \log\left(\pi^t(1|s) e^{-\alpha Q^t(s,1)}\right) + \log\left(\frac{1}{2}\right) \geq -\alpha Q^t(s,i).$$

which holds as all the terms in the summation are positive, $\pi^t(j|s) e^{-\alpha Q^t(s,j)} > 0 \;\; \forall\, j \in \{1,2,\ldots,k\}$ and $\log(\cdot)$ is a monotonic transformation. Our result holds by noting

$$\frac{1}{2}\left(\sum_{j=1}^{k} \pi^t(j|s) e^{-\alpha Q^t(s,j)}\right) \geq e^{-\alpha Q^t(s,i)} \quad \Rightarrow \quad \frac{\pi^{t+1}(i|s)}{\pi^t(i|s)} = \frac{e^{-\alpha Q^t(s,i)}}{\sum_{j=1}^{k} \pi^t(j|s) e^{-\alpha Q^t(s,j)}} \leq \frac{1}{2}.$$

□

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 15 shows how the NPG update decays the probabilities of *sub-optimal* actions (in the set $O_t^-(s)$) by a multiplicative factor instead of zeroing them out[13]. Thus intuitively, the NPG update resemble 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 the policy iteration update.

[13]This defintion of sub-optimal actions based on action gap threshold, $\epsilon = \log(2)/\alpha$, is essentially an artifact that we are taking gradient steps with finite stepsizes. As $\alpha \to \infty$, the threshold $\epsilon \to 0$ making the NPG update equivalent to a soft-policy iteration update.

Lemma 16 (Progress quantification)**.** *Let $J_{\pi^t}(s)$ denote the cost-to-go function for policy π^t from any starting state s. Then,*

$$T_{\pi^{t+1}} J_{\pi^t}(s) - J_{\pi^t}(s) \leq \frac{1}{2} \cdot (T J_{\pi^t}(s) - J_{\pi^t}(s)) + \epsilon$$

Proof. Recall, we assumed that: $Q^t(s,1) < Q^t(2) \ldots < Q^t(s,k)$ which implies that the policy iteration update, π_t^+ puts the entire mass on action 1. That is, $\pi_+^t(1|s) = 1$ and $\pi_+^t(i|s) = 0 \ \ \forall \ i \neq 1$. Consider,

$$\begin{aligned}
T_{\pi^{t+1}} J_{\pi^t}(s) - T J_{\pi^t}(s) &= \langle \pi^{t+1}(\cdot|s) - \pi_t^+(\cdot|s), Q^t(s,\cdot) \rangle \\
&= (\pi^{t+1}(1|s) - 1) Q^t(s,1) + \sum_{j=2}^{k} \pi^{t+1}(j|s) Q^t(s,j) \\
&= -\sum_{j=2}^{k} \pi^{t+1}(j|s) Q^t(s,1) + \sum_{j=2}^{k} \pi^{t+1}(j|s) Q^t(s,j) \\
&= \sum_{j=2}^{k} \pi^{t+1}(j|s) \left(Q^t(s,j) - Q^t(s,1) \right) \\
&= \sum_{j \in \mathcal{O}_t^-} \pi^{t+1}(j|s) \left(Q^t(s,j) - Q^t(s,1) \right) + \sum_{j \in \mathcal{O}_t^+} \pi^{t+1}(j|s) \left(Q^t(s,j) - Q^t(s,1) \right) \\
&= \sum_{j \in \mathcal{O}_t^-} \frac{\pi^{t+1}(j|s)}{\pi^t(j|s)} \pi^t(j|s) \left(Q^t(s,j) - Q^t(s,1) \right) + \sum_{j \in \mathcal{O}_t^+} \pi^{t+1}(j|s) \underbrace{\left(Q^t(s,j) - Q^t(s,1) \right)}_{\leq \epsilon} \\
&\leq \frac{1}{2} \sum_{j \in \mathcal{O}_t^-} \pi^t(j|s) \left(Q^t(s,j) - Q^t(s,1) \right) + \epsilon \\
&\leq \frac{1}{2} \left(\sum_{j=2}^{k} \pi^t(j|s) (Q^t(s,j) - Q^t(s,1)) \right) + \epsilon \\
&= \frac{1}{2} \left(\sum_{j=2}^{k} \pi^t(j|s) Q^t(s,j) - \sum_{j=2}^{k} \pi^t(j|s) Q^t(s,1) \right) + \epsilon \\
&= \frac{1}{2} \left(\left(\pi^t(j|s) - 1 \right) Q^t(s,1) + \sum_{j=2}^{k} \pi^t(j|s) Q^t(s,j) \right) + \epsilon \\
&= \frac{1}{2} \langle \pi^t(\cdot|s) - \pi_t^+(\cdot|s), Q^t(s,\cdot) \rangle + \epsilon \\
&= \frac{1}{2} \left(J_{\pi^t}(s) - T J_{\pi^t}(s) \right) + \epsilon \qquad (36)
\end{aligned}$$

where we used that $\frac{\pi^{t+1}(j|s)}{\pi^t(j|s)} \leq \frac{1}{2} \ \ \forall \, j \in \mathcal{O}_t^-(s)$ as shown above in Lemma 15 along with the fact that $\left(Q^t(s,j) - Q^t(s,1) \right) \leq \epsilon \ \ \forall \, j \in \mathcal{O}_t^+(s)$ which follows by definition. Subtracting $J_{\pi^t}(s)$ from both sides in Equation (36) and rearranging terms gives our desired result

$$T_{\pi^{t+1}} J_{\pi^t}(s) - J_{\pi^t}(s) \leq \frac{1}{2} \cdot (T J_{\pi^t}(s) - J_{\pi^t}(s)) + \epsilon$$

□

Step 3: Completing the proof: Lemma 16 clearly quantifies the relationship between an NPG update with constant stepsize α and a soft policy iteration update with an additive error ϵ. With this connection, we give a simple proof of geometric convergence for the natural policy gradient method. We claim that $J_{\pi^{t+1}}(s) \leq J_{\pi^t}(s)$. To see this, note that for any stepsize α, the NPG update for state s can be equivalently written as:

$$\pi^{t+1}(s) = \arg\min_{a \in \Delta^{k-1}} \left[Q^t(s,a) + \frac{\eta_{\pi^t(s)}}{\alpha} D_{\mathrm{KL}}(a || \pi^t(s)) \right]$$

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 monotonicty property of $T_{\pi^{t+1}}$, we have $J_{\pi^t}(s) \geq T_{\pi^{t+1}} J_{\pi^t}(s) \geq T^2_{\pi^{t+1}} J_{\pi^t}(s) \geq \ldots \geq J_{\pi^{t+1}}(s)$ by noting that $J_{\pi^{t+1}}(s) = \lim_{n \to \infty} T^n_{\pi^{t+1}} J_{\pi^t}(s)$. From Lemma 16, we get

$$J_{\pi^{t+1}}(s) - J_{\pi^t}(s) \leq T_{\pi^{t+1}} J_{\pi^t}(s) - J_{\pi^t}(s) \leq \frac{1}{2} \cdot (T J_{\pi^t}(s) - J_{\pi^t}(s)) + \epsilon$$

Subtracting $J^*(s)$ from both sides and rearranging terms,

$$\begin{aligned} J_{\pi^{t+1}}(s) - J^*(s) &\leq \frac{1}{2} J_{\pi^t}(s) + \frac{1}{2} T J_{\pi^t}(s) - J^*(s) + \epsilon \\ &= \frac{1}{2} \left(J_{\pi^t}(s) - J^*(s) \right) + \frac{1}{2} \left(T J_{\pi^t}(s) - J^*(s) \right) + \epsilon \end{aligned}$$

which implies,

$$\begin{aligned} \| J_{\pi^{t+1}} - J^* \|_\infty &\leq \frac{1}{2} \| J_{\pi^t} - J^* \|_\infty + \frac{1}{2} \| T J_{\pi^t} - J^* \|_\infty + \epsilon \\ &\leq \left[\frac{1}{2} + \frac{\gamma}{2} \right] \| J_{\pi^t} - J^* \|_\infty + \epsilon \end{aligned}$$

where we used that $\| T J_{\pi^t} - J^* \|_\infty = \| T J_{\pi^t} - T J^* \|_\infty \leq \gamma \| J_{\pi^t} - J^* \|_\infty$ which follows from the contraction property of, $T(\cdot)$. 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.

$$\begin{aligned} \| J_{\pi^{t+1}} - J^* \|_\infty &\leq \left(1 - \frac{(1-\gamma)}{2} \right) \| J_{\pi^t} - J^* \|_\infty + \epsilon \\ &\leq \left(1 - \frac{(1-\gamma)}{2} \right)^t \| J_{\pi^0} - J^* \|_\infty + \epsilon \sum_{i=0}^{t-1} \left(1 - \frac{(1-\gamma)}{2} \right)^i \\ &\leq \left(1 - \frac{(1-\gamma)}{2} \right)^t \| J_{\pi^0} - J^* \|_\infty + \frac{2\epsilon}{(1-\gamma)} \end{aligned}$$

As $\alpha \to \infty$, constant error term $\epsilon \to 0$ and therefore NPG update is exactly the policy iteration step.

E Omitted Proofs

E.1 Proof of Lemma 5

For the reader's convenience, we restate the claim.

Lemma 5 (Policy gradient theorem)**.** *$\ell(\theta)$ is differentiable and*

$$\nabla \ell(\theta) = \nabla_{\overline{\theta}} \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})\Big|_{\overline{\theta}=\theta} = \sum_{s\in\mathcal{S}} \eta_{\pi_\theta}(s) \left[\nabla_{\overline{\theta}} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))\Big|_{\overline{\theta}=\theta}\right].$$

Proof. We have already shown

$$\begin{aligned}\ell(\overline{\theta}) &= \ell(\theta) + \eta_\theta \left(T_{\overline{\theta}} J_\theta - J_\theta\right) + \left[(\eta_{\overline{\theta}} - \eta_\theta)\left(T_{\overline{\theta}} J_\theta - T_\theta J_\theta\right)\right] \\ &= \ell(\theta) + \eta_\theta J_\theta + \underbrace{\sum_{s\in\mathcal{S}} \eta_\theta(s) Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))}_{(a)} + \underbrace{\left[(\eta_{\overline{\theta}} - \eta_\theta)\left(T_{\overline{\theta}} J_\theta - T_\theta J_\theta\right)\right]}_{(b)}\end{aligned}$$

where the second equality applies the basic relation between Q functions and Bellman operators, $(T_{\pi'} J_\pi)(s) = Q_\pi(s, \pi'(s))$. We show the gradient $\nabla \ell(\overline{\theta})|_{\overline{\theta}=\theta}$ exists and calculate it, treating separately the terms (a) and (b).

Step 1: *Recall the Leibniz rule:*
Recall the Leibniz rule states that, for any base measure μ and function f, $\int_{\mathcal{S}} f(\theta, s) d\mu(s)$ is continuously differentiable with respect to θ, with

$$\frac{\partial}{\partial \theta_i} \int_{\mathcal{S}} f(\theta, s) d\mu(s) = \int_{\mathcal{S}} \frac{\partial}{\partial \theta_i} f(\theta, s) d\mu(s),$$

provided $\frac{\partial}{\partial \theta_i} f(\theta, s)$ is continuous and integrable, meaning $\int_{\mathcal{S}} \left|\frac{\partial}{\partial \theta_i} f(\theta, s)\right| d\mu(s) < \infty$. Here we make repeated use of this fact when μ is either the counting measure or $\mu(s) = \eta_\theta(s)$.

Step 2: *$\frac{\partial}{\partial \overline{\theta}_i} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))$ is continuous in $\overline{\theta}$ and satisfies $\sup_{s\in\mathcal{S}, \overline{\theta}\in\Theta} \left|\frac{\partial}{\partial \overline{\theta}_i} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))\right|$*
Write,

$$Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) = g(s, \pi_{\overline{\theta}}(s)) + \sum_{s'\in\mathcal{S}} P(s'|s, \pi_{\overline{\theta}}(s)) J_\theta(s').$$

This first term, $g(s, \pi_{\overline{\theta}}(s))$, is easily seen to be continuously differentiable, since $g(s, a)$ is continuously differentiable with respect to a and $\pi_{\overline{\theta}}(s)$ is continuously differentiable with respect to $\overline{\theta}$. In addition,

$$\left|\frac{\partial}{\partial \overline{\theta}_i} g(s, \pi_{\overline{\theta}}(s))\right| = \left|\left\langle \frac{\partial}{\partial \overline{\theta}_i} \pi_{\overline{\theta}}(s)\,,\, \frac{\partial}{\partial a} g(s, a)\Big|_{a=\pi_{\overline{\theta}}(s)} \right\rangle\right| \le \sup_{s\in\mathcal{S}, \overline{\theta}\in\Theta} \left\|\frac{\partial}{\partial \overline{\theta}_i} \pi_{\overline{\theta}}(s)\right\|_2 \left\|\frac{\partial}{\partial a} g(s, a)\Big|_{a=\pi_{\overline{\theta}}(s)}\right\|_2 < \infty$$

is bounded uniformly by Assumption 1.

Now consider the second term. Observe that $\overline{\theta} \mapsto P(s'|a, \pi_{\overline{\theta}}(s))$ is also continuously differentiable by the chain rule. Its partial derivatives are integrable, since

$$\sum_{s'\in\mathcal{S}} \left|\frac{\partial}{\partial \overline{\theta}_i} P(s'|s, \pi_{\overline{\theta}}(s)) J_\theta(s')\right| \le \|J_\theta\|_\infty \sum_{s'\in\mathcal{S}} \left|\frac{\partial}{\partial \overline{\theta}_i} P(s'|s, \pi_{\overline{\theta}}(s))\right| < \infty$$

where Assumption 1 implies the infinite sum is finite. This shows $\overline{\theta} \mapsto \sum_{s' \in \mathcal{S}} P(s'|s, \pi_{\overline{\theta}}(s)) J_\theta(s')$ is continuously differentiable with

$$\frac{\partial}{\partial \overline{\theta}_i} \sum_{s' \in \mathcal{S}} P(s'|s, \pi_{\overline{\theta}}(s)) J_\theta(s') = \sum_{s' \in \mathcal{S}} \frac{\partial}{\partial \overline{\theta}_i} P(s'|s, \pi_{\overline{\theta}}(s)) J_\theta(s')$$

Assumption 1 shows these partial derivatives are uniformly bounded, with,

$$\sup_{s \in \mathcal{S}, \overline{\theta} \in \Theta} \left| \frac{\partial}{\partial \overline{\theta}_i} \sum_{s' \in \mathcal{S}} P(s'|s, \pi_{\overline{\theta}}(s)) J_\theta(s') \right| < \frac{\|g\|_\infty}{1-\gamma} \sup_{s \in \mathcal{S}, \overline{\theta} \in \Theta} \left| \sum_{s' \in \mathcal{S}} \frac{\partial}{\partial \overline{\theta}_i} P(s'|s, \pi_{\overline{\theta}}(s)) \right| < \infty,$$

where we used the bound $\|J_\theta\|_\infty \leq \|g\|_\infty / (1-\gamma)$.

Step 3 *Calculating the derivative of term (a).*
The result of step (2), together with the Leibniz rule implies we can exchange the summation and derivative, to find

$$\frac{\partial}{\partial \overline{\theta}_i} \sum_{s \in \mathcal{S}} \eta_\theta(s) Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) = \sum_{s \in \mathcal{S}} \eta_\theta(s) \frac{\partial}{\partial \overline{\theta}_i} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))$$

and this partial derivative is continuous in $\overline{\theta}$.

Step 4: *Show that* $\lim_{\overline{\theta} \to \theta} \left\| T_{\overline{\theta}} J_\theta - T_\theta J_\theta \right\|_\infty = 0$.
We have,

$$\left| T_{\overline{\theta}} J_\theta(s) - T_\theta J_\theta(s) \right| = \left| Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) - Q_{\pi_\theta}(s, \pi_\theta(s)) \right| \leq \|\overline{\theta} - \theta\| \sup_{s \in \mathcal{S}, \overline{\theta} \in \Theta} \left| \nabla_{\overline{\theta}} Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) \right| < \infty$$

where the derivative on the right hand side is bounded uniformly by Step 1.

Step 4: *Show that the operator $\eta \to \eta P_\theta$ is Lipschitz with respect to the maximum norm, meaning that there exists $c < \infty$ such that for any distribution η over $\mathcal{S}$ and any $\theta, \overline{\theta} \in \Theta$*

$$\|\eta P_\theta - \eta P_{\overline{\theta}}\|_\infty \leq c \|\overline{\theta} - \theta\|.$$

We have,

$$\begin{aligned}
\|\eta P_\theta - \eta P_{\overline{\theta}}\|_\infty &= \sup_{s' \in \mathcal{S}} \left| \sum_{s \in \mathcal{S}} \eta(s) \left(P(s'|s, \pi_{\overline{\theta}}(s) - P(s'|s, \pi_\theta(s)) \right) \right| \\
&\leq \sup_{s, s' \in \mathcal{S}} \left| P(s'|s, \pi_{\overline{\theta}}(s) - P(s'|s, \pi_\theta(s)) \right| \\
&= \sup_{s, s' \in \mathcal{S}} \left| P(s'|s, \pi_{\overline{\theta}}(s) - P(s'|s, \pi_\theta(s)) \right| \\
&\leq \underbrace{\left(\sup_{s, s' \in \mathcal{S}} \sup_{\tilde{\theta} \in \Theta} \|\nabla P(s'|s, \pi_{\tilde{\theta}}(s))\|_2 \right)}_{:= c} \|\overline{\theta} - \theta\|_2.
\end{aligned}$$

That $c < \infty$ is a direct consequence of Assumption 1. The final equality can be derived by Taylor's theorem, which ensures that for each $s, s' \in \mathcal{S}$ there exists $\tilde{\theta} \in \{t\theta + (1-t)\overline{\theta} \, : \, t \in [0,1]\}$ on the line segment joining θ and $\overline{\theta}$ such that

$$P(s'|s, \pi_{\overline{\theta}}(s) - P(s'|s, \pi_\theta(s)) = \langle \nabla_{\tilde{\theta}} P(s'|s, \pi_{\tilde{\theta}}(s)) \, , \, \overline{\theta} - \theta \rangle.$$

The right hand side is then bounded by Cauchy Swartz.

Step 5: *Treat term (b) by showing* $\nabla_{\overline{\theta}} \left[\left(\eta_{\overline{\theta}} - \eta_\theta \right) \left(T_{\overline{\theta}} J_\theta - T_\theta J_\theta \right) \right] \Big|_{\overline{\theta}=\theta} = 0$.

To show this, we show $\left\| \eta_{\overline{\theta}} - \eta_\theta \right\|_\infty = O(\left\| \overline{\theta} - \theta \right\|_2)$, which together with step 4 shows,

$$\lim_{\overline{\theta} \to \theta} \frac{\left\| \left(\eta_{\overline{\theta}} - \eta_\theta \right) \left(T_{\overline{\theta}} J_\theta - T_\theta J_\theta \right) \right\|_\infty}{\| \overline{\theta} - \theta \|_2} = 0,$$

implying our claim. We have

$$\eta_\theta = (1-\gamma)\rho + \gamma \eta_\theta P_\theta$$

That is, η_θ is the unique solution to the fixed point equation

$$\eta_\theta = F_\theta(\eta_\theta)$$

where $F_\theta(\eta) := (1-\gamma)\rho + \eta P_\theta$. Viewed as a left operator $\eta \mapsto \eta P_\theta$ is a mapping from ℓ_1 to ℓ_1 defined by

$$(\eta P_\theta(s')) = \sum_{s \in \mathcal{S}} \eta(s) P(s'|s, \pi_\theta(s)).$$

This shows $\|\eta P_\theta\|_\infty \leq \|\eta\|_\infty$, i.e. P_θ is a non-expansion in the maximum norm. From this, we have that F_θ is a contraction with modulus γ with respect to the maximum norm since

$$\|F_\theta(\eta') - F_\theta(\eta)\|_\infty = \gamma \|(\eta' - \eta) P_\theta\|_\infty \leq \gamma \|\eta' - \eta\|_\infty.$$

Contractivity implies,

$$\begin{aligned} \|\eta_{\overline{\theta}} - \eta_\theta\|_\infty &= \left\| F_{\overline{\theta}}\left(\eta_{\overline{\theta}}\right) - F_\theta\left(\eta_\theta\right) \right\|_\infty \\ &\leq \left\| F_{\overline{\theta}}\left(\eta_{\overline{\theta}}\right) - F_{\overline{\theta}}\left(\eta_\theta\right) \right\|_\infty + \left\| F_{\overline{\theta}}\left(\eta_\theta\right) - F_\theta\left(\eta_\theta\right) \right\|_\infty \\ &\leq \gamma \|\eta_{\overline{\theta}} - \eta_\theta\|_\infty + \left\| F_{\overline{\theta}}\left(\eta_\theta\right) - F_\theta\left(\eta_\theta\right) \right\|_\infty, \end{aligned}$$

which can be rewritten as

$$\|\eta_{\overline{\theta}} - \eta_\theta\|_\infty \leq \frac{1}{1-\gamma} \left\| F_{\overline{\theta}}\left(\eta_\theta\right) - F_\theta\left(\eta_\theta\right) \right\|_\infty \leq c \|\overline{\theta} - \theta\|$$

where the inequality follows from step 4. □

E.2 Proof of Theorem 2

For the reader's convenience, we restate Condition 3 and Theorem 2.

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 costless 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. For any fixed* $s \in \mathcal{S}_h$, $\pi_\theta(s)$ *depends only on* θ_h.

Theorem 2. *Suppose Conditions 3 and 4 hold. Further assume that* ρ *is supported over each* $\mathcal{S}_h$ *for* $h \leq H$. *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 $h \leq H$,

$$\frac{\partial}{\partial \theta_h} \ell(\theta) = 0 \iff \frac{\partial}{\partial \theta_h} \mathcal{B}(\theta | \eta_{\pi_\theta}, J_{\pi_\theta}) = 0$$

using the policy gradient theorem in lemma 5. By definition, $J_{\pi_\theta}(s) = J^*(s) = 0$ for $s \in \mathcal{S}_{H+1}$ (as $\mathcal{S}_{H+1}$ contains a single, costless 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,

$$\frac{\partial}{\partial \theta_H} \mathcal{B}(\theta | \eta_{\pi_\theta}, J_{\pi_\theta}) = 0 \Rightarrow \frac{\partial}{\partial \theta_H} \mathcal{B}(\theta | \eta_{\pi_\theta}, J^*) = 0$$

Hence, for $S \sim \eta_{\pi_\theta}$ such that $S \in \mathcal{S}_H$

$$\mathbb{E}[J_{\pi_\theta}(S)] = \min_{\bar{\theta}_H \in \Theta_H} \mathbb{E}[Q^*(S, \pi_{\bar{\theta}_H}(S))] = \mathbb{E}[Q^*(S, \pi_{\theta_H^*}(S))] = \mathbb{E}[J^*(S)]$$

where the first equality follows by assumption that $\theta \to \mathcal{B}(\theta | \eta_{\pi_\theta}, J^*)$ has no suboptimal stationary points and the second eqaulity uses the assumption that policy class Π_θ contains the optimal policy. As $\rho(s) > 0$ for all $s \in \mathcal{S}_H$, our desired resutlt 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,

$$Q_{\pi_\theta}(s, a) = g(s, a) + \gamma \sum_{s' \in \mathcal{S}_{h+1}} P(s' | s, a) J_{\pi_\theta}(s') = g(s, a) + \gamma \sum_{s' \in \mathcal{S}_{h+1}} P(s' | s, a) J^*(s') = Q^*(s, a),$$

Again, $\frac{\partial}{\partial \theta_h} \mathcal{B}(\theta | \eta_{\pi_\theta}, J_{\pi_\theta}) = 0$ (which holds as θ is a stationary point) implies $\frac{\partial}{\partial \theta_h} \mathcal{B}(\theta | \eta_{\pi_\theta}, J^*) = 0$. By exactly the same argument as above,

$$\mathbb{E}[J_{\pi_\theta}(S)] = \min_{\bar{\theta}_h \in \Theta_h} \mathbb{E}[Q^*(S, \pi_{\bar{\theta}_h}(S))] = \mathbb{E}[Q^*(S, \pi_{\theta_h^*}(S))] = \mathbb{E}[J^*(S)]$$

for any $S \sim \eta_{\pi_\theta}$ such that $S \in \mathcal{S}_h$. Our result follows by noting that $\rho(s) > 0$ for all $s \in \mathcal{S}_h$.

□

The following lemma shows how Condition 4 holds for the finite horizon inventory control problem described in Example 5.

Lemma 17. *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 η supported over S, the weighted policy iteration objective $\mathcal{B}(\theta | \eta, J^*)$ has no suboptimal stationary points.*

Proof. Let $Q^*(s,a)$ be the Q-function corresponding to the optimal policy. It can be easilly shown that $Q^*(s,a)$ is convex in a (using results in chapter 3 of Bertsekas [1995])[14]. We want to show that for any θ such that $\nabla\mathcal{B}(\theta|\eta, J^*) = 0$, the base-stock policy π_θ is optimal. For $S \in \mathcal{S}_h$, we have $\frac{\partial}{\partial\theta_h}\pi_\theta(S) = 0$ if $S > \theta_h$. Therefore,

$$\frac{\partial}{\partial\theta_h}\mathcal{B}(\theta|\eta, J^*) = \frac{\partial}{\partial\theta_h}\mathbb{E}_{s\sim\eta}\left[Q^*(s,\pi_\theta(s))\right] = \int_{S<\theta_h}\left[\frac{\partial}{\partial a}Q^*(S,a)\Big|_{a=\pi_\theta(s)}\frac{\partial}{\partial\theta_h}\pi_\theta(s)\right]\eta(S)$$

Note that $Q^*(S,a)$ is convex with a minimum at $a = \pi_{\theta_h^*}(S)$. Thus, $\frac{\partial}{\partial a}Q^*(S,a)\Big|_{a=\pi_\theta(S)} > 0$ for $\theta_h < \theta_h^*$ (as we are ordering more) and $\frac{\partial}{\partial a}Q^*(S,a)\Big|_{a=\pi_\theta(S)} < 0$ otherwise. Using this along with the fact that $\frac{\partial}{\partial\theta_h}\pi_\theta(S) = 1$ for $S < \theta_h$, we get

$$\frac{\partial}{\partial\theta_h}\mathcal{B}(\theta|\eta, J^*) = 0 \iff \frac{\partial}{\partial a}Q^*(S,a)\Big|_{a=\pi_\theta(S)} = 0 \quad \forall\, S < \theta_h \tag{37}$$

Convexity of Q^* along with (37) implies that $\pi_\theta(S)$ is optimal for $S < \theta_h$ which in turn implies that the thresholds must match, i.e. $\theta_h = \theta_h^*$ (as $\pi_\theta(S) > 0$ for $S < \theta_h$). □

E.3 Concentrability coefficients

Lemma 8. *Let π^* denote any optimal stationary policy. Then,*

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

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 of some policy. Let $J^* = J_{\pi^*}$, Then, the variational form of Bellman's inequality in (1) gives

$$J_\pi - J^* = (I - \gamma P_{\pi^*})^{-1}(J_\pi - T_{\pi^*}J_\pi) \preceq (I - \gamma P_{\pi^*})^{-1}(J_\pi - TJ_\pi)$$

Left multiplying by ρ, using the definition $\eta_{\pi^*} = (1-\gamma)\rho(I - \gamma P_{\pi^*})^{-1}$ (see Equation (4) in Section 3) and that $J_\pi - TJ_\pi \succeq 0$ and $J_\pi \succeq J^*$ gives the result:

$$\begin{aligned}\|J_\pi - J^*\|_{1,\rho} = \rho(J_\pi - J^*) \le \frac{1}{(1-\gamma)}\eta_{\pi^*}(J_\pi - TJ_\pi) &\le \frac{\left(\sup_{s\in\mathcal{S}}\frac{\eta_{\pi^*}(s)}{\rho(s)}\right)}{(1-\gamma)}\rho(J_\pi - TJ_\pi) \\ &= \frac{\left(\sup_{s\in\mathcal{S}}\frac{\eta_{\pi^*}(s)}{\rho(s)}\right)}{(1-\gamma)}\|J_\pi - TJ_\pi\|_{1,\rho}.\end{aligned}$$

□

Lemma 11 (Concentrability in LQ control)**.** *Consider the linear quadratic control problem in Example 2. Suppose $\Sigma = \mathbb{E}_{s\sim\rho}[ss^\top] \succ 0$. Then, $\kappa_\rho \le n \cdot \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$.*

[14] This follows as the optimal cost-to go function, $J^*(\cdot)$ and the per step costs (of ordering and holding/backlogging) are convex.

Proof. Recall that for the LQ control problem, we assume the system to be controllable and search in the space of stable linear policies, Π_{Θ_S}. The induced set of finite cost-to-go functions is

$$\mathcal{J}_\Theta = \mathcal{J}_q = \{J : J(s) = s^\top K s \ \forall\, s \in \mathcal{S}\}$$

for some $K \succ 0$. We showed that for $J \in \mathcal{J}_\Theta$, we have $J \in \mathcal{J}_q$, $TJ \in \mathcal{J}_q$, and of course $J^* \in \mathcal{J}_q$ (see Section 5.1 for details). Importantly, for any $J \in \mathcal{J}_q$, we also showed that the Bellman operator is a contraction in the weighted maximum norm, $\|J\|_{\infty,w} = \sup_{s\in\mathbb{R}^n} |J(s)|w(s)$ for $w(s) = \|s\|_2^2$ implying $\|J\|_{\infty,w} = \|K\|_2$ where $\|K\|_2$ is the spectral norm.

Fix $J \in \mathcal{J}_q$ with $J(s) = s^\top K s$ for some $K \succ 0$. Let $S \sim \rho$ and let $v_1, \cdots v_n$ denote an orthonormal basis of eigenvectors of $\Sigma = \mathbb{E}[SS^\top]$ and $\lambda_1 \le \cdots \le \lambda_n$ denote the corresponding eigenvalues. Our proof uses the following standard result in matrix algebra (see Fang et al. [1994] for example). For any two positive semidefinite symmetric matrices A, B:

$$\lambda_{\min}(A)\mathrm{Trace}(B) \le \mathrm{Trace}(AB) \le \lambda_{\max}(A)\mathrm{Trace}(B) \tag{38}$$

where $\lambda_{\min}(A), \lambda_{\max}(A)$ are the minimum and maximum eigenvalues of A respectively. We have

$$\|J\|_{1,\rho} = \mathbb{E}[S^\top K S] = \mathrm{Trace}(K\Sigma) \le \lambda_n(\Sigma)\mathrm{Trace}(K) \le n\lambda_n(\Sigma)\|K\|_2 = n\lambda_n(\Sigma)\|J\|_{\infty,w} \tag{39}$$

where the first inequality uses (38) and the second inequality uses that $\mathrm{Trace}(K) = \sum_{i=1}^n \lambda_i(K) \le n\lambda_n(K) = n\|K\|_2$. Similarly, in the reverse direction, we have

$$\|J\|_{1,\rho} = \mathbb{E}[S^\top K S] = \mathrm{Trace}(K\Sigma) \ge \lambda_1(\Sigma)\mathrm{Trace}(K) \ge \lambda_1(\Sigma)\|K\|_2 = \lambda_1(\Sigma)\|J\|_{\infty,w} \tag{40}$$

which againn uses (38) and the fact that $\mathrm{Trace}(K) = \sum_{i=1}^n \lambda_i(K) \ge \lambda_n(K) = \|K\|_2$. Equations (39) and (40) establish norm equivalence for $J \in \mathcal{J}_q$.

$$\lambda_1(\Sigma)\|J\|_{\infty,w} \le \|J\|_{1,\rho} \le n\lambda_n(\Sigma)\|J\|_{\infty,w}.$$

Using Lemma 9, we have $\kappa_\rho \le n \cdot \frac{\lambda_n(\Sigma)}{\lambda_1(\Sigma)}$ as desired. □

Lemma 10 (Concentrability in optimal stopping)**.** *Suppose $\mathcal{S} = \mathcal{S}_C \cup \{T\}$ consists of a finite set of continuing states $\mathcal{S}_C$ and terminal state T that is absorbing ($P(T|T,a) = 1$) and costless ($g(T,a) = 0$). There are two actions $\mathcal{A} = \{0,1\}$, stop ($a = 0$) and continue ($a = 1$). Consider the policy 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 $\rho = \mu$, $\kappa_\rho \le 1$.*

Proof. We show that the Bellman opetaror T is a contraction with modulus γ in $\|\cdot\|_{1,\mu}$. The proof follows immediately using Lemma 9. Note that for any scalars (x_1, x_2, y), we have $|\min\{y, x_1\} -$

$\min\{y, x_2\}| \leq |x_1 - x_2|$. Then,

$$\begin{aligned}
\|TJ - TJ'\|_{1,\mu} &= \sum_{s\in\mathcal{S}_C} \mu(s) \left|TJ(s) - TJ'(s)\right| \\
&= \sum_{s\in\mathcal{S}_C} \mu(s) \left|\min\left\{g(s,0)\,,\, \gamma \sum_{s'\in\mathcal{S}} P(s'|s,1)J(s')\right\} - \min\left\{g(s,0)\,,\, \gamma \sum_{s'\in\mathcal{S}} P(s'|s,1)J'(s')\right\}\right| \\
&\leq \gamma \sum_{s\in\mathcal{S}_C} \mu(s) \left|\sum_{s'\in\mathcal{S}} P(s'|s,1)\left(J(s') - J'(s')\right)\right| \\
&\leq \gamma \sum_{s\in\mathcal{S}_C} \mu(s) \sum_{s'\in\mathcal{S}} P(s'|s,1)\left|J(s') - J'(s')\right| \\
&= \gamma \sum_{s'\in\mathcal{S}_C} \mu(s')\left|J(s') - J'(s')\right| \\
&= \gamma \|J - J'\|_{1,\mu}.
\end{aligned}$$

For $\rho = \mu$, we have $C, c = 1$ in Lemma 9 implying $\kappa_\rho \leq 1$. □

F 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 Gordon [1995], Tsitsiklis and Van Roy [1996], Van Roy [2006]. It is common across several academic communities [e.g Rust, 1997, Whitt, 1978]. 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 Ortner and Ryabko, 2012]. 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 6 (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 eucidean 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 partion is effective such that the approximation error, $\inf_{\pi\in\Pi} \|T_{\pi'} - TJ\|_{1,\eta_\pi}$, is small for any cost-to-go function $J : \mathcal{S} \to \mathbb{R}$.*