Global Optimality Guarantees For Policy Gradient Methods

Jalaj Bhandari and Daniel Russo

Columbia University

Abstract

Policy gradients 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 standard dynamic programming techniques, policy gradient algorithms face non-convex optimization problems and are widely understood to converge only to a stationary point. This work identifies structural properties – shared by several classic control problems – that ensure the policy gradient objective function has no suboptimal stationary points despite being non-convex. When these conditions are strengthened, this objective satisfies a Polyak-lojasiewicz (gradient dominance) condition that yields convergence rates. We also provide bounds on the optimality gap of any stationary point when some of these conditions are relaxed.

Keywords: Reinforcement learning, policy gradient methods, policy iteration, dynamic programming, gradient dominance.

1 Introduction

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

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

Stochastic gradients can be generated by monte-carlo simulation, even in complex environments and with policies represented by deep neural networks [Schulman et al., 2015a,b, 2017]. This approach is especially appealing when one has an inductive bias about the form of an effective policy. For example, Glasserman and Tayur [1995] use gradient descent to optimize over simple structured policies in a realistic simulator of a multi-echelon inventory control problem. Direct policy search has been used in control problems in robotics [Peters and Schaal, 2006], manufacturing [Caramanis and Liberopoulos, 1992], arcade games [Schulman et al., 2015a], revenue management [Talluri and Van Ryzin, 2006, Section 3.5.1], ambulance redeployment [Maxwell et al., 2013], scheduling in queues [L'Ecuyer and Glynn, 1994, L'Ecuyer et al., 1994], and many other areas.

Unfortunately, while policy gradient methods can be applied to a very broad class of problems, it is not clear whether they adequately address even simple control problems solvable by classical methods. A key challenge is that the total cost $\ell(\cdot)$ is a non-convex function of the chosen policy. Typical of results concerning black-box optimization of non-convex functions, policy gradient methods are widely understood to converge asymptotically to a stationary point or a local minimum.

(a) Two state, two action MDP

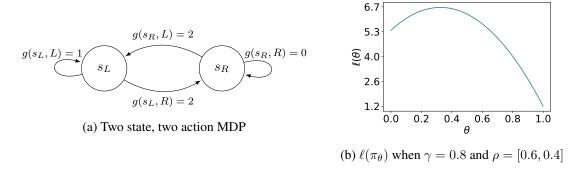(b) $\ell(\pi_\theta)$ when $\gamma = 0.8$ and $\rho = [0.6, 0.4]$

Figure 1: Presence of bad local minima with a constrained policy class.

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 literature seems to provide almost no guarantees into the *quality* of these stationary points. Worse yet, Example 1 shows that policy gradient methods can get stuck in a bad local minimum in very simple examples even when the policy class contains the optimal policy.

Example 1. *Consider the MDP depicted in Figure 1a. There are two states, left (s_L) and right (s_R), and two possible actions, L and R, which move the agent to the desired state in the next period. Staying in the state L incurs a cost $g(s_L, L) = 1$ per period, whereas staying in the right state is costless with $g(s_R, R) = 0$. Moving between states incurs a per-period cost of 2. When the discount factor exceeds $1/2$, it is easy to show that the optimal policy chooses action R in either state. In that case, it is reasonable to search in a constrained policy class, $\{\pi_\theta : \theta \in [0, 1]\}$ that plays action R with probability $\theta \in [0, 1]$ regardless of the current state. Setting $\theta = 1$ yields an optimal policy. Unfortunately, as shown in Figure 1, the total discounted cost incurred, $\ell(\pi_\theta)$, is a nonconvex function of θ. When initialized with small value of θ, cost is locally increasing as a function of θ, and so a gradient method updates the policy toward a bad local minimum at $\theta = 0$. Once there, any local policy search approach gets stuck as there are no descent directions that reduce cost. It is worth noting here that in general, policy gradient methods face many additional challenges, for instance due to unsophisticated exploration or policy parameterization. This example instead highlights the risk of bad local minima due to the non-convexity of the infinite horizon cost function $\ell(\cdot)$.*

In marked contrast to the example above, important recent work of Fazel et al. [2018] showed that for the deterministic linear quadratic control problem, policy gradient with the class of linear policies converges to the global optimum, despite non-convexity of the objective. Here, the authors provided an intricate analysis, leveraging a variety of closed form expressions available for linear-quadratic problems. Separately, in the operations literature, Kunnumkal and Topaloglu [2008] propose a stochastic approximation method for setting base-stock levels in inventory control. In this example too, the objective is non-convex, but the authors establish convergence to the globally optimal solution using an intricate analysis quite different than that of Fazel et al. [2018]. How do we reconcile these success stories with the simple counterexample given in Example 1?

2

1.1 Our Contribution

Policy gradient methods aim to directly minimize the *multi-period* total discounted cost by applying first-order optimization methods. By contrast, classical dynamic programming methods, like value and policy iteration, indirectly minimize the total cost by solving a sequence of simpler *single period* problems. We uncover an *indirect* analysis of policy gradient methods, deducing global convergence properties from conditions on the single period problems solved by policy iteration.

As a consequence of our general framework, we show that for several classic control problems, policy gradient methods performed with respect to natural structured policy classes face 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:

- Finite state and action MDPs with the class of all stochastic policies.

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

- An optimal stopping problem with the class of threshold policies.

- A finite horizon inventory control problems with the class of non-stationary base-stock policies.

These canonical control problems provide an important benchmark and sanity check. But why do policy gradient methods avoid suboptimal local minima in these cases as opposed to the simple case in Example 1? Interestingly, the examples above share some important structural properties. Consider the following properties of the LQ control example:

1. The policy class is *closed* under policy improvement. That is, starting with a linear policy and performing a policy iteration step yields another linear policy.

2. The single period optimization problem defining a policy iteration update has no suboptimal stationary points. In particular, the objective is a simple convex quadratic that can be easily optimized using first-order methods.

These same properties hold for finite MDPs as well as the optimal stopping problem. Our result in Theorem 1 shows that these two properties, together with mild regularity conditions, imply that any stationary point of the policy gradient loss function is globally optimal.

We remark that the closure condition is much weaker than requiring the policy class to contain all possible policies. This is useful, for example in problems where simple structured policy classes may be naturally aligned with the problem objective. However, the closure property is stronger than only requiring the policy class to contain (near) optimal policies. The presence of bad local minima in Example 1 illustrates why such a condition is necessary. In that case, the policy class contains the optimal policy, but it is not closed under policy improvement (see Section 5.1).

We extend these results in several ways. Theorem 2 studies a strengthening of the second condition above that leads to fast converge rates. In particular, when the single period objective defined by a (weighted) policy iteration problem satisfies a Polyak-lojasiewicz (PL) condition (also popularly known as "gradient dominance"), this property is inherited by the policy gradient objective

---

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

$\ell(\cdot)$. The PL conditions are relaxations of (strong) convexity that guarantee fast global convergence of first-order methods even for non-convex objectives [Nesterov and Polyak, 2006, Polyak, 1963].

Next, in Theorem 3, we show that for finite horizon problems with non-stationary policy classes, like the finite horizon inventory control problem, we only require that the policy class contains an optimal policy instead of the closure condition. In addition, we only require a weaker version of condition 2 above. See Section 5 for more details. Finally, Theorem 5, studies a weakening of the closure condition more generally. It assumes that the policy class is *approximately* closed under policy improvement – meaning that the policy iteration update can be solved in the given policy class up to a small error. In this case, we show how bound the optimality gap of any stationary point. Many of our intermediate results may also be of interest, especially our approach to concentrability coefficients in Section 7 and the corresponding bounds in Theorem 4.

1.2 Limited scope of this work

This paper is focused on understanding the optimization landscape of the non-convex policy gradient objective $\ell(\cdot)$, which is a fundamental challenge for local policy search methods. Such an investigation is simultaneously relevant to many strategies for searching locally over the policy space, including policy gradient methods [Sutton et al., 2000], natural gradient methods [Kakade, 2002], finite difference schemes [Riedmiller et al., 2007], evolutionary strategies [Salimans et al., 2017] etc.

However, we do sidestep many other issues that could lead to poor performance for specific local policy search methods. One such notable issue is that of *exploration*. It is well known that the naive random exploration provided by stochastic policies may be insufficient to guarantee convergence to an optimal policy. See Kakade and Langford [2002] or our discussion in Appendix A for examples. We imagine that we have access to a simulator where a restart distribution ensures that important regions of the state space are visited under all policies (see Assumption 1).

Another issue involves the choice of *policy parameterization*. In particular, for the popular softmax parameterization, the Jacobian matrix becomes ill conditioned near corners of the probability simplex, making a policy nearly insensitive to changes in the parameter space. Successful policy gradient methods must perform a local change of variables, like the popular natural policy gradient methods proposed by Kakade [2002]. See Section 6 for a detailed discussion on softmax policies. In addition, we do not address the choice of a *gradient estimator*, which is the subject of a large literature on stochastic simulation and reinforcement learning. See Sutton and Barto [2018], Mohamed et al. [2020], Glasserman and Ho [1991] or Fu [2006] for reviews from different perspectives.

For the rest of this paper, we will consider an idealized policy gradient update with access to exact gradient evaluations, $\theta_{k+1} = \theta_k - \alpha_k \nabla \ell(\theta_k)$. Generalizations to treat stochastic noise in gradient evaluations are possible using classical results from stochastic approximation literature which show that under regularity conditions and appropriately decaying step-sizes, most noisy iterative algorithms converge to the same limit as their deterministic counterparts [see e.g. Bertsekas and Tsitsiklis, 1996, Borkar, 2009]. We do not pursue such extensions for brevity.

1.3 Related Literature

Prior work on analysis of policy gradient methods. Apart from the aforementioned works of Kunnumkal and Topaloglu [2008] and Fazel et al. [2018], there has been limited prior work in theoretical guarantees for policy gradient methods, especially beyond tabular MDPs. See Agarwal

et al. [2020] for a detailed literature review of results for the tabular setting. One notable exception is the work by Scherrer and Geist [2014] which provides guarantees on the quality of (approximate) stationary points obtained by local policy search algorithms, such as Conservative Policy Iteration (CPI) [Kakade and Langford, 2002], Policy Search by Dynamic Programming (PSDP) [Bagnell et al., 2004], with a *convex policy class*. Relative to that work, our result in Theorem 1 is more general as it applies to problem settings with deterministic policies, infinite action spaces, structured cost functions, and most importantly parameterized policy classes that are not convex, such as the class of threshold or base-stock policies.

Concurrent work.   Concurrently with this work, Agarwal et al. [2020] provide a detailed study of policy gradient methods, primarily focusing on convergence rates in the tabular case for specific algorithms and with different policy parameterizations. They also extend their insights to the function approximation setting, focusing on natural gradient descent with a restricted class of log-linear policies in finite action spaces. Along similar lines, Shani et al. [2020] also provide convergence rates for trust-region based policy optimization methods for regularized tabular MDPs using ideas from analysis of the classic mirror descent method [Beck and Teboulle, 2003]. Although we do not analyze specific algorithms, intellectually we view regularized MDPs as satisfying a second-order gradient dominance condition, for which basic results in optimization theory imply fast convergence rates with first-order methods. See Section 3 for details. This highlights the strength of our general approach as it applies to a range of different problem settings. Finally, a recent paper by Wang et al. [2019] studies global convergence properties of actor-critic methods in the compatible function approximation setting of Sutton et al. [2000] with over-parameterized two layer neural networks, a regime in which the linearization error is bounded and neural networks essentially behave like kernel functions [Hofmann et al., 2008]. Complimentary to these works, we focus primarily on novel insights into when, why, and how policy search methods succeed with first-order algorithms. Our results apply to classic control problems, including those with continuous state and action spaces, and structured classes of deterministic policies.

On non-convex optimization in machine learning.   Beyond reinforcement learning, our work connects to an emerging body of literature on first-order methods for non-convex optimization problems, giving rates of convergence to stationary-points for first-order methods [Carmon et al., 2018, 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] and ensuring convergence to approximate local minima rather than saddle points [Agarwal et al., 2017, Jin et al., 2017, Lee et al., 2016]. A complementary line of research studies the optimization landscape of specific problems to essentially ensure that any local minima is globally optimal [Bhojanapalli et al., 2016, Ge et al., 2015, 2016, Kawaguchi, 2016, Sun et al., 2017]. Taken together, these results show how interesting non-convex optimization problems can be efficiently solved using first-order methods. Our work contributes to the second line of research, offering insight into landscape of the optimization objective $\ell(\cdot)$.

2   Problem formulation

We defer measurability assumptions required for a rigorous presentation until the end of this section, keeping most of the formulation less formal but more accessible. A Markov decision process (MDP)

is a six-tuple $(\mathcal{S}, (\mathcal{A}_s)_{s\in\mathcal{S}}, g, P, \gamma, \rho)$, consisting of a state space $\mathcal{S} \subset \mathbb{R}^n$, action spaces $(\mathcal{A}_s)_{s\in\mathcal{S}}$, cost function g, transition kernel P, discount factor $\gamma \in (0,1)$ and initial distribution ρ. 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 a probability distribution $P(\cdot|s,a)$ over $\mathcal{S}$ for any given state $s \in \mathcal{S}$ and action $a \in \mathcal{A}_s$. The cost function $g(\cdot)$ encodes the instantaneous expected cost $g(s,a)$ incurred when selecting action a in state s. We assume that per-period costs are uniformly bounded, meaning $\sup_{s\in\mathcal{S}, a\in\mathcal{A}_s} |g(s,a)| < \infty$.

A stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ is a function that prescribes a feasible action[2] $\pi(s) \in \mathcal{A}_s$ for each state $s \in \mathcal{S}$. Let Π denote the set of all (measurable) stationary polices and $\mathcal{M} \subset \mathcal{S}$ be any measurable set. For any $\pi \in \Pi$, let $g_\pi(s) = g(s, \pi(s))$ denote the per step cost function and $J_\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^\infty \gamma^t g_\pi(s_t) \right]$ to be the corresponding cost-to-go function. Here, the notation $\mathbb{E}_s^\pi[\cdot]$ indicates that expectation is taken over the Markovian sequence of states $(s_0, s_1, \cdots)$ with $s_0 = s$ and the transition kernel $\mathbb{P}^\pi(s_{t+1} \in \mathcal{M}|s_t, \cdots, s_0) = P(\mathcal{M}|s_t, \pi(s_t))$. More generally, we write $\mathbb{E}_\nu^\pi[\cdot]$ when the initial state is randomly drawn from some distribution ν over $\mathcal{S}$ and let $\mathbb{P}_\nu^\pi(\cdot) = \mathbb{P}^\pi(s_1 \in \mathcal{M}|s_0 \sim \nu)$ denote the corresponding probability measure.

Loss function and policy gradient methods. Under the initial distribution ρ, a stationary policy $\pi \in \Pi$ has discounted average cost

$$\ell(\pi) = (1-\gamma)\mathbb{E}_\rho^\pi \left[\sum_{t=0}^\infty \gamma^t g_\pi(s_t) \right] = (1-\gamma) \int J_\pi(s)\rho(ds), \tag{1}$$

and *discounted state-occupancy measure* given by

$$\eta_\pi(\mathcal{M}) = (1-\gamma) \sum_{t=0}^\infty \gamma^t \mathbb{P}_\rho^\pi(s_t \in \mathcal{M}) \qquad \mathcal{M} \subset \mathcal{S}. \tag{2}$$

Since $\mathbb{P}_\rho^\pi(s_t \in \cdot)$ is the distribution of the state at time t under the policy π and initial distribution ρ, the discounted state occupancy measure gives the discounted fraction of time the system spends in a given part of the state space. Note that $\ell(\pi)$ can be equivalently written as $\ell(\pi) = \int g_\pi(s)\, \eta_\pi(ds)$. Hence, it's natural to normalize by $(1-\gamma)$ in (1).

Policy gradient methods directly apply first-order optimization methods to minimize $\ell(\cdot)$ over a chosen parameterized class of policies, $\Pi_\Theta = \{\pi_\theta(\cdot) : \theta \in \Theta\} \subset \Pi$. For this reason, we often refer to ℓ as the policy gradient objective function. We assume $\Theta \subset \mathbb{R}^d$ is convex and denote $\mathcal{J}_\Theta = \{J_{\pi_\theta} : \theta \in \Theta\}$ to be the set of cost-to-go functions corresponding to Π_Θ. 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 overload notation, writing $\ell(\theta) = \ell(\pi_\theta)$. We will later impose smoothness conditions to ensure differentiability of $\ell(\theta)$.

The optimal cost-to-go function is defined as $J^*(s) = \inf_{\pi\in\Pi} J_\pi(s)$. A stationary policy π is said to be optimal if $J_\pi(s) = J^*(s)$ for every $s \in \mathcal{S}$. Under the technical conditions stated below in Assumption 2, at least one optimal policy exists, which we denote by π^* throughout the paper. The next lemma establishes that optimal policies are minimizers of the average cost function ℓ and that the reverse direction essentially holds when ρ places positive weight on all parts of the state space. For example, if $\mathcal{S}$ is discrete and $\rho(s) > 0$ for all $s \in \mathcal{S}$, then any minimizer of ℓ is an optimal policy. The proof can be found in Appendix B.

---

[2]Rather than develop separate notation for randomized and deterministic policies, in specific settings like Example 3, we accommodate randomized policies by letting $a \in \mathcal{A}_s$ denote a choice of some probability vector.

Lemma 1. *A policy satisfies $\pi \in \arg\min_{\pi' \in \Pi} \ell(\pi')$ if and only if $J_\pi = J^*$ ρ–almost surely, i.e.*
$\rho(\{s \in \mathcal{S} : J_\pi(s) = J^*(s)\}) = 1$.

Exploratory initial distribution. Policy gradient methods have poor convergence properties if applied without an exploratory initial distribution. See Appendix A for a full discussion. Inspired by Kakade and Langford [2002], we assume that the discounted state occupancy measure under an optimal policy is absolutely continuous with respect to the initial distribution, mathematically denoted as $\eta_{\pi^*} \ll \rho$. Roughly, an assumption of this form ensures that the policy gradient loss function in (1) is sensitive to policy performance in all important parts of the state space. When $\mathcal{S}$ is discrete, it suffices to assume $\rho(s) > 0$ for each $s \in \mathcal{S}$. When ρ and η_π both possess probability density functions (PDFs) over $\mathcal{S}$, it suffices to assume that ρ is supported over $\mathcal{S}$. In Section 7, we discuss more about the dependence of our results on specific choices of the initial distribution.

Assumption 1. *We assume that η_{π^*} is absolutely continuous with respect to ρ. That is, for any $\mathcal{M} \subset \mathcal{S}$ such that $\rho(\mathcal{M}) = 0$, we have $\eta_{\pi^*}(\mathcal{M}) = 0$.*

By (2), $\eta_\pi \succeq (1 - \gamma)\rho$ for any stationary policy $\pi \in \Pi$ where $\succeq$ is used to denote an inequality that holds element-wise. Therefore, under Assumption 1, $\eta_{\pi^*} \ll \eta_\pi$ for any policy π. Many of our results actually rely on this consequence of Assumption 1, rather than Assumption 1 itself.

Bellman operators. Let $\mathcal{J}$ denote the set of bounded (measurable) functions on the state space. Define the Bellman operator $T_\pi : \mathcal{J} \to \mathcal{J}$ and Bellman optimality operator $T : \mathcal{J} \to \mathcal{J}$ by

$$(T_\pi J)(s) := g(s, \pi(s)) + \int J(s')P(ds'|s, \pi(s)), \tag{3}$$

$$(TJ)(s) := \min_{a \in \mathcal{A}_s} \left[g(s, a) + \int J(s')P(ds'|s, a) \right]. \tag{4}$$

The Bellman optimality operator in (4) can be equivalently defined as $TJ(s) = \min_{\pi \in \Pi} T_\pi J(s)$. Suitable measurability condition given in assumption 2 ensures that the minimum in (4) is attained and in particular there exists policy $\pi \in \Pi$ such that $T_\pi J = TJ$. It is well known that when the per period costs are uniformly bounded (as we assumed), T and T_π are monotone and contraction operators with respect to the maximum norm. Their unique fixed points are J_π and J^*, respectively. Throughout, we make repeated use of the standard element-wise inequalities which hold for any policy π and $J, J_\pi \in \mathcal{J}$ and can be deduced from (3) and (4) above.

$$TJ \preceq T_\pi J \quad \text{and} \quad TJ_\pi \preceq J_\pi \tag{5}$$

The state-action cost-to-go function ("Q-function") corresponding to a policy $\pi \in \Pi$ is given by,

$$Q_\pi(s, a) = g(s, a) + \gamma \int J_\pi(s')P(ds' \mid s, a). \tag{6}$$

Define $Q^*(\cdot) = Q_{\pi^*}(\cdot)$. Notice that for any polices $\pi, \pi' \in \Pi$, we have the following relations,

$$Q_\pi(s, \pi(s)) = J_\pi(s), \qquad Q_\pi(s, \pi'(s)) = (T_{\pi'} J_\pi)(s), \qquad \min_{a \in \mathcal{A}_s} Q_\pi(s, a) = (TJ_\pi)(s). \tag{7}$$

Notation. For any $J \in \mathcal{J}$, we define the weighted p-norm as, $\|J\|_{p,\nu} = \left(\int |J(s)|^p \nu(ds) \right)^{1/p}$ for a given probability distribution ν over $\mathcal{S}$ and $p \geq 1$. Similarly, the maximum norm is given by $\|J\|_{\infty} = \sup_{s \in \mathcal{S}} |J(s)|$. For a matrix $A \in \mathbb{R}^{n \times m}$, we write $\|A\|_p = \max_{x:\|x\|_p=1} \|Ax\|_p$ for the operator norm and, in the case where $m = n$, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalue. Let $\langle x, y \rangle = x^\top y = \sum_i x_i y_i$ denote the standard inner product. We write element-wise inequalities as $\preceq$ or $\succeq$, so $J \preceq J'$ if and only if $J(s) \leq J'(s)$ for each $s \in \mathcal{S}$.

Measurability assumptions. Here we state assumptions which avoid pathological measurability issues that can arise in dynamic programming with general state and action spaces [Bertsekas and Shreve, 1978, Blackwell, 1965], i.e. when both state and actions spaces are uncountably infinite. Readers unfamiliar with measure theory can safely skip this while still understanding the paper's main insights. The key condition is the existence of a measurable selection rule, i.e. a measurable rule that associates each state with an action attaining the minimum in a Bellman update. All of our examples satisfy appropriate smoothness and compactness conditions that ensure such a condition holds. We refer the readers to [Hernández-Lerma and Lasserre, 2012, Section 3.3] for a detailed account. A brief introduction is also given in [Puterman, 2014, Section 6.2.5]. We use the term measurable to refer to a Borel measurable set or function.

Assumption 2 (Measurable selection). *Assume the sets $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{K} := \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$ as well as the function $g : \mathcal{K} \to \mathbb{R}$ are measurable. The transition kernel P is a stochastic kernel on $\mathcal{S}$ given $\mathcal{K}$, meaning that for each $(s, a) \in \mathcal{K}$, $P(\cdot|s, a)$ is a probability measure on $\mathcal{S}$ and for each measurable set $\mathcal{M} \subset \mathcal{S}$, $P(\mathcal{M}|\cdot)$ is a measurable function. For each bounded measurable function J on $\mathcal{S}$, there is a measurable policy $\pi \in \Pi$ such that*

$$g(s, \pi(s)) + \gamma \int J(s') P(ds'|s, \pi(s)) = \inf_{a \in \mathcal{A}_s} \left[g(s, a) + \gamma \int J(s') P(ds'|s, a) \right].$$

3 Background on smooth nonconvex optimization

Convergence to stationary points. 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 that under appropriate smoothness conditions many algorithms will converge to stationary points of the objective, i.e. those satisfying the first-order necessary conditions for optimality as in Definition 1. 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 properties of specific algorithms.

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 if $\langle x' - x, \nabla f(x) \rangle \geq 0$ for all $x' \in \mathcal{X}$. For $\mathcal{X} = \mathbb{R}^d$, a stationary point satisfies $\nabla f(x) = 0$.*

We include an illustrative result showing that projected gradient descent converges asymptotically to a stationary point under appropriate smoothness and regularity conditions. This result can be generalized in numerous ways. For example, [Beck, 2017, Theorem 10.15] provides rates of convergence. A more complete treatment can be found in textbooks on nonlinear optimization [see e.g Bertsekas, 1997].

The result below covers two cases. The first assumes ∇f is Lipschitz, or, equivalently for twice differentiable functions, that the maximum eigenvalue of its Hessian is bounded above. The second relaxes this condition, only requiring regularity properties on the sublevel set of the initial iterate. This accommodates functions like $f(x) = x^4$, whose second derivative is unbounded. This result is possible because projected gradient descent with sufficiently small step-sizes is guaranteed to reduce cost in each iteration, so all iterates lie in certain sublevel sets. The restriction that f has bounded sublevel sets is satisfied if the feasible region $\mathcal{X}$ is itself a bounded set or if the function is coercive, meaning $f(x) \to \infty$ as $\|x\| \to \infty$. In problems where this is not naturally satisfied, it can sometimes be enforced by adding a small penalty function (e.g. an entropy regularizer) to the objective. Recall that a point x_∞ is said to be a limit point of a sequence $\{x_k\}$ if some subsequence converges to x_∞.

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 and its β–sublevel set $\{x \in \mathcal{X} : f(x) \leq \beta\}$ is bounded for each $\beta \in \mathbb{R}$. Consider the sequence $x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha \nabla f(x_k))$ for $k \in \mathbb{N}$.*

1. *[Beck, 2002, 2017] Assume f is differentiable on an open set containing $\mathcal{X}$ and ∇f is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant L. If $\alpha \in (0, 1/L]$, the sequence $\{x_k\}$ has at least one limit point and any limit point x_∞ is a stationary point of $f(\cdot)$ on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.*

2. *Suppose f is continuously twice differentiable on an open set containing the sublevel set $\{x \in \mathcal{X} : f(x) \leq f(x_0)\}$. For a sufficiently small $\alpha > 0$, the sequence $\{x_k\}$ has at least one limit point and any limit point x_∞ is a stationary point of $f(\cdot)$ on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.*

Proof. The proof for the second part closely follows the proof for the first part as shown in [Beck, 2002, 2017] with slight modifications. For completeness, we give a proof sketch in Appendix C. □

Convergence rates under gradient dominance. Results like Lemma 2 ensure that first order methods converge asymptotically to stationary points under mild regularity conditions. Then, even if the cost function is non-convex, such algorithms will converge toward a global optimum if all stationary points are optimal. A stronger property, called the Polyak-Lojasiewicz (PL) inequality and also commonly known as *gradient dominance*, effectively requires that approximate stationary points are also approximately optimal. Combined with regularity conditions, this yields rates of convergence for first-order methods. Below, we introduce a notion of gradient dominance which might seem somewhat nonstandard, since many authors treat only unconstrained problems. For the unconstrained case, this result reduces to the well known PL inequality of Polyak [1963], $\min_{x' \in \mathbb{R}^d} f(x) \geq f(x) - \frac{c^2}{2\mu} \|\nabla f(x)\|_2^2$.

Definition 2. *For $\mathcal{X} \subseteq \mathbb{R}^d$, we say f is (c, μ)–gradient dominated over $\mathcal{X}$ if there exists constants $c > 0$ and $\mu \geq 0$ such that*

$$\min_{x' \in \mathcal{X}} f(x') \geq f(x) + \min_{x' \in \mathcal{X}} \left[c \langle \nabla f(x), x' - x \rangle + \frac{\mu}{2} \|x - x'\|_2^2 \right] \quad \forall x \in \mathcal{X}. \tag{8}$$

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

Any stationary point of a gradient dominated function is globally optimal. To see this, note that if a point x is stationary, meaning $\langle \nabla f(x), x' - x \rangle \geq 0$ for every $x' \in \mathcal{X}$, then the minimizer of

9

the right hand side in (8) is x, implying $\min_{x' \in \mathcal{X}} f(x') \geq f(x)$. More broadly, note how in (8), the optimality gap $\min_{x' \in \mathcal{X}} f(x') - f(x)$ can be bounded by a measure of how far x is from stationarity, which is captured by the minimization problem on the right hand side.

Convex and strongly convex functions are gradient dominated. Recall that a differentiable function f is said to be μ–strongly–convex if it satisfies the inequality

$$f(x') \geq f(x) + \langle \nabla f(x), \, x' - x \rangle + \frac{\mu}{2} \| x - x' \|_2^2, \tag{9}$$

for every $x, x' \in \mathcal{X}$. Minimizing over x' on each side of (9) shows that a μ–strongly-convex function is $(1, \mu)$–gradient dominated. Similarly, convex functions satisfy (9) with $\mu = 0$, and therefore are $(1, 0)$–gradient dominated. Thus, gradient dominated functions satisfy a critical property of convex functions (relating the optimality gap to distance from stationarity). However, there are important classes of functions that are gradient dominated despite being nonconvex.

Under gradient dominance conditions, popular first order optimization algorithms are assured to converge to the global minimum and a simple analysis provides finite time rates of convergence. As an illustrative result, part (a) of Lemma 3 strengthens Lemma 2 by providing an $\mathcal{O}(1/\sqrt{T})$ convergence rate. While part (b) of this lemma assumes $\mathcal{X} = \mathbb{R}^d$, it is also possible to show geometric rates for projected gradient descent on constrained subsets $\mathcal{X} \neq \mathbb{R}^d$. We do not consider this case for brevity.

Lemma 3 (Convergence rates for gradient dominated smooth functions). *Consider the problem,* $\min_{x \in \mathcal{X}} f(x)$ *where* $\mathcal{X} \subseteq \mathbb{R}^d$ *is nonempty. Assume* ∇f *is* L–*Lipschitz continuous on* $\mathcal{X}$. *Denote* $f^* = \inf_{x' \in \mathcal{X}} f(x')$. *Consider the sequence* $x_{t+1} = \mathrm{Proj}_{\mathcal{X}} (x_t - \alpha \nabla f(x_t))$.

1. *Let* $\mathcal{X} \subset \mathbb{R}^d$ *be bounded. Set* $R = \sup_{x, x' \in \mathcal{X}} \| x - x' \|_2$ *and* $k = \sup_{x \in \mathcal{X}} \| \nabla f(x) \|_2$. *If* $\alpha \leq \min\{\frac{1}{k}, \frac{1}{L}\}$ *and* f *is* $(c, 0)$–*gradient-dominated, then,*

$$f(x_T) - f^* \leq \sqrt{\frac{2R^2 c \left(f(x_0) - f^* \right)}{\alpha T}}.$$

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

$$f(x_T) - f^* \leq \left(1 - \frac{\mu}{c^2 L} \right)^T \left(f(x_0) - f^* \right).$$

Proof. See Appendix C for a detailed proof of Part (1). The proof of part (2) can be found in [Karimi et al., 2016, Polyak, 1963]. $\square$

4 Motivation from linear quadratic control

We first motivate and instantiate our general results for the special case of linear quadratic (LQ) control. Leveraging many of the closed form expressions available in this case, recent work of Fazel et al. [2018] showed that policy gradient methods converge to the globally optimal policy under some technical conditions. The key to their result is showing that the infinite horizon cost function, despite being non-convex, has no suboptimal stationary points (and is in fact gradient dominated). Given the presence of bad stationary points in Example 1, there must be some special problem structure driving this, but what? Quite different from Fazel et al. [2018], our arguments involve classical understanding

of the single period cost function underlying policy iteration, avoiding the complications of directly analyzing the infinite horizon cost function $\ell(\cdot)$.

We highlight the two key properties of LQ control identified in Section 1.1. That is, we show that (a) the class of linear policies is closed under policy improvement and (b) the policy iteration problem can be solved to optimality by a gradient method, since it is convex quadratic and therefore has no suboptimal stationary points. A short proof then shows how these two conditions imply that $\ell(\cdot)$ has no suboptimal station points. Like Fazel et al. [2018], we simplify the presentation by studying deterministic LQ control, but it is easy to allow for noisy dynamics.

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

$$\text{Minimize} \quad \sum_{t=0}^{\infty} \gamma^t \left(a_t^\top R a_t + s_t^\top C s_t \right)$$

$$\text{Subject to} \quad s_{t+1} = A s_t + B a_t, \quad s_0 \sim \rho$$

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 that the second moment of the initial distribution $\mathbb{E}_\rho \left[s_0 s_0^\top \right]$ is finite and positive definite. In this setting, 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 searching for the optimal θ via a gradient method. Unfortunately, the loss function $\ell(\theta) = (1-\gamma)\mathbb{E}_\rho[J_{\pi_\theta}(s_0)]$ is non-convex (see Appendix B in Fazel et al. [2018]), making it unclear if gradient descent on $\ell(\theta)$ would reach the global minimum.

For LQ control, if a linear policy π_θ is applied from a state s_0, then unrolling the linear dynamics we have $s_t = (A + B\theta)s_{t-1} = \cdots = (A + B\theta)^t s_0$. Then, we can write the cost-to-go function as:

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

A linear policy π_θ is said to be stable *if its cost-to-go is finite from all initial states, or equivalently, if all eigenvalues of the matrix $\sqrt{\gamma}(A + B\theta)$ lie strictly within the unit circle. Let $\Theta_S \subset \mathbb{R}^{k \times n}$ denote the set of all parameters defining stable linear policies. One can show that $\ell(\theta) = \infty$ if $\theta \notin \Theta_S$. Note that when $\gamma = 1$ our definition reduces to the more standard definition of a stable linear policy in undiscounted problems. See Bertsekas [1995] for further discussion on stability in discounted LQ control. 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, it can be shown that starting from a stable linear policy, policy iteration (PI) converges to an optimal policy by solving a sequence of simpler single period optimization problems [Hewer, 1971, Kleinman, 1968]. The optimization problems arise when applying the Bellman operator, which in this case can be written as

$$(T J_{\pi_\theta})(s) = \min_{a \in \mathbb{R}^k} \left[\underbrace{a^\top R a + s^\top C s + \gamma(As + Ba)^\top K_\theta (As + Ba)}_{= Q_{\pi_\theta}(s,a)} \right]. \tag{10}$$

A single step iteration of PI updates the stable linear policy π_θ to a new policy π^+ that selects the action $\pi^+(s) = \arg\min_a Q_{\pi_\theta}(s, a)$. This can equivalently be expressed as $T_{\pi^+} J_{\pi_\theta} = T J_{\pi_\theta}$.

Typically, a PI update requires solving a unique optimization problem for each state, but given the convex quadratic nature of the problem in (10), *it can be easily checked that* $\pi^+(s) = \overline{\theta}s$ *where* $\overline{\theta} = -\gamma(R + \gamma B^\top K_\theta B)^{-1}B^\top K_\theta A$. *Thus, starting from a linear policy, a PI update yields yet another linear policy implying that for LQ control, the class of linear policies is closed under policy improvement. Policy iteration steps are sometimes called policy improvement steps because the new policy* π^+ *is assured to have lower lower cost-to-go. In particular,* $J_{\pi^+}(s) \leq J_{\pi_\theta}(s)$ *for all* $s \in \mathcal{S}$ *and the improvement is strict at some set of states if* π_θ *is not an optimal policy. Crucially, this implies that the PI update* π^+ *is also a stable policy (see Lemma 4 for a formal statement).*

A useful interpretation of the PI update is to view it as the minimizer of a weighted policy iteration cost,

$$\mathcal{B}(\theta'|\eta, J_{\pi_\theta}) := \int (T_{\pi_{\theta'}} J_\theta)(s)\,\eta(ds) = \int Q_{\pi_\theta}(s, \pi_{\theta'}(s))\,\eta(ds), \tag{11}$$

over policy parameters θ'. *Under appropriate conditions on the distribution* η, *the solution is unique*[3] *From* (10), *note that* $Q_{\pi_\theta}(s, a)$ *is a convex quadratic function of action* a, *which implies* $Q_{\pi_\theta}(s, \theta's)$ *is a convex quadratic function of* θ' *and so is* (11). *This shows that the weighted PI objective has no suboptimal stationary points, the second key property we identified in Section 1.1.*

Because the per-stage cost functions in LQ control are unbounded, this example is technically beyond the scope of the problem formulation in Section 2. Thankfully, the properties of Bellman operators that underlie our analysis hold for LQ control. See Section D.2.1 for details. As a consequence, *the proofs* of our general results will essentially apply without modification to stable policies in LQ control. Nevertheless, to be formal, any results about LQ control will clearly specify that they apply to Example 2 and standalone proofs are given for completeness.

To discuss policy gradient methods, we first need some smoothness properties of $\ell(\cdot)$. Beginning with an initial stable policy θ_0 which incurs finite cost, first-order algorithms with an appropriate step-size are assured to decrease cost on every iteration, meaning that iterates remain in the sublevel set $\{\theta \in \Theta : \ell(\theta) \leq \ell(\theta_0)\} \subset \Theta_S$. The next lemma establishes regularity conditions on these sublevel sets which are sufficient to apply optimization results, like Lemma 2, to show that gradient descent converges to a stationary point of $\ell(\cdot)$.

Lemma 4. *Consider the LQ control problem formulated in Example 2. The set* Θ_S *is open and* ℓ *is twice continuously differentiable on* Θ_S. *For any* $\alpha \in \mathbb{R}$, *the sublevel set* $C_\alpha := \{\theta \in \mathbb{R}^{n \times k} : \ell(\theta) \leq \alpha\}$ *is a compact subset of* Θ_S *and* $\sup_{\theta \in C_\alpha} \|\nabla^2 \ell(\theta)\| < \infty$.

Proof. These properties follow from [Rautert and Sachs, 1997, Toivonen, 1985]. Some additional details are provided in Appendix D.2.2 □

With this background, a simple proof shows that for LQ control, the policy gradient loss function has no suboptimal stationary points despite being non-convex. Essentially, for any stable linear policy that is suboptimal, we show that moving along the line segment toward a policy iteration update forms a descent direction, implying that it cannot be a stationary point. The two properties we identified, convexity of the weighted PI objective and closure with respect to the class of linear policies, are critical for this argument.

---

[3]Due to the quadratic structure of the objective, it is enough η has a finite and strictly positive definite second moment matrix. This is similar to our assumption about the initial distribution ρ and it can be established along the same lines as the end of the proof of Lemma 5.

Lemma 5. *For the LQ control problem formulated in Example 2, any stable linear policy θ satisfies $\nabla \ell(\theta) = 0$ if and only if $J_{\pi_\theta} = J^*$.*

Proof sketch. Consider a stable linear policy π_θ and take $\pi_{\overline{\theta}}$ to be a policy iteration update. Standard analysis of policy iteration, using monotonicty of the Bellman operator, shows that $J_{\pi_{\overline{\theta}}} \preceq T J_{\pi_\theta} \preceq J_{\pi_\theta}$. Here, the second inequality is strict at some set of states unless π_θ is an optimal policy. This implies $\ell(\overline{\theta}) < \ell(\theta)$ when π_θ is suboptimal. Now, consider a soft policy iteration update $\theta^\alpha = (1 - \alpha)\theta + \alpha\overline{\theta} \ \forall \alpha \in [0, 1]$. Leveraging convexity of the policy iteration objective in (11), a short argument shows $\frac{d}{d\alpha} \ell(\theta^\alpha)|_{\alpha=0} \leq 0$ and this inequality is strict unless π_θ is an optimal policy. See Appendix D.2.3 for a detailed proof. $\square$

This idea of constructing a descent direction is strongly reminiscent to the arguments in Kakade and Langford [2002] for finite MDPs. While we find this proof to be intuitive, it relies not just on the closure property, but on convexity of the policy class[4] as well as convexity of the policy iteration cost function, which will not hold in all of our examples. One contribution of this paper is to find a clean generalization of this argument, relaxing convexity conditions into Condition 2 in the next section.

5 General results

We now generalize some of the insights discussed above for the LQ control example to identify properties which ensure the policy gradient objective has no suboptimal stationary points. In the next section, we show these properties hold for various problems settings beyond LQ control.

5.1 Conditions on the policy iteration cost function

Consider the weighted policy iteration or the "Bellman" cost function introduced in (11).

$$\mathcal{B}(\bar{\pi} \mid \eta, J_\pi) = \int (T_{\bar{\pi}} J_\pi)(s) \, \eta(ds) = \int Q_\pi(s, \bar{\pi}(s)) \, \eta(ds),$$

for a probability distribution η over $\mathcal{S}$ and $J_\pi \in \mathcal{J}$. Here, the final equality follows by noting that $(T_{\bar{\pi}} J_\pi)(s) \equiv Q_\pi(s, \bar{\pi}(s))$ from (7). This Bellman cost function is a *single period* objective, considering the cost-to-go of following $\bar{\pi}$ for a single period and following π thereafter. We overload notation to write $\mathcal{B}(\theta \mid \eta, J) = \mathcal{B}(\pi_\theta \mid \eta, J)$. When the state space is discrete and $\eta(s) > 0$ for all $s \in \mathcal{S}$, classic policy iteration update can be equivalently written as,

$$\pi_{k+1} = \arg\min_{\pi \in \Pi} \ \mathcal{B}(\pi | \eta, J_{\pi_k}). \tag{12}$$

Policy iteration, like value iteration, indirectly optimizes the infinite horizon cost-to-go, $\ell(\cdot)$ by solving the sequence of simpler single period problems in (12). On the other hand, policy gradient methods aim to directly minimize $\ell(\pi_\theta)$. Despite this crucial difference, our approach is to infer properties of the complex multi-period objective $\ell(\cdot)$ using some structure present in the single period problems. We outline this below.

---

[4]The class of linear policies is convex. That is, the policy $\alpha\pi_\theta + (1 - \alpha)\pi_{\overline{\theta}}$ is a linear policy for any given linear policies $\pi_\theta, \pi_{\overline{\theta}}$ and some $\alpha \in [0, 1]$. However, the class of threshold policies, used in the optimal stopping and inventory control problems is not convex. If $\pi_\theta(s) = \mathbf{1}(s \leq \theta)$ for $\theta \in \mathbb{R}$ is a threshold policy, then $\frac{1}{2}\pi_\theta + \frac{1}{2}\pi_{\theta'}$ is not a threshold policy when $\theta \neq \theta'$.

Differentiability. Before arguing about any convergence properties, we first need conditions for the policy gradient itself to be well defined. Condition 0 below states smoothness conditions, related to partial differentiability of the Bellman objective $\mathcal{B}(\cdot)$, that ensure $\ell(\cdot)$ is differentiable and its gradients satisfy a convenient formula used in practical implementations [Marbach and Tsitsiklis, 2001, Silver et al., 2014, Sutton and Barto, 2018].

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

As $\mathcal{B}(\overline{\theta}|\eta_{\pi_\theta}, J_{\pi_\theta}) = \int Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s)) \eta(ds)$, differentiability in $\overline{\theta}$ follows if $Q_{\pi_\theta}(s, \pi_{\overline{\theta}}(s))$ is differentiable almost everywhere and the exchange of derivative and integral is permitted. Also note that $\mathcal{B}(\theta|\eta_{\pi_{\overline{\theta}}}, J_{\pi_\theta}) = \int J_{\pi_\theta}(s) \eta_{\pi_{\overline{\theta}}}(ds)$. Thus, differentiability in $\overline{\theta}$ is related to the existence of a weak derivative of the state occupancy measure [Pflug, 1990, 1988]. A large literature studies sufficient conditions that imply differentiability [Asmussen and Glynn, 2007, Glasserman and Ho, 1991, Rhee and Glynn, 2017]. We do not try to advance that literature, instead focusing on the convergence of policy gradient methods when they are well defined.

Condition 0 arises quite naturally in calculating the derivatives of $\ell(\theta)$ as shown in the policy gradient theorem below. To see this, we refer the readers to a short proof in Appendix B.

Lemma 6 (Policy gradient theorem). *Under Condition 0, $\ell(\theta)$ is continuously differentiable and*

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

Remark 1. *Our presentation differs from the familiar form of the policy gradient theorem [Sutton and Barto, 2018], which is written as $\nabla \ell(\theta) = \mathbb{E}[Q_{\pi_\theta}(s, a)\nabla \log \pi_\theta(a|s)]$ where $\pi_\theta(a|s)$ is the probability of selecting a deterministic action a in state a and the expectation is taken over the distribution of states and actions under π_θ. This form is useful for gradient estimation, but it applies only to stochastic policies and seems to obscure connections with the gradient of the weighted PI objective. The expression in Lemma 6 however is more general and can be applied to both deterministic and stochastic policies (by taking $\pi_\theta(s)$ to be a probability vector). Lemma 6 was first derived under somewhat stringent regularity conditions by Silver et al. [2014]. Condition 0 and our short proof appear to be new.*

Closure under policy improvement. We now introduce one of our main conditions, which we call closure under policy improvement. This is a consistency condition which essentially says that the policy improvement update can be solved within the policy class.

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

This closure assumption accommodates interesting examples in which a restricted class of policies is naturally aligned with the decision task, like the class of linear policies in LQ control or threshold policies for optimal stopping. We emphasize that this condition is weaker than requiring the policy class to contain nearly all stochastic policies. However, it is stronger than just requiring the policy class to contain an optimal policy. An extension in section 8 bounds the optimality gap of stationary points when the policy is not close but satisfies a relaxed closure condition.

Some condition along these lines appears to be necessary. Indeed, Example 1 in the introduction showed an extremely simple problem for which policy gradient methods can get stuck in a bad local minima even though the policy class contains an optimal policy. There, we consider a two state ($\mathcal{S} = \{s_L, s_R\}$) deterministic MDP with actions corresponding to moving left (L) and right (R) respectively. An action $a \in \mathcal{A} = [0, 1]$ indicates a probability of choosing Right. We consider a restricted class of policies of the form $\pi_\theta(s_L) = \pi_\theta(s_R) = \theta$, which plays action R in either state with probability $\theta \in [0, 1]$. Simple calculations show why this policy class is not closed under policy improvement. For any policy π_θ we can write the Q-function as

$$Q_{\pi_\theta}(s_L, a) = (1 - a) \cdot 1 + a \cdot 2 + \gamma \left((1 - a)J_{\pi_\theta}(s_L) + aJ_{\pi_\theta}(s_R)\right)$$
$$Q_{\pi_\theta}(s_R, a) = (1 - a) \cdot 2 + a \cdot 0 + \gamma \left((1 - a)J_{\pi_\theta}(s_L) + aJ_{\pi_\theta}(s_R)\right)$$

Consider a policy where θ is nearly zero, so π_θ moves left with high probability. It is easy to check that $0 = \arg\min_{a \in [0,1]} Q_{\pi_\theta}(s_L, a)$. That is, in state s_L, it is optimal for the decision maker to move left assuming that the policy π_θ (which almost always move left) will be applied in future periods. A similar argument shows that $1 = \arg\min_{a \in [0,1]} Q_{\pi_\theta}(s_R, a)$. Clearly, this policy iteration update is not contained in the one dimensional policy class $\{\pi_\theta : \theta \in [0, 1]\}$.

Stationary points of the weighted PI objective. As a first order method, policy gradients require additional local optimization structure to succeed. The following conditions ensure that first order methods are suitable for solving the weighted policy iteration problem.

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

Condition 2.B (Gradient dominance of the weighted PI objective). *For any $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ is (c, μ)–gradient-dominated over Θ.*

It is worth emphasizing that the single period Bellman objective $\bar{\theta} \mapsto \mathcal{B}(\bar{\theta}|\eta_{\pi_\theta}, J_{\pi_\theta})$ is often much simpler than the infinite horizon objective $\ell(\bar{\theta})$. In LQ control it is a convex quadratic function and is therefore gradient dominated. For finite state action MDPs, say for instance in Example 1, it is linear and hence is also gradient dominated, even though we showed that the total cost function $\ell(\theta)$ is non-convex and can have suboptimal local minima. Even for complex neural networks, a very active literature studies the quality of stationary points and local minima for certain single period loss functions [Du and Lee, 2018, Livni et al., 2014].

5.2 Closed policy classes and optimality of stationary points.

Our first result establishes that the policy gradient objective has no suboptimal stationary points when the policy class is closed under policy improvement and the single period Bellman objective has no suboptimal stationary points.

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

Proof of Theorem 1. We first give a key lemma which establishes a Bellman-type equation that holds when the single period objective $\bar{\theta} \mapsto \mathcal{B}(\bar{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})$ has no bad stationary points.

Lemma 7. *Suppose Condition 2.A is satisfied. If θ is a stationary point of $\ell : \Theta \to \mathbb{R}$, then*

$$\int J_{\pi_\theta} d\eta_{\pi_\theta} = \min_{\pi \in \Pi_\Theta} \int (T_\pi J_{\pi_\theta}) \, d\eta_{\pi_\theta}.$$

Proof. If θ is a stationary point of $\ell : \Theta \to \mathbb{R}$, then by the policy gradient theorem in Lemma 6, it is also a stationary point of the function $\bar{\theta} \mapsto \mathcal{B}(\bar{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})$. Since Condition 2.A holds, this implies

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

Recalling the definition of $\mathcal{B}(\theta \mid \eta, J_\pi)$ in (11) lets us rewrite both sides of this equation as,

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

$\square$

Next, in Lemma 8, we state an "average" form of Bellman's equation which shows that under an exploratory initial distribution (see Assumption 1), an optimal policy has zero average Bellman error.

Lemma 8 (On average Bellman equation). *For any $\pi \in \Pi$,*

$$\ell(\pi) = \ell(\pi^*) \iff \int (J_\pi - T J_\pi) \, d\rho = 0$$

Proof. See Appendix B.1 for a detailed proof. $\square$

We now complete the proof of Theorem 1.

Proof of Theorem 1. To show the first direction, note that $\ell(\theta) = \ell(\pi^*)$ implies that θ is a minimizer of $\ell(\cdot)$. By the first order necessary conditions of optimality, θ must be a stationary point of $\ell(\cdot)$. To prove the other direction, suppose that θ is a stationary point of $\ell(\cdot)$. Then,

$$\int J_{\pi_\theta} d\eta_{\pi_\theta} = \min_{\pi \in \Pi_\Theta} \int [T_\pi J_{\pi_\theta}] \, d\eta_{\pi_\theta} = \int [T J_{\pi_\theta}] \, d\eta_{\pi_\theta}.$$

where the first equality follows from Lemma 7 (implied by Condition 2.A) while the second equality uses the closure property in Condition 1. By the definition in (2), $\eta_\pi \succeq (1 - \gamma)\rho$. Using that $J_\pi \succeq T J_\pi$, we have

$$0 = \int [J_{\pi_\theta} - T J_{\pi_\theta}] \, d\eta_{\pi_\theta} \geq (1 - \gamma) \int [J_{\pi_\theta} - T J_{\pi_\theta}] \, d\rho \geq 0.$$

The "on average Bellman equation" in Lemma 8 let's us conclude that $\ell(\theta) = \ell(\pi^*)$. $\square$

5.3 Convergence rates for policy gradient methods.

Theorem 1, which guarantees that the policy gradient objective has no suboptimal stationary points, is only an asymptotic result when viewed in context of Lemma 2. Our next result strengthens this by showing that $\ell(\theta)$ is gradient dominated (though possibly non-convex) for closed policy classes if the simpler weighted PI objective is gradient dominated. Recall, gradient dominance and smoothness of $\ell(\cdot)$ often imply fast convergence rates for first-order methods, using results like Lemma 3. Our investigation here is inspired by Fazel et al. [2018], who showed gradient dominance for LQ control by a careful manipulation of closed form expressions available in linear systems.

Our result in Theorem 2 below relies on a constant that measures the efficacy of an exploratory initial distribution in a more refined way than Assumption 1. We call this the *effective concentrability coefficient*, since it plays a role similar to the concentrability coefficients that are widely used 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]. 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. The somewhat opaque definition in (13) is precisely the quantity we need in our analysis. Section 7 provides more insights into the definition along with interpretable bounds for κ_ρ.

Definition 3. *Define the effective concentrability coefficient κ_ρ for the class of cost-to-go functions $\mathcal{J}_\Theta = \{J_{\pi_\theta} : \theta \in \Theta\}$ to be the smallest scalar such that*

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

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

Theorem 2 gives a gradient dominance condition on $\ell(\cdot)$ for closed policy classes under Condition 2.B. Subsequently, corollary 1 holds as (strongly) convex functions are gradient dominated.

Theorem 2. *If Conditions 0, 1, and 2.B hold, then $\ell(\cdot)$ is $\left(\frac{1-\gamma}{\kappa_\rho} \cdot c, \frac{1-\gamma}{\kappa_\rho} \cdot \mu\right)$–gradient dominated.*

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

Proof of Theorem 2. Our proof can be divided into two key steps. First, we use closure property (Condition 1) to bound the optimality gap of a policy, $\ell(\pi) - \ell(\pi^*)$, by the improvement under a weighted PI update. The second step uses the policy gradient theorem in Lemma 6 to translate this inequality into a gradient dominance condition on $\ell(\cdot)$. It is noteworthy that our results crucially depend on using an exploratory initial distribution under which $\kappa_\rho < \infty$.

Proof of Theorem 2. We first derive a consequence of the closure condition:

$$\ell(\pi_\theta) - \min_{\pi \in \Pi} \ell(\pi) = (1 - \gamma) \int [J_{\pi_\theta} - J^*] \, d\rho \stackrel{(a)}{=} (1 - \gamma) \|J_{\pi_\theta} - J^*\|_{1,\rho}$$

$$\stackrel{(b)}{\leq} \kappa_\rho \|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\rho}$$

$$\stackrel{(c)}{\leq} \frac{\kappa_\rho}{(1 - \gamma)} \|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1, \eta_{\pi_\theta}}$$

$$= \frac{\kappa_\rho}{(1 - \gamma)} \int [J_{\pi_\theta} - TJ_{\pi_\theta}] \, d\eta_{\pi_\theta}$$

$$\stackrel{(d)}{=} \frac{\kappa_\rho}{(1 - \gamma)} \left(\int J_{\pi_\theta} d\eta_{\pi_\theta} - \min_{\pi \in \Pi_\Theta} \int [T_\pi J_{\pi_\theta}] \, d\eta_{\pi_\theta} \right)$$

$$= \frac{\kappa_\rho}{(1 - \gamma)} \left(\mathcal{B}(\theta \mid \eta_{\pi_\theta}, J_{\pi_\theta}) - \min_{\theta' \in \Theta} \mathcal{B}(\theta' \mid \eta_{\pi_\theta}, J_{\pi_\theta}) \right).$$

Here (a) uses that $J_{\pi_\theta} \succeq J^*$, (b) applies the definition of κ_ρ in (13), (c) uses that $\eta_{\pi_\theta} \succeq (1 - \gamma)\rho$ (see definition in (2)) and (d) uses closure property of the policy class.

By Condition 2.B, $\bar{\theta} \mapsto \mathcal{B}(\bar{\theta} \mid \eta_\theta, J_{\pi_\theta})$ is (c, μ)–gradient dominated. Using gradient dominance and the policy gradient theorem in Lemma 6, we find

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

$$\leq - \min_{v \in \Theta} \left[c \left\langle \nabla_\theta \ell(\theta), v - \theta \right\rangle + \frac{\mu}{2} \|v - \theta\|_2^2 \right].$$

Combining this with the preceding calculation yields the desired result. □

5.4 Beyond closed policy classes: the case of non-stationary policies.

For finite horizon problems with a class of non-stationary policies, we can guarantee that the policy gradient objective has no spurious local minima under a much weaker condition. Rather than require the policy class to be closed under improvement, it is sufficient that the policy class contains an optimal policy[5]. For this reason, our theory will cover a broad variety of finite horizon dynamic programming problems for which structured policy classes are known to be optimal. Interestingly, this result relies critically on the use of a non-stationary policy class. Recall how Example 1 shows that for stationary policy classes, policy gradient methods can get stuck in bad local minima even if the policy class contains an optimal policy.

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-inhomogeneous 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, essentially encoding time-inhomogeneous policies.

---

[5]This is clearly a weaker property as closure of the policy class implies that it contains an optimal policy.

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 $h \leq H$, $P(\mathcal{S}_{h+1}|s, a) = 1$ 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 for this subsection which applies under conditions much weaker than those for Theorem 1. First, we only require the policy class to contain the optimal policy. Second, we relax Condition 2.A, which considered stationary points of single-period Bellman objective $\theta \mapsto \mathcal{B}(\theta|\eta_\pi, J_\pi)$, induced by any policy $\pi \in \Pi_\Theta$. Instead, we only need to impose a regularity condition on the Bellman objective corresponding to the optimal cost-to-go function, $\mathcal{B}(\theta|\eta, J^*)$.

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

However, for our argument to work only with Condition 4, we do require a stronger regularity property of the initial distribution ρ as compared to Assumption 1.

Assumption 3. *For $\pi \in \Pi_\Theta$, η_π is absolutely continuous with respect to ρ.*

Assumption 3 is crucial for our proof as it enables us to relate stationary points of $\mathcal{B}(\theta|\eta, J_\pi)$ to those of $\mathcal{B}(\theta|\eta, J^*)$. See Appendix B.2 for details. Also note that in context of finite horizon problems, Assumption 3 implies that the agent may begin in a sub-problem with fewer than H periods remaining. This would typically be possible in simulation based optimization and it is necessary for the results we derive.

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

Proof Sketch. The proof is given in Appendix B.2 and proceeds by backward induction. We first show that all stationary points must act according to an optimal policy from any state in $\mathcal{S}_H$. We then argue that at a stationary point, the policy must act optimally from any state in $\mathcal{S}_h$, for all $h < H$. $\square$

Note that Theorem 3 only gives a characterization of the stationary points of $\ell(\cdot)$. We leave the study of a gradient dominance condition for finite horizon problems as future work.

6 Examples

We have already discussed the linear quadratic control example in Section 4. It is noteworthy that even though $\ell(\cdot)$ is gradient dominated with degree two for LQ control, the convergence rate result in Lemma 3 doesn't directly apply as the smoothness properties only hold over sublevel sets. However, we believe a geometric convergence rate can be shown with a constant step-size chosen appropriately to ensure that iterates always remain in sublevel sets (similar to the proof of Lemma 2). In what follows, we apply our general results in Section 5 to several additional examples.

Finite state and action MDPs with natural parameterization.

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

$$g(s, a) = \sum_{i=1}^{k} g(s, e_i)\, a_i \qquad\qquad P(s'|s, a) = \sum_{i=1}^{k} P(s'|s, e_i)\, a_i. \qquad (14)$$

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

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

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

$$\mathcal{B}(\pi'|\eta_\pi, J_\pi) = \sum_{s \in \mathcal{S}} \eta_\pi(s) \sum_{i=1}^{k} Q_\pi(s, e_i)\, \pi'_{s,i} \qquad (15)$$

is convex (linear) in π'. This implies Condition 2.A along with the gradient dominance property in Condition 2.B. Therefore, Theorems 1 and 2 confirm that for tabular MDPs, $\ell(\cdot)$ has no suboptimal stationary points and is gradient dominated. Convergence rates like those in Lemma 3 also require smoothness properties. This can be found in Lemma E.3 of Agarwal et al. [2020]. Condition 0 is verified in Appendix D.1.

Regularized finite state and action MDPs with natural parameterization.

It is common to add 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 $R(a) = \frac{1}{\alpha} D_{\mathrm{KL}}(U||a)$ where U is the uniform distribution (i.e. $U_i = 1/k$ for each i), and D_{KL} denotes the Kullback-Leibler divergence. Note that we have chosen to regularize the single stage cost functions rather than $\ell(\theta)$ directly because this form of regularization is can be viewed as a special case of our problem formulation. In this case, the weighted Bellman objective takes the form

$$\mathcal{B}(\pi' \mid \eta_\pi, J_\pi) = \sum_{s \in \mathcal{S}} \eta_\pi(s) \sum_{i=1}^{k} Q_\pi(s, e_i)\, \pi'_{s,i} + \sum_{s \in \mathcal{S}} \eta_\pi(s) R(\pi'(s)), \qquad (16)$$

which is the sum of a linear function and strongly convex regularizer. Since the weighted Bellman objective is strongly convex, Corollary 1 implies that $\ell(\cdot)$ is gradient dominated with degree two, implying appropriate first order methods attain a linear rate of converge to the global optimum of the regularized MDP.

Regularized finite state and action MDPs with softmax parameterization.

Example 4 (softmax policies). *For tabular MDPs, it is quite common to use a softmax policy parameterized by $\theta \in \mathbb{R}^{n \times k}$ where for any state s, the $\pi_\theta(s) \in \Delta^{k-1}$ 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, so $\theta \in \mathbb{R}^{n \times k}$. This means that the mapping $\theta \mapsto \pi_\theta$ is invertible, so each policy in the policy class corresponds to a unique parameter vector.

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, the optimization problem $\min_{\theta \in \Theta} \ell(\theta)$ has no optimal solution and in fact $\ell(\cdot)$ has no stationary points. Convergence to an optimal policy only occurs 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 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. For concreteness, one might consider again the the KL divergence regularizer described above, defined by $R(a) = \frac{1}{\alpha} D_{\mathrm{KL}}(U||a)$ where U is the uniform distribution. For such a regularizer, $R(\pi_\theta(s)) \to \infty$ if $\|\theta_s\| \to \infty$, implying that $\ell(\theta)$ is coercive and therefore has a non-degenerate minimizer (corresponding to a policy in the interior of the probability simplex). Lemma 2 shows that gradient descent converges to a stationary point of $\ell(\theta)$.

It is straightforward to verify that $\theta \mapsto \ell(\pi_\theta)$ has no suboptimal stationary points. For this discussion, we view θ and π as stacked vectors rather than matrices. Note that $\frac{\partial}{\partial \theta} \ell(\pi_\theta) = \frac{\partial \ell(\pi_\theta)}{\partial \pi} \cdot \frac{\partial \pi_\theta}{\partial \theta}$ where $\frac{\partial \pi_\theta}{\partial \theta}$ is a Jacobian matrix. Our argument uses two facts. First, we showed in the previous section that $\pi \mapsto \ell(\pi)$ has no suboptimal stationary points, so $G = \frac{\partial}{\partial \pi} \ell(\pi_\theta) \neq 0$ if π_θ is suboptimal. Second, the Jacobian matrix $\frac{\partial \pi_\theta}{\partial \theta}$ has full rank, so there exits a (natural gradient) direction $N \in \Theta$ with $\frac{\partial \pi_\theta}{\partial \theta} N = G$. (Essentially, this means we can move the policy in direction G by moving the parameter in direction N.) Then, the directional derivative can be calculated as $\langle \frac{\partial}{\partial \theta} \ell(\pi_\theta), N \rangle = \|G\|^2 > 0$, showing that $\frac{\partial}{\partial \theta} \ell(\pi_\theta) \neq 0$, showing that θ is not a stationary point.

Optimal stopping with threshold policies.

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

Example 5 (Optimal Stopping). *The optimal stopping problem is most naturally formulated as a reward maximization problem[6]. In each round the agent observes a state variable x_t taking*

---
[6]One can imagine costs as negative reward to be consistent with our formulation.

values in a finite context set $\mathcal{X}$, which evolves according to an uncontrolled Markov chain with time-homogeneous transition probabilities $\mathbb{P}(x_{t+1} = x'|x_t = x) = p(x'|x)$. Conditioned on x_t, the agent receives an offer y_t drawn i.i.d from some density $q_{x_t}(\cdot)$ supported over $\mathcal{Y}$, i.e. $q_x(y) > 0 \; \forall \, y \in \mathcal{Y}$. We assume q_{x_t} has a continuous derivative and the offer set $\mathcal{Y} = [y_{\min}, y_{\max}]$ is an interval in $\mathbb{R}$ with $y_{\min} > 0$. If the offer is accepted in round t, the process terminates and the agent accrues a reward of $\gamma^t y_t$. Rejecting the offer in any round is costless. The agent's objective is to maximize the expected revenue.

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

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

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

Details for this example are provided in Appendix D.4. For instance, it is easy to show closure of the policy class. For a threshold policy π, a policy iteration step updates to yet another threshold policy which accepts an offer y in x if and only if it exceeds the continuation value, $c_\pi(x) = \gamma \mathbb{E}[J_\pi(x_{t+1}, y_{t+1})|x_t = x]$. See Appendix D.4 for details where we also show how Conditions 0, 1 and 2.A hold. The next lemma shows that Condition 2.B also applies. The gradient dominance constant depends on a measure of the degree of uniformity in the offer distribution $q_x(\cdot)$. We know β is finite because $\mathcal{X}$ is finite, $q_x(y) > 0$ for each $y \in \mathcal{Y}$ (by assumption), and $\mathcal{Y}$ is compact.

Lemma 9 (Gradient dominance for optimal stopping). *Consider the optimal stopping problem formulated in Example 5. For any $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta|\eta_\pi, J_\pi)$ is $(\beta, 0)$–gradient-dominated where $\beta = \max_{x \in \mathcal{X}, y \in \mathcal{Y}} q_x(y) / \min_{x \in \mathcal{X}, y \in \mathcal{Y}} q_x(y)$.*

Finally, we also establish smoothness properties for the optimal stopping problem, ensuring that the convergence rates in Lemma 3 apply here as well.

Lemma 10. *For the optimal stopping problem in Example 5, $\max_{\theta \in \Theta} \|\nabla^2 \ell(\theta)\| < \infty$.*

Finite horizon inventory control with base stock policies.

We now apply Theorem 3 to a finite horizon inventory control problem with the class of multi-period base stock policies. Kunnumkal and Topaloglu [2008] previously showed through a somewhat intricate analysis that a stochastic approximation algorithm converges to the optimal policy, despite non-convexity of the objective.

Example 6 (Finite horizon inventory control). *Consider a multi-period inventory control problem (also popularly known as the multi-period newsvendor problem) with backlogged demands. In time period t, the seller selects a non-negative quantity $a_t \geq 0$ of inventory to order on the basis of the current inventory level $x_t \in \mathbb{R}$. After ordering inventory, a random i.i.d demand w_t is realized and*

the inventory level evolves as: $x_{t+1} = x_t + a_t - w_t$. We assume the demand distribution has a density supported over some bounded set $[0, w_{\max}]$. Negative inventory levels correspond to backlogged demand that is filled when additional inventory becomes available.

The seller begins at some stage h_0 *in the initial period and the stage advances by one in each time period, i.e. $h_{t+1} = h_t + 1$, until a final state H is reached[7]. The seller's objective is to minimize total expected cost over the horizon,*

$$\mathbb{E}\left[\sum_{t=0}^{H-h_0} \left(ca_t + b\max\{x_t + a_t - w_t, 0\} + p\max\{-x_t - a_t + w_t, 0\}\right)\right] \tag{17}$$

where $c, b, p > 0$ denote the per unit costs of ordering, holding and backlogging items, respectively. We assume that $p > c$. Otherwise, the optimal policy may never order in any period. It is well known that a (multi-period) base-stock policy is optimal for this setting [Bertsekas, 1995].

Let $s_t = (x_t, h_t)$ denote the state at time t, encoding all information needed to make an optimal ordering decision. A base-stock policy depends on a vector of target inventory levels $(\theta_0, \cdots, \theta_{H-1}) \in \mathbb{R}_+^H$ At state (x, h), the policy orders inventory $\pi_\theta((x, h)) = \max\{0, \theta_h - x\}$. That is, it orders enough inventory to reach a target level θ_h, whenever feasible. We restrict to the policy class $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$ with bounded parameter space $\Theta = [0, 2w_{\max}]^H$. With a selection of $\theta_{h_t} = 2w_{\max}$, the seller can ensure $x_{t+1} \geq w_{\max}$ is large enough to meet all demand with probability one. Inventory levels above this are clearly suboptimal, due to their excessive holding costs. As the demand distribution is bounded and the target inventory level is in $[0, 2w_{\max}]$ the feasible inventory levels at every period are trivially bounded in $\mathcal{I} = [-w_{\max}, 2w_{\max}]$, as long as the initial inventory level does not fall outside this set. The initial state (x_0, h_0) is drawn from the initial distribution ρ. We assume that h_0 has a positive probability of taking on any value in $\{1, \cdots, H\}$ and that, conditioned on h_0, the distribution of x_0 has a continuous density that is supported on $\mathcal{I}$.

This is clearly an example of a finite horizon problem with a non-stationary policy class and hence has the structure noted in Condition 3. The differentiability conditions follows essentially from using the Leibniz rule along with the fact that base-stock policies are differentiable everywhere except at a single point. Condition 4 can easily verified too, using a classic result in inventory control theory which shows that the optimal state-action cost-to-go function, $Q^*(s, a)$ is convex in a [Bertsekas, 1995].

7 Bounds on the concentrability coefficient

To show a gradient dominance result, we need to relate the optimality gap to the magnitude of errors in the Bellman equation. In Section 5.3, we defined the effective concentrability coefficient κ_ρ for the set of cost-to-go functions $\mathcal{J}_\Theta = \{J_{\pi_\theta} : \theta \in \Theta\}$ to be the smallest scalar such that

$$\|J - J^*\|_{1,\rho} \leq \frac{\kappa_\rho}{(1-\gamma)}\|J - TJ\|_{1,\rho} \quad \forall J \in \mathcal{J}_\Theta. \tag{18}$$

To motivate our definition, note that whenever the Bellman operator is a contraction in some norm $\|\cdot\|$ with modulus γ, the following inequality holds: $\|J - J^*\| \leq (1-\gamma)^{-1}\|J - TJ\|$ (see (21)

---

[7]This construct of stages to be consistent with Section 5.4, where the simulator may start at later stages of the decision problem.

in Appendix B). For bounded cost problems, T is a contraction in the maximum norm and can sometimes also be shown to be contractive in a weighted norm which reflects state relevance. See Bertsekas [1995] or Puterman [2014]. Essentially, the constant κ_ρ enables the above inequality in the weighted norm $\| \cdot \|_{1,\rho}$, in which T is typically not contractive.

The focus on this norm is motivated by two factors. First, the optimality gap 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 (18) modulo a constant factor. Second, the policy gradient theorem in Lemma 6 reveals a natural dependence on the errors in Bellman's equation weighted under the state occupancy measure η_π. As $\eta_\pi \succeq (1 - \gamma)\rho$, it makes sense to measure the Bellman errors in $\| \cdot \|_{1,\rho}$. It is worth noting that, because our definition of κ_ρ depends only on the subclass of cost-to-go functions, $\mathcal{J}_\Theta$ allows for stronger bounds when these functions obey certain regularity properties. We now provide several useful upper bounds on κ_ρ. See Appendix B.3 for proof details.

Theorem 4. *The following results apply under the general problem formulation in Section 2.*

   (a) If $\mathcal{S}$ is finite, then $\kappa_\rho \leq 1/\left(\min_{s \in \mathcal{S}} \rho(s)\right)$.

   (b) Let π^* denote any optimal stationary policy. Then, $\kappa_\rho \leq \left\| \frac{d\eta_{\pi^*}}{d\rho} \right\|_\infty$.

   (c) $\kappa_\rho \leq C/c$ if T is a contraction with modulus γ in a norm $\| \cdot \|$ that satisfies

$$c\|J\| \leq \|J\|_{1,\rho} \leq C\|J\| \qquad \forall J \in \mathcal{J}_\Theta. \tag{19}$$

The bound in part (a) is simple and can be derived as a special case of the result in part (b). The bound in part (b) depends on the worst-case likelihood ratio between the state occupancy measure under the optimal policy and the initial distribution. Note, $\frac{d\eta_{\pi^*}}{d\rho}$ is the Radon-Nikodym derivative term, which exists because of Assumption 1. Put differently, this result implies that $\kappa_\rho \leq C$ if $\eta_{\pi^*}(\mathcal{M}) \leq C\rho(\mathcal{M})$ for every measurable set $\mathcal{M} \subset \mathcal{S}$. This result is, essentially, a restatement of a key observation in Kakade and Langford [2002]. Such distributional mismatch terms also appears in the works of Agarwal et al. [2020], Scherrer and Geist [2014].

The result in part (c) gives an alternative approach to bounding κ_ρ. It 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 which aligns with that norm's state weighting. The optimal stopping problem is one such special case where it can be shown that T is a contraction in $\| \cdot \|_{1,\mu}$, where μ is the stationary distribution of the underlying Markov chain – assuming it is never interrupted by stopping. Choosing $\rho = \mu$ implies $\kappa_\rho \leq 1$ using (19). In practical problems, one could easily sample initial states from ρ by simulating this Markov process.

Lemma 11. *For the optimal stopping problem in Example 5, consider a policy π_C that never stops, i.e. $\pi_C(s) = 1$ for all $s \in \mathcal{S}_C$. Let μ be a stationary distribution of the induced Markov process, meaning $\mu(\mathcal{M}) = \int P(\mathcal{M}|s', 1)\mu(ds')$ for any $\mathcal{M} \subset \mathcal{S}$. Then, choosing $\rho = \mu$ implies $\kappa_\rho \leq 1$.*

The LQ control problem technically falls outside the scope of our general formulation as per-stage costs are not bounded. Therefore, we restrict our attention to the cost-to-go functions corresponding to stable linear policies. For this result, we are able to leverage certain regularity properties of this which imply better bounds than the generic bound implied by part (b) of Theorem 4. In particular, because the class of cost-to-go functions induced by linear policies are quadratic, we need only the initial distribution to explore the basis of the state space sufficiently, rather than requiring it to almost perfectly mimic the steady state distribution of the (unknown) optimal policy.

Lemma 12. *Consider the LQ control problem formulated in Example* 2. *Define* $\Sigma_\rho := \mathbb{E}_\rho \left[s_0 s_0^\top \right]$ *and let* $\theta^* \in \mathbb{R}^{n \times k}$ *denote the parameter of an optimal policy. Then,*

$$\|J - J^*\|_{1,\rho} \leq \frac{\kappa}{(1 - \gamma)} \|J - TJ\|_{1,\rho} \qquad \forall J \in \{J_{\pi_\theta} : \theta \in \Theta_S\}$$

when

$$\kappa = \frac{(1 - \gamma)}{1 - \gamma \|A + B\theta^*\|_2^2} \cdot \frac{\lambda_{\max}(\Sigma_\rho)}{\lambda_{\min}(\Sigma_\rho)}. \tag{20}$$

The bound above depends on $\|A + B\theta^*\|_2$ for the optimal linear policy θ^* as well as the condition number of the second moment matrix under ρ. Recall, we assumed the system to be controllable which implies that an optimal policy satisfies $\|A + B\theta^*\|_2 < 1/\sqrt{\gamma}$. This ensures that the first term $\frac{(1-\gamma)}{1 - \gamma \|A + B\theta^*\|_2^2}$ in (20) is finite, less than 1 when $\|A + B\theta^*\|_2 \leq 1$, and becomes large in problems where θ^* is only barely stable.

8 Closure under approximate policy improvement.

So far, our results crucially depend on the closure property of the policy class, which applies to many classical dynamic programming problems with structured policy classes. A natural question to ask is whether we can relax this closure condition. In this section, we present results for the case where our policy class is only *approximately* closed under improvement. One would expect expressive policy classes such as 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] to follow this condition. Recall that Π denotes the class of all stationary policies and Π_Θ denotes the parameterized policy class over which we search.

Condition 5 (Closure under approximate policy improvement). *There exists $\epsilon \geq 0$ such that for every $\pi \in \Pi_\Theta$,*

$$\min_{\pi^+ \in \Pi_\Theta} \mathcal{B}(\pi^+ | \eta_\pi, J_\pi) \leq \min_{\pi' \in \Pi} \mathcal{B}(\pi' | \eta_\pi, J_\pi) + \epsilon.$$

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 $T_{\pi^+} J_\pi(s) = TJ_\pi(s)$ for every $s \in \mathcal{S}$. Condition 5 measures the deviation from this ideal case, in a norm that weights states by the discounted state occupancy measure under the current policy. We refer to ϵ as the *inherent Bellman error* of the policy class.

To motivate our theory, let's take the example of state aggregation which has a long history of being employed in reinforcement learning [Tsitsiklis and Van Roy, 1996, Van Roy, 2006]. For example, consider a continuous state, finite action MDP. A state-aggregation is a partition of of the state space $\mathcal{S} = \cup_{i=1}^m \mathcal{S}_i$ into m disjoint subsets. A state-aggregated policy π_θ can be described by a parameter $\theta = (\theta_1, \cdots, \theta_m)$ where $\pi_\theta(s) = \theta_i \in \Delta^{k-1}$ prescribes a common action distribution for every $s \in \mathcal{S}_i$. Our result applies when this policy class has low inherent Bellman error.

We state our formal result in Theorem 5 below to show that for highly flexible policy classes, any stationary point of the policy gradient objective is nearly optimal, where this optimality gap is a function of the inherent Bellman error in Condition 5. Our result 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. Recall the definition of κ_ρ in (13), which relates the optimality gap to errors in the Bellman equation weighted under ρ.

Theorem 5. *Suppose Conditions 0, 2.A and 5 hold. Then, ℓ is continuously differentiable and any stationary point θ of $\ell(\cdot)$ satisfies,*

$$\ell(\pi_\theta) - \ell(\pi^*) \leq \frac{\kappa_\rho}{(1-\gamma)} \cdot \epsilon$$

Proof. Suppose θ is a stationary point of $\ell : \Theta \to \mathbb{R}$. Under Conditions 2.A and 5, we have

$$\min_{\pi^+ \in \Pi_\Theta} \mathcal{B}(\pi^+ | \eta_{\pi_\theta}, J_{\pi_\theta}) - \min_{\pi' \in \Pi} \mathcal{B}(\pi' | \eta_{\pi_\theta}, J_{\pi_\theta}) = \min_{\pi \in \Pi_\Theta} \left(\int [T_\pi J_{\pi_\theta}] \, d\eta_{\pi_\theta} \right) - \int [T J_{\pi_\theta}] \, d\eta_{\pi_\theta}$$

$$= \int [J_{\pi_\theta} - T J_{\pi_\theta}] \, d\eta_{\pi_\theta}$$

$$= \|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1, \eta_{\pi_\theta}} \leq \epsilon.$$

where the second equality follows from Lemma 7 and the final equality uses that $J_{\pi_\theta} \succeq T J_{\pi_\theta}$ for any $\pi_\theta \in \Pi_\Theta$. Then, we have

$$\ell(\pi_\theta) - \min_\pi \ell(\pi) = (1-\gamma) \int [J_{\pi_\theta} - J^*] \, d\rho = (1-\gamma)\|J_{\pi_\theta} - J^*\|_{1, \rho}$$

$$\leq \kappa_\rho \|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1, \rho}$$

$$\leq \frac{\kappa_\rho}{(1-\gamma)} \|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1, \eta_{\pi_\theta}}$$

$$= \frac{\kappa_\rho \cdot \epsilon}{(1-\gamma)}.$$

The first inequality follows from the definition of κ_ρ while the second uses that $\eta_{\pi_\theta} \succeq (1-\gamma)\rho$. $\square$

9 Conclusion

In this paper, we uncover structural properties of the underlying MDP which guarantee that policy gradient methods converge to globally optimal solutions as well as characterize their convergence rates, even though the optimization objective is non-convex. Our results rely on a key connection with policy iteration, a well studied dynamic programming algorithm which solves a single period optimization problem at every step and often has special structure. We show how the policy gradient objective inherits this nice structure, making it amenable for gradient based algorithms to find the optimal policy at a fast rate. There are a number of research directions to extend our work, including results for the function approximation setting with neural network or kernel based parameterization of the cost-to-go function as well as designing principled exploration approaches which can be efficiently combined with policy gradient methods.

10 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 measure functions on $\mathcal{S}$.
$g(s, a)$	$\triangleq$	Single period expected cost of action a in state s.
$P(\mathcal{M}|s, a)$	$\triangleq$	Transition probability to set $\mathcal{M} \subset \mathcal{S}$
g_π	$\triangleq$	Single period cost function 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.
η_π	$\triangleq$	The discounted state occupancy measure under policy π.
$\ell(\pi) = \int J_\pi d\rho$	$\triangleq$	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 $\ell(\pi_\theta)$.
$\mathcal{B}(\pi'|\eta, J_\pi)$	$\triangleq$	"Bellman" objective or the weighted policy iteration objective
$\mathcal{B}(\theta|\eta, J_\pi) = \mathcal{B}(\pi_\theta|\eta, J_\pi)$	$\triangleq$	Overloaded notation for the policy iteration objective at π_θ.
κ_ρ	$\triangleq$	Effective concentrability coefficient.
$\|J\|_\infty$	$\triangleq$	Max-norm $\sup_s |J(s)|$.
$\|J\|_{1,\eta}$	$\triangleq$	Weighted 1-norm $\int |J(s)| \, d\eta$.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 09–12 Jul 2020.

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.

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.

Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media, 2007.

Mohammad Gheshlaghi Azar, Rémi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Reinforcement learning with a near optimal rate of convergence. 2011.

J Andrew Bagnell, Sham M Kakade, Jeff G Schneider, and Andrew Y Ng. Policy search by dynamic programming. In *Advances in neural information processing systems*, pages 831–838, 2004.

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

Amir Beck. *Convergence rate analysis of gradient based algorithms*. PhD thesis, Tel-Aviv University, 2002.

Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Dimitir P Bertsekas and Steven Shreve. *Stochastic optimal control: the discrete-time case*. 1978.

Dimitri P Bertsekas. *Dynamic programming and optimal control*. 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.

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.

David Blackwell. Discounted dynamic programming. *The Annals of Mathematical Statistics*, 36(1): 226–235, 1965.

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

Michael Caramanis and George Liberopoulos. Perturbation analysis for the design of flexible manufacturing system flow controllers. *Operations Research*, 40(6):1107–1125, 1992.

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.

Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

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.

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational inverse control with events: A general framework for data-driven reward definition. In *Advances in Neural Information Processing Systems*, pages 8538–8547, 2018.

Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13: 575–616, 2006.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online 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.

Paul Glasserman and Yu-Chi Ho. *Gradient estimation via perturbation analysis*, volume 116. Springer Science & Business Media, 1991.

Paul Glasserman and Sridhar Tayur. Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Science*, 41(2):263–281, 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.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.

G Hewer. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 16(4):382–384, 1971.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

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.

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.

Sven Koenig and Reid G Simmons. Complexity analysis of real-time reinforcement learning. In *AAAI*, pages 99–107, 1993.

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.

Pierre L'Ecuyer, Nataly Giroux, and Peter W Glynn. Stochastic optimization by simulation: numerical experiments with the m/m/1 queue in steady-state. *Management science*, 40(10):1245–1261, 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.

Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information Fusion*, 58:52–68, 2020.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pages 855–863, 2014.

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

Matthew S Maxwell, Shane G Henderson, and Huseyin Topaloglu. Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic Systems*, 3(2): 322–361, 2013.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

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.

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

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

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.

G Ch Pflug. On-line optimization of simulated markovian processes. *Mathematics of Operations Research*, 15(3):381–395, 1990.

Georg Ch Pflug. Derivatives of probability measures-concepts and applications to the optimization of stochastic systems. In *Discrete Event Systems: Models and Applications*, pages 252–274. Springer, 1988.

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.

Tankred Rautert and Ekkehard W Sachs. Computational design of optimal output feedback controllers. *SIAM Journal on Optimization*, 7(3):837–852, 1997.

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.

Chang-Han Rhee and Peter Glynn. Lyapunov conditions for differentiability of markov chain expectations: the absolutely continuous case. *arXiv preprint arXiv:1707.03870*, 2017.

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.

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 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.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

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.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

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.* 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.

Kalyan T Talluri and Garrett J Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.

Sebastian B Thrun. E cient exploration in reinforcement learning. Technical report, Technical Report CMU-CS-92-102, School of Computer Science, Carnegie Mellon . . . , 1992.

Hannu T Toivonen. A globally convergent algorithm for the optimal constant output feedback problem. *International Journal of Control*, 41(6):1589–1599, 1985.

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

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

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

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

Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real-world robotic reinforcement learning. *arXiv preprint arXiv:2004.12570*, 2020.

A   On the necessity of an exploratory initial distribution

Our results critically rely on using an exploratory initial distribution (see Assumption 1). This is not an artifact of the proof techniques and it is well known that, in the absence of strong assumptions on the transition kernel, policy gradient methods have poor convergence properties if applied without some form of sophisticated exploration. While this aspect of policy gradient methods is not always highlighted in the literature, many applied papers assume access to a diverse set of starting states using either explicit restarts [Fu et al., 2018, Haarnoja et al., 2018] or some form of continual learning that aims to increase the support of a training distribution [Lesort et al., 2020, Zhu et al., 2020].

The following example, which is commonly known as a "chain" MDP [Kakade and Langford, 2002, Thrun, 1992] or the "river swim" problem [Osband et al., 2013, Strehl and Littman, 2008], illustrates the challenges for policy gradient in the absence of sufficient exploration. Many other examples in the reinforcement learning literature, like the "combination lock" problem [Koenig and Simmons, 1993] and the "grid world" problem [Azar et al., 2011] highlight the same issue. While these examples are typically used to highlight a statistical challenge, here we focus on the optimization landscape. This example is partly inspired by one in Kakade and Langford [2002]. A similar discussion appears also in Agarwal et al. [2020]. We include this section to keep the paper self contained. In addition, it does not seem that past work has shown clearly that $\ell(\cdot)$ may have suboptimal local minima in the absence of an exploratory initial distribution, instead showing the existence of suboptimal polices with small but nonzero gradient norm.
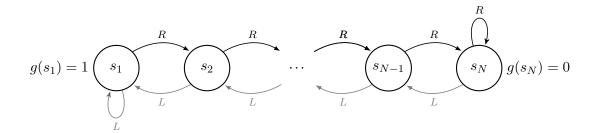
Figure 2: A simple chain MDP example to illustrate how policy gradient methods face suboptimal local minima in the absence of an exploratory initial distribution.

Example 7. *Consider the MDP shown in Figure 2. There are N states and the agent can move either left (L) or right (R) from each state. The agent always begins in the leftmost state (i.e. $\rho(s_1) = 1$). She incurs a cost of 2 per-period when in any state other than the leftmost or rightmost state, a cost $g(s_1) = 1$ from the leftmost state and a cost of $g(s_N) = 0$ per period in the rightmost state. A stationary policy $\pi \in [0,1]^N$ is a vector[8] where $\pi(s)$ specifies the probability of choosing the action R in state s. When the horizon is sufficiently long, the optimal policy moves right in each period. From Lemma 6, one can calculate the policy gradient as*

$$\frac{\partial \ell(\pi)}{\partial \pi(s)} = \eta_\pi(s) \left(Q_\pi(s, R) - Q_\pi(s, L) \right).$$

We argue that a suboptimal policy π that always moves left, i.e. $\pi(s_i) = 0 \;\forall\, i \in \{1, \dots, N\}$, is a local minimum of $\ell(\cdot)$. To see this, first note that the agent will always start and stay in the leftmost state, so $\eta_\pi(s_i) = 0$ when $i \geq 2$. The only possible nonzero component of $\nabla \ell(\pi)$ is the first term corresponding to state s_1. Therefore, for any policy $\pi' \in [0,1]^N$,

$$\langle \nabla \ell(\pi), \pi' - \pi \rangle = \eta_\pi(s_1) \left(Q_\pi(s_1, R) - Q_\pi(s_1, L) \right) \left(\pi'(s_1) - \pi(s_1) \right) \geq 0,$$

which follows as $Q_\pi(s_1, R) > Q_\pi(s_1, L)$, given that moving to s_2 for a single period is more costly than staying in s_1 and the fact that $\pi(s_1) = 0$, so $\pi'(s_1) - \pi(s_1) \geq 0$ for any feasible policy π'.

Similar issues arise under a (non-degenerate) stochastic policy. The main idea is that policies which are more likely to move left from every state are expected to require exponentially (in the number of states) many periods to reach the rightmost state. An explicitly bound confirming that the policy gradient can be exponentially small in N is shown in Agarwal et al. [2020].

B Omitted proofs.

In this section, we provide proofs for some of the main results along with the supporting lemmas.

B.1 General results.

We prove some key lemmas which are used to show our general results in Theorems 1 and 2. We start with some useful background on Bellman operators.

---

[8]Note that unlike Example 1, this policy class is closed under policy improvement.

Bellman operators. For bounded cost-to-go functions, Bellman operators are monotone, meaning that $J \preceq J'$ implies $TJ \preceq TJ'$ and $T_\pi J \preceq T_\pi J'$, and contractive in $\|\cdot\|_\infty$ with modulus γ. A useful consequence of contractivity relates optimality gap to errors in the cost-to-go functions.

$$\|J_\pi - J^*\|_\infty \leq \frac{1}{1-\gamma} \|J_\pi - TJ_\pi\|_\infty \tag{21}$$

where J^* is the optimal cost-to-go function. A simple argument [Bertsekas, 1995] shows (21).

$$\|J_\pi - J^*\|_\infty = \|T_\pi J_\pi - TJ_\pi + TJ_\pi - J^*\|_\infty \leq \|T_\pi J_\pi - TJ_\pi\|_\infty + \|TJ_\pi - TJ^*\|_\infty$$
$$\leq \|T_\pi J_\pi - TJ_\pi\|_\infty + \gamma\|J_\pi - J^*\|_\infty.$$

Optimal policies and minimizers of the policy gradient loss. For the reader's convenience, we recall Lemma 1, which relates minimizers of $\ell(\cdot)$ to the classic definition of optimal policies in dynamic programming.

Lemma 1. *A policy satisfies $\pi \in \arg\min_{\pi' \in \Pi} \ell(\pi')$ if and only if $J_\pi = J^*$ ρ–almost surely, i.e. $\rho(\{s \in \mathcal{S} : J_\pi(s) = J^*(s)\}) = 1$.*

Proof. Recall $\ell(\pi) = \int J_\pi(s)\rho(ds)$. An optimal policy π^* satisfies $J_{\pi^*}(s) = J^*(s)$ for every state $s \in \mathcal{S}$. Since $J_\pi(s) \geq J^*(s)$ for each $s \in \mathcal{S}$, we have

$$\ell(\pi) - \ell(\pi^*) = \int \left(J_\pi(s) - J^*(s)\right) \rho(ds) \geq 0. \tag{22}$$

Since this holds for every policy π, it is clear that $\ell(\pi^*) = \min_{\pi \in \Pi} \ell(\pi)$.

A basic fact in measure theory states that, for a non-negative function $J : \mathcal{S} \to \mathbb{R}_+$, $\int J d\rho = 0$ if and only if $J = 0$ ρ-almost surely. Since $J_\pi(s) \geq J^*(s)$ for each $s \in \mathcal{S}$, applying this fact with a choice of $J = J_\pi - J^*$ implies equality holds in (22) if and only if $J_\pi - J^* = 0$ ρ-almost-surely. $\quad\square$

Performance difference and telescoping sums. Throughout the analysis, we use a basic result which relates the difference in cost-to-go functions to the gap in Bellman's equation at future states. For any two cost-to-go functions, $J_\pi, J \in \mathcal{J}$, and any starting state $s_0 \in \mathcal{S}$, we have

$$J_\pi(s_0) - J(s_0) = T_\pi J(s_0) - J(s_0) + T_\pi J_\pi(s_0) - T_\pi J(s_0)$$
$$= T_\pi J(s_0) - J(s_0) + \gamma P_\pi \left(J_\pi(s_0) - J(s_0)\right)$$

where we use that $(P_\pi J)(s) = \int J(s')P(ds'|s, \pi(s))$. Taking expectation over some initial distribution ν and noting that $(P_\pi J)(s_t) = \mathbb{E}^\pi [J(s_{t+1})|s_t]$, we have

$$\mathbb{E}_\nu^\pi [J_\pi(s_0) - J(s_0)] = \mathbb{E}_\nu^\pi \left[\sum_{t=0}^\infty \gamma^t \left[T_\pi J(s_t) - J(s_t)\right]\right], \tag{23}$$

where we use the tower property of conditional expectation to simplify the telescoping sum. Kakade and Langford [2002] use this to give a particularly convenient form, which is commonly known as the *performance difference lemma*. Choosing $J = J_{\bar\pi}, \nu = \rho$ in (23) and recalling that $\ell(\pi) = (1-\gamma)\mathbb{E}_\rho[J_\pi(s_0)]$ gives

$$\ell(\pi) - \ell(\bar\pi) = (1-\gamma)\mathbb{E}_\rho^\pi \left[\sum_{t=0}^\infty \gamma^t \left(T_\pi J_{\bar\pi}(s_t) - J_{\bar\pi}(s_t)\right)\right] = \int [T_\pi J_{\bar\pi} - J_{\bar\pi}] \, d\eta_\pi. \tag{24}$$

The second equality follows using the definition of the discounted state occupancy measure, $\eta_\pi(\mathcal{M}) = (1-\gamma)\mathbb{E}_\rho^\pi\left[\sum_{t=0}^\infty \gamma^t \mathbf{1}(s_t \in \mathcal{M})\right]$ for any measurable set $\mathcal{M} \subset \mathcal{S}$.

A policy gradient formula. We give a short of the policy gradient theorem in Lemma 6 assuming the differentiability conditions hold.

Lemma 6 (Policy gradient theorem). *Under Condition 0, $\ell(\theta)$ is continuously differentiable and*

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

Proof. Recall that $\mathcal{B}(\theta|\eta_{\bar{\theta}}, J_\theta) = \int J_{\pi_\theta} d\eta_{\pi_{\bar{\theta}}}$. From the performance difference lemma in (24), we have

$$\ell(\bar{\theta}) - \ell(\theta) = \int \left[T_{\bar{\theta}}J_\theta - J_\theta\right] d\eta_{\bar{\theta}} = \mathcal{B}(\bar{\theta}|\eta_{\bar{\theta}}, J_\theta) - \mathcal{B}(\theta|\eta_{\bar{\theta}}, J_\theta).$$

Expanding the total derivative in terms of partial derivatives gives,

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

$\square$

On average Bellman equation. We prove the on average Bellman equation which is a key lemma we used for proving Theorem 1.

Lemma 8 (On average Bellman equation). *For any $\pi \in \Pi$,*

$$\ell(\pi) = \ell(\pi^*) \iff \int (J_\pi - TJ_\pi)\, d\rho = 0$$

Proof. Recall that a non-negative function f satisfies $\int f d\mu = 0$ if and only if $f = 0$ almost surely under the probability distribution μ. We also use the fact that $J_\pi \succeq J^*$, by definition of the optimal cost-to-go, and $J_\pi \succeq TJ_\pi$, as shown in (5).

To show the left hand side, recall that by definition, $\ell(\pi) - \ell(\pi') = (1-\gamma)\int (J_\pi - J_{\pi'})\, d\rho$. Therefore, $\ell(\pi) = \ell(\pi^*)$ implies that,

$$0 = \int (J_\pi - J^*)\, d\rho \geq \int (J_\pi - J_{\pi^+})\, d\rho \overset{(a)}{=} (1-\gamma)^{-1}\int (J_\pi - T_{\pi^+}J_\pi)\, d\eta_{\pi^+}$$

$$= (1-\gamma)^{-1}\int (J_\pi - TJ_\pi)\, d\eta_{\pi^+}$$

$$\geq \int (J_\pi - TJ_\pi)\, d\rho \geq 0.$$

where we take π^+ to be the policy iteration update at π, i.e. $T_{\pi^+}J_\pi = TJ_\pi$ and (a) follows by using the performance difference lemma in (24) with $\bar{\pi} = \pi^+$. The penultimate inequality uses that $\eta_{\pi^+} \succeq (1-\gamma)\rho$ while the final inequality follows by using that $J_\pi \succeq TJ_\pi$.

To show the other side, suppose $\int (J_\pi - TJ_\pi)d\rho = 0$. Let $S_0 = \{s : J_\pi(s) - TJ_\pi(s) = 0\}$ and S_0^c denote its complement. As we assumed η_{π^*} to be absolutely continuous with respect to ρ, we have that $\rho(S_0^c) = 0 \implies \eta_{\pi^*}(S_0^c) = 0$. Therefore,

$$\int (J_\pi - TJ_\pi)d\rho = 0 \implies \int (J_\pi - TJ_\pi)d\eta_{\pi^*} = 0.$$

As $J_\pi \succeq J^*$, we have $\ell(\pi) - \ell(\pi^*) = (1 - \gamma)\int (J_\pi - J^*)\,d\rho \geq 0$. Then, we get our result by noting

$$0 \leq \ell(\pi) - \ell(\pi^*) \overset{(b)}{=} \int (J_\pi - T_{\pi^*}J_\pi)\,d\eta_{\pi^*} \leq \int (J_\pi - TJ_\pi)\,d\eta_{\pi^*} = 0$$

where (b) follows from the performance difference lemma in (24). $\qquad\square$

B.2 Non-stationary policy classes: Proof of Theorem 3

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

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

Proof. To give a more transparent proof, it is helpful to develop some notation that highlights a (limited) sense in which the problem decomposes across time periods. With some abuse of notation[9], for any parameter vector $\theta = (\theta_1, \cdots, \theta_H)$, define $\pi_{\theta_h} : \mathcal{S}_h :\to \mathcal{A}$ to be the restriction of the policy π_θ to $\mathcal{S}_h$, i.e. $\pi_{\theta_h}(s) = \pi_\theta(s)$ for all $s \in \mathcal{S}_h$.

Single period PI objectives. Similarly, define

$$\mathcal{B}_h(\theta_h \mid \eta, J_\pi) = \int_{\mathcal{S}_h} Q_\pi(s, \pi_{\theta_h}(s))\eta(ds)$$

so that

$$\mathcal{B}(\theta \mid \eta, J_\pi) = \sum_{h=1}^{H} \mathcal{B}_h(\theta_h \mid \eta, J_\pi).$$

Because the parameter space factorizes as $\Theta = \Theta_1 \times \cdots \times \Theta_H$, this separability of the weighted Bellman objective implies

$$\theta \in \arg\min_{\overline{\theta} \in \Theta} \mathcal{B}(\overline{\theta} \mid \eta, J_\pi) \iff \theta_h \in \arg\min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h(\overline{\theta}_h \mid \eta, J_\pi) \; \forall h. \qquad (25)$$

Single period characterization of stationary points. The policy gradient formula in Lemma 6 states $\nabla_\theta \ell(\theta) = \nabla_{\overline{\theta}} \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})|_{\overline{\theta}=\theta}$. Therefore, θ is a stationary point of $\min_{\overline{\theta} \in \Theta} \ell(\overline{\theta})$ if and only if it is a the stationary points of the optimization problem $\min_{\overline{\theta} \in \Theta} \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J_{\pi_\theta})$. Because $\Theta = \Theta_1 \times \cdots \times \Theta_H$, this problem separates across time periods, and we find:

$$\theta \text{ is stationary for } \min_{\overline{\theta} \in \Theta} \ell(\overline{\theta}) \iff \theta_h \text{ is a stationary for } \min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h(\overline{\theta}_h \mid \eta_{\pi_\theta}, J_{\pi_\theta}) \; \forall h \qquad (26)$$

---

[9]Technically, for this to be appropriate we should imagine $\Theta_1, \cdots, \Theta_H$ are disjoint, which we could assume without loss of generality.

Inductive proof. We argue that any stationary point θ of $\ell(\cdot)$ must satisfy $J_{\pi_\theta} = J^*$ ρ-almost-surely. By Lemma 1, this implies it is a minimizer of $\ell(\cdot)$. By assumption 3, $\eta_\pi \ll \rho$ for any $\pi \in \Pi$. In the reverse direction, $\eta_\pi \succeq (1-\gamma)\rho$ by definition, which implies $\rho \ll \eta_\pi$ for each $\pi \in \Pi$. Therefore, we can throughout claim that certain events hold "almost surely" without reference to whether the base measure is ρ or η_π.

We proceed by backward induction, showing $J_{\pi_\theta}(s) = J^*(s)$ almost surely for $s \in \mathcal{S}_h$. As a base of induction, consider $h = H + 1$, By definition, $J_{\pi_\theta}(s) = J^*(s) = 0$ for all $s \in \mathcal{S}_{H+1}$ as $\mathcal{S}_{H+1} = \{\tau\}$ contains a single costless absorbing state.

Now, for any $h \leq H$, suppose $J_{\pi_\theta}(s) = J^*(s)$ almost surely for $s \in \mathcal{S}_{h+1}$. We first claim that $\mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J_{\pi_\theta}) = \mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J^*)$ for all $\overline{\theta}_h \in \Theta_h$. This is a consequence of our induction hypothesis and Assumption 3. In particular,

$$
\begin{aligned}
0 \geq \mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J_{\pi_\theta}) - \mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J^*) &= \int_{\mathcal{S}_h} \left[Q_{\pi_\theta}(s, \pi_{\overline{\theta}_h}(s)) - Q^*(s, \pi_{\overline{\theta}_h}(s)) \right] \rho(ds) \\
&= \gamma \int_{\mathcal{S}_h} \int_{s' \in \mathcal{S}_{h+1}} [J_{\pi_\theta}(s') - J^*(s')] \, P(ds'|s, \pi_{\overline{\theta}_h}(s)) \, \rho(ds) \\
&\stackrel{(a)}{\geq} \frac{1}{(1-\gamma)} \int_{s' \in \mathcal{S}_{h+1}} [J_{\pi_\theta}(s') - J^*(s')] \, \eta_{\pi_{\overline{\theta}}}(ds') \\
&\stackrel{(b)}{=} 0,
\end{aligned}
$$

where we use throughout the pointwise non-positivity of $Q_{\pi_\theta} - Q^*$ or $J_{\pi_\theta} - J^*$. Inequality (a) is justified by specializing (2) for the finite horizon setting, finding that for any measurable set $\mathcal{M} \subset \mathcal{S}_{h+1}$,

$$
(1-\gamma)^{-1}\eta_{\pi_{\overline{\theta}}}(\mathcal{M}) = \left[\rho(\mathcal{M}) + \gamma \int_{s \in \mathcal{S}_h} P(\mathcal{M}|s, \pi_{\overline{\theta}}(s)) \, \rho(ds) + \dots \right] \succeq \gamma \int_{s \in \mathcal{S}_h} P(\mathcal{M}|s, \pi_{\overline{\theta}}(s)) \, \rho(ds).
$$

Inequality (b) uses that $J_{\pi_\theta} = J^*$ almost surely.

To complete the induction step, note that combining the conclusion above with the characterization of stationary points in (26) shows θ_h is a stationary point of the optimization problem $\min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J^*)$. Recall that Condition 4 assumed that $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta} | \eta_{\pi_\theta}, J^*)$ has no suboptimal stationary points. Because of the separability structure highlighted in equations (25) and (26), this condition implies that $\overline{\theta}_h \mapsto \mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J^*)$ has no bad stationary points. Hence, we have shown that $\theta_h \in \arg\min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J^*)$. Putting it all together, we have shown,

$$
\int_{\mathcal{S}_h} J_{\pi_\theta}(s)\eta_{\pi_\theta}(ds) = \int_{\mathcal{S}_h} Q_{\pi_\theta}(s, \pi_{\theta_h}(s))\eta_{\pi_\theta}(ds) = \mathcal{B}_h(\theta_h | \eta_{\pi_\theta}, J_{\pi_\theta}) = \mathcal{B}_h(\theta_h | \eta_{\pi_\theta}, J^*)
$$

$$
\begin{aligned}
&= \min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h(\overline{\theta}_h | \eta_{\pi_\theta}, J^*) \\
&\stackrel{(c)}{=} \min_{\overline{\theta}_h \in \Theta_h} \int_{\mathcal{S}_h} Q^*\left(s, \pi_{\overline{\theta}_h}(s)\right) \eta_{\pi_\theta}(ds) \\
&= \int_{\mathcal{S}_h} J^*(s)\eta_{\pi_\theta}(ds),
\end{aligned}
$$

where equality (c) applies our assumption that the policy class contains an optimal policy, i.e. there exists $\theta_h \in \Theta_h$ such that $Q^*(s, \pi_{\theta_h}(s)) = \min_{a \in \mathcal{A}_s} Q^*(s, a) = J^*(s)$ for all $s \in \mathcal{S}_h$. Since $J_{\pi_\theta} - J^*$ is pointwise non-negative and $\int_{\mathcal{S}_H} [J_{\pi_\theta} - J^*] \, d\eta_{\pi_\theta} = 0$, these functions are equal almost surely, completing the induction step. $\qquad\square$

B.3 Concentrability coefficients

Theorem 4. *The following results apply under the general problem formulation in Section 2.*

(a) *If $\mathcal{S}$ is finite, then $\kappa_\rho \leq 1/\left(\min_{s \in \mathcal{S}} \rho(s)\right)$.*

(b) *Let π^* denote any optimal stationary policy. Then, $\kappa_\rho \leq \left\|\frac{d\eta_{\pi^*}}{d\rho}\right\|_\infty$.*

(c) *$\kappa_\rho \leq C/c$ if T is a contraction with modulus γ in a norm $\|\cdot\|$ that satisfies*

$$c\|J\| \leq \|J\|_{1,\rho} \leq C\|J\| \qquad \forall J \in \mathcal{J}_\Theta. \tag{19}$$

Proof. The proof of part (a) follows as a simple corollary of the result in part (b).

Proof of part (b). Recall that π^* denotes an optimal policy. Using that $J_\pi \succeq J^*$ and the performance difference lemma in (24), we get

$$
\begin{aligned}
(1-\gamma)\int (J_\pi - J^*) \, d\rho = (1-\gamma)\|J_\pi - J^*\|_{1,\rho} = \ell(\pi) - \ell(\pi^*) &= \int (J_\pi - T_{\pi^*} J_\pi) \, d\eta_{\pi^*} \\
&\overset{(a)}{\leq} \int (J_\pi - TJ_\pi) \, d\eta_{\pi^*} \\
&\overset{(b)}{=} \int (J_\pi - TJ_\pi)\left(\frac{d\eta_{\pi^*}}{d\rho}\right) d\rho \\
&\leq \left\|\frac{d\eta_{\pi^*}}{d\rho}\right\|_\infty \int (J_\pi - TJ_\pi) \, d\rho \\
&\overset{(c)}{=} \left\|\frac{d\eta_{\pi^*}}{d\rho}\right\|_\infty \|J_\pi - TJ_\pi\|_{1,\rho}
\end{aligned}
$$

where (a) follows by using that $T_{\pi'} J \succeq TJ$ for any policy π' and each $J \in \mathcal{J}$, (b) uses definition of the Radon-Nikodym derivative and (c) follows as $J_\pi \succeq TJ_\pi$.

Proof of part (c). From (21), if T is contraction with modulus γ in $\|\cdot\|$, then $\|J - J^*\| \leq \frac{1}{(1-\gamma)}\|J - TJ^*\|$. Our result follows by noting,

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

$$\square$$

Lemmas 11 and 12 which state the concentrability coefficients for optimal stopping and LQ control are proved in Appendix D where both these examples are treated in detail.

C Convergence proofs for first order methods.

To start, let us define some standard notions from first order optimization. For a convex set $\mathcal{X} \subset \mathbb{R}^d$, we say a function $f : \mathcal{X} \to \mathbb{R}$ is k-Lipshitz if $\|f(x) - f(y)\|_2 \leq k\|x - y\|_2$ for every $x, y \in \mathcal{X}$. We say a function is L-smooth if f is differentiable throughout $\mathcal{X}$ and ∇f is L-Lipschitz. A consequence of smoothness that will be useful throughout our proofs is often called the *descent lemma*. It implies a quadratic upper bound on function values. The proof follows by Taylor expansion and the mean-value theorem [Bertsekas, 1997].

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

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

The following interpretation of projected gradient updates will be very useful for our proof. Recall the notation for orthogonal projection: $\text{Proj}_{\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \|y - x\|_2^2$. The projected gradient descent iteration can be equivalently written as

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

giving a "proximal" interpretation of projection as minimizing a local quadratic approximation. See Beck [2017] for a simple proof.

C.1 Asymptotic convergence to stationary points: proof of Lemma 2

For convenience, we first restate the claim.

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 and its β–sublevel set $\{x \in \mathcal{X} : f(x) \leq \beta\}$ is bounded for each $\beta \in \mathbb{R}$. Consider the sequence $x_{k+1} = \text{Proj}_{\mathcal{X}}\left(x_k - \alpha \nabla f(x_k)\right)$ for $k \in \mathbb{N}$.*

1. *[Beck, 2002, 2017] Assume f is differentiable on an open set containing $\mathcal{X}$ and ∇f is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant L. If $\alpha \in (0, 1/L]$, the sequence $\{x_k\}$ has at least one limit point and any limit point x_∞ is a stationary point of $f(\cdot)$ on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.*

2. *Suppose f is continuously twice differentiable on an open set containing the sublevel set $\{x \in \mathcal{X} : f(x) \leq f(x_0)\}$. For a sufficiently small $\alpha > 0$, the sequence $\{x_k\}$ has at least one limit point and any limit point x_∞ is a stationary point of $f(\cdot)$ on $\mathcal{X}$ satisfying $f(x_k) \downarrow f(x_\infty)$.*

Proof. Part 1 follows from the simple proofs in [Beck, 2002, 2017]. We show the claim in part 2. Throughout, let $\|x\|$ denotes the Euclidean norm of a vector x and $\|A\| = \max_{\|x\| \leq 2} \|Ax\|$ be induced operator norm of a matrix A. Note that the sub-level set $S := \{x \in \mathcal{X} : f(x) \leq f(x_0)\}$ is compact (continuity of $f(\cdot)$ implies its closed and we assume it to be bounded). Also, for a sufficiently small ϵ, $f(\cdot)$ is twice continuously differentiable over the compact set,

$$S_\epsilon := \{x + y : x \in S_1, \|y\| \leq \epsilon\}.$$

which follows by our assumption that $f(\cdot)$ is twice continuously differentiable on an open set containing S. We denote $G = \max_{x \in S} \|\nabla f(x)\|$ and $L = \max_{x \in S_\epsilon} \|\nabla^2 f(x)\|$. Note that G and L

are finite since $\|\nabla f\|$ and $\|\nabla^2 f\|$ are continuous over the compact sets S and S_ϵ. Fix the step-size $\alpha = \min\{\epsilon/G, 1/L\}$. For any $x \in S_1$, define $x^+ = \text{Proj}_{\mathcal{X}}(x - \alpha \nabla f(x))$. For this choice of step-size, $x^+ \in S_\epsilon$ since

$$\|x^+ - x\|_2 = \|\text{Proj}_{\mathcal{X}}(x - \alpha \nabla f(x)) - \text{Proj}_{\mathcal{X}}(x)\| \le \|\alpha \nabla f(x)\| \le \alpha G \le \epsilon,$$

which follows as projection operators are non-expansive. The optimality conditions for projection onto a convex set yield the standard property that $\hat{x} = \text{Proj}_{\mathcal{X}}(x)$ if and only if $\langle \hat{x} - x, y - \hat{x} \rangle \ge 0$ for all $y \in \mathcal{X}$. Using this and some algebra, we get

$$\langle x - \alpha \nabla f(x) - x^+, x - x^+ \rangle \le 0 \implies \|x - x^+\|^2 - \alpha \langle \nabla f(x), x - x^+ \rangle \le 0.$$

As $x^+ \in S_\epsilon$,

$$f(x^+) \le f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{L}{2}\|x^+ - x\|^2 \qquad [\text{smoothness of} f(\cdot) \text{ over } S_\epsilon]$$

$$\le f(x) + \left(\frac{L}{2} - \frac{1}{\alpha}\right)\|x^+ - x\|^2$$

$$\le f(x). \qquad [\alpha \le 1/L]$$

Since the projected gradient update reduces cost, we know $x^+ \in S$. Repeating this argument inductively shows that $f(x_{k+1}) \le f(x_k)$ and $x_k \in S$ for all k. Since $\{x_k\}$ is contained in a compact set S, it has a convergent sub-sequence, $\{x_{k_i}\}$ with some limit x_∞. We have ,

$$\lim_{k\to\infty} f(x_k) = \lim_{i\to\infty} f(x_{k_i}) = f(x_\infty),$$

where the first limit exists since $\{f(x_k)\}$ is monotone-decreasing and bounded below and the final inequality uses continuity of $f(\cdot)$. The proof to show that any limit point is a stationary point follows from [Beck, 2002, 2017]. See also [Bertsekas, 1997, Figure 3.3.2]. We omit this for brevity. $\qquad\square$

C.2 Convergence rates under gradient dominance: Proof of Lemma 3.

We first restate the claim.

Lemma 3 (Convergence rates for gradient dominated smooth functions). *Consider the problem,* $\min_{x\in\mathcal{X}} f(x)$ *where* $\mathcal{X} \subseteq \mathbb{R}^d$ *is nonempty. Assume* ∇f *is* L–*Lipschitz continuous on* $\mathcal{X}$. *Denote* $f^* = \inf_{x'\in\mathcal{X}} f(x')$. *Consider the sequence* $x_{t+1} = \text{Proj}_{\mathcal{X}}(x_t - \alpha \nabla f(x_t))$.

1. *Let* $\mathcal{X} \subset \mathbb{R}^d$ *be bounded. Set* $R = \sup_{x,x'\in\mathcal{X}} \|x - x'\|_2$ *and* $k = \sup_{x\in\mathcal{X}} \|\nabla f(x)\|_2$. *If* $\alpha \le \min\{\frac{1}{k}, \frac{1}{L}\}$ *and* f *is* $(c, 0)$–*gradient-dominated, then,*

$$f(x_T) - f^* \le \sqrt{\frac{2R^2 c\,(f(x_0) - f^*)}{\alpha T}}.$$

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

$$f(x_T) - f^* \le \left(1 - \frac{\mu}{c^2 L}\right)^T (f(x_0) - f^*).$$

Proof of Lemma 3. Recall, by Definition 2 that a function f is defined to be (c, μ)–*gradient domi- nated* over $\mathcal{X}$ if there exists a constant $c > 0$ and $\mu \ge 0$ such that

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

41

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

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

Therefore, for any $x \neq x^*$, we have $\min_{y \in \mathcal{D}} \langle \nabla f(x_t), y - x \rangle < 0$. Let $\{x_t\}$ be the iterates produced by projected gradient descent. At iterate x_t, let $\bar{y} = \arg\min_{y \in \mathcal{X}} \langle \nabla f(x_t), y - x_t \rangle$ and denote $\delta_t = \min_{y \in \mathcal{X}} \langle \nabla f(x_t), y - x_t \rangle$. Note that $\delta_t \leq 0$ and $|\delta_t| \leq \|\nabla f(x_t)\| \|y - x_t\| \leq kR$ as f is assumed to be k-Lipschitz. We take a constant stepsize, $\alpha_t = \alpha \leq \min\{\frac{1}{k}, \frac{1}{L}\}$. Then,

$$
\begin{aligned}
f(x_{t+1}) - f(x_t) &\overset{(a)}{\leq} \min_{y \in \mathcal{D}} \left[\langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2\alpha} \|y - x_t\|_2^2 \right] \\
&\overset{(b)}{=} \min_{\beta \in [0,1]} \left[\langle \nabla f(x_t), x_t + \beta(\bar{y} - x_t) - x_t \rangle + \frac{1}{2\alpha} \|x_t + \beta(\bar{y} - x_t) - x_t\|_2^2 \right] \\
&= \min_{\beta \in [0,1]} \left[\beta \langle \nabla f(x_t), (\bar{y} - x_t) \rangle + \frac{\beta^2}{2\alpha} \|\bar{y} - x_t\|_2^2 \right] \\
&\leq \min_{\beta \in [0,1]} \left[\beta \delta_t + \frac{\beta^2 R^2}{2\alpha} \right] = \frac{-\alpha \delta_t^2}{2R^2}
\end{aligned}
\tag{29}
$$

where the minimizer $\beta^* = -\delta_t \alpha / R^2 \leq k\alpha/R \leq 1$ as $\alpha \leq \min\{\frac{1}{k}, \frac{1}{L}\}$ (we assume $R > 1$ without loss of generality as we can take any upper bound while minimizing in (29)). Here (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 $x_t \rightarrow y$. Using (28), we get

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

Rearranging, we get our desired result

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

Since also $f(x_T) \leq f(x_{T-1}) \leq \cdots f(x_1)$, we have

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

Proof of Part (b): We refer readers to the proof in Karimi et al. [2016], which can be dated back to Polyak [1963]. $\qquad\square$

D   Example details.

D.1   Finite MDPs

We verify Condition 0 for the tabular MDP case as formulated in Example 3. Recall the policy class in that case is the set of all stationary randomized policies, $\Pi = \{\pi \in \mathbb{R}_+^{n \times k} : \sum_{i=1}^{k} \pi_{s,i} = 1 \ \forall s \in \{1, \cdots, n\}\}$. As argued in Section 6, for any $\pi, \bar{\pi} \in \Pi$ and $s \in \mathcal{S}$, the Q-function is linear in $\bar{\pi}$, as $Q_\pi(s, \bar{\pi}(s)) = \langle Q_\pi(s, e_i), \bar{\pi}(s)$. Therefore, $\mathcal{B}(\bar{\pi}|\eta, J_\pi) = \sum_{s \in \mathcal{S}} \eta(s) Q_\pi(s, \bar{\pi}(s))$ is linear and hence differentiable in $\bar{\pi}$.

Similarly, for any $s, s' \in \mathcal{S}$, the probability transition matrix $P_\pi(s'|s) \in \mathbb{R}^{n \times n}$ is linear, $P_\pi(s'|s) = \langle P(s'|s, e_i), \pi'(s) \rangle$ and hence differentiable in π'. Also note that P_π is a stochastic matrix, $\|P_\pi\|_\infty \leq 1$ and hence $(I - \gamma P_\pi)$ is invertible. Therefore, $\eta_{\pi'}(\cdot) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \rho P_{\pi'}^t = (1 - \gamma)\rho(I - \gamma P_\pi)^{-1}$ is differentiable[10] in $\bar{\pi}$ (using the inverse function theorem).

D.2   LQ control

D.2.1   Preliminaries.

We consider the LQ control problem as described in Example 2 with all the notations and assumptions introduced there. Even though this example doesn't fit our general formulation as the per period costs are not uniformly bounded, the important properties of Bellman operators that are used in our proofs all hold when restricting attention to stable linear policies and quadratic value functions. Define the set of strictly convex quadratic cost to go functions as

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

Lemma 14 (Bellman operators for LQ control). *Consider the LQ control problem formulated in Example 2. For $J, \bar{J} \in \mathcal{J}_q$ and a stable linear policy π, the following properties hold:*

1. *(Closure on the set of quadratic cost-to-go functions) $T_\pi J \in \mathcal{J}_q$ and $TJ \in \mathcal{J}_q$.*

2. *(Monotonicity) If $J \preceq \bar{J}$, then $T_\pi J \preceq T_\pi \bar{J}$ and $TJ \preceq T\bar{J}$.*

3. *(Bellman equation) $J_\pi = T_\pi J_\pi$ and $J_\pi = \lim_{k \to \infty} T_\pi^k J$. Moreover, $J = TJ$ if and only if $J = J^*$.*

We use these properties extensively for our analysis but omit the proofs as these results can be found[11] in standard textbooks [e.g. Bertsekas, 1995].

Beyond these facts of Bellman operators, in the proof of several results we will use the following standard property of the trace operator. This can be found in [Fang et al., 1994], for example. Recall that $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a symmetric matrix.

Lemma 15. *For any two symmetric and positive semi-definite symmetric matrices $A, B \in \mathbb{R}^{n \times n}$:*

$$\lambda_{\min}(A)\text{Trace}(B) \leq \text{Trace}(AB) \leq \lambda_{\max}(A)\text{Trace}(B)$$

---

[10]For finite states, it is easy to see that discounted state occupancy measure in (2) can be simplified to this form.

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

D.2.2 Smoothness properties for LQ control.

We do not verify Condition 0 for LQ control as it is well known that for any stable policy, $\ell(\theta)$ is twice continuously differentiable [Rautert and Sachs, 1997]. Studying only stable linear policies is appropriate, as we argued in Section 4 that beginning with a stable linear policy, iterates for first order methods with appropriate step-sizes are assured to stay in the sublevel set, which by Lemma 4 only contain stable policies.

Lemma 4. *Consider the LQ control problem formulated in Example 2. The set Θ_S is open and ℓ is twice continuously differentiable on Θ_S. For any $\alpha \in \mathbb{R}$, the sublevel set $C_\alpha := \left\{\theta \in \mathbb{R}^{n \times k} : \ell(\theta) \leq \alpha\right\}$ is a compact subset of Θ_S and $\sup_{\theta \in C_\alpha} \|\nabla^2 \ell(\theta)\| < \infty$.*

Proof. We first show that any sublevel set only contains stable policies, i.e. $C_\alpha \subset \Theta_S$. Recall that by our assumption on the initial distribution, $\Sigma := \mathbb{E}_\rho[s_0 s_0^\top] \succ 0$. As shown in Section 4, we can write the total cost function corresponding to a linear policy π_θ as $\ell(\theta) = (1 - \gamma)\mathbb{E}_\rho[s_0^\top K_\theta s_0]$. Therefore,

$$(1 - \gamma)^{-1}\ell(\theta) = \mathbb{E}_\rho\left[s_0^\top K_\theta s_0\right] = \mathbb{E}\left[\text{Trace}\left(K_\theta s_0 s_0^\top\right)\right] = \text{Trace}(K_\theta \Sigma) \geq \text{Trace}(K_\theta)\,\lambda_{\min}(\Sigma), \tag{30}$$

where we applied Lemma 15. We have the following bound on the trace of the cost-to-go matrix K_θ:

$$
\begin{aligned}
\text{Trace}(K_\theta) = \text{Trace}&\left(\sum_{t=0}^\infty \gamma^t \left[(A + B\theta)^t\right]^\top (\theta^\top R\theta + C)\left[(A + B\theta)^t\right]\right)\\
&\geq \lambda_{\min}(\theta^\top R\theta + C)\sum_{t=0}^\infty \gamma^t \text{Trace}\left(\left[(A + B\theta)^t\right]^\top \left[(A + B\theta)^t\right]\right)\\
&\geq \lambda_{\min}(\theta^\top R\theta + C)\sum_{t=0}^\infty \gamma^t \lambda_{\max}\left(\left[(A + B\theta)^t\right]^\top \left[(A + B\theta)^t\right]\right)\\
&= \lambda_{\min}(\theta^\top R\theta + C)\sum_{t=0}^\infty \gamma^t \|(A + B\theta)^t\|_2^2\\
&\geq \lambda_{\min}(\theta^\top R\theta + C)\sum_{t=0}^\infty \gamma^t \|A + B\theta\|_2^{2t}
\end{aligned}
$$

where the first inequality uses Lemma 15 and the equality uses a basic fact about the operator norm: $\|\tilde{A}\|_2 = \sqrt{\lambda_{\max}\left(\tilde{A}\tilde{A}^\top\right)}$ for any matrix $\tilde{A}$. Clearly, $\sum_{t=0}^\infty \gamma^t \|A + B\theta\|_2^{2t} = \infty$ for any $\theta \notin \Theta_S$. (Recall by definition $\theta \in \Theta_S$ if and only if $\|A + B\theta\|_2^2 < 1/\gamma$). As $R, C \succ 0$, this implies that $\ell(\theta) = \infty$ for any $\theta \notin \Theta_S$. Using (30), we conclude that a sublevel set cannot contain an unstable policy.

With a little bit of algebraic simplification (see Rautert and Sachs [1997]), it is easy to see that $\ell(\theta)$ is twice continuously differentiable for any $\theta \in \Theta_S$ and hence over sublevel sets (as $C_\alpha \subset \Theta_S$). We show that sublevel sets are compact by showing that they are closed and bounded. As ℓ is continuous, by definition its sublevel sets are closed. For the class of linear policies, $\pi_\theta(s) = \theta s$, we can show $\ell(\theta)$ is a coercive function, meaning $\lim_{\|\theta\|_2 \to \infty} \ell(\theta) = \infty$. To see this, consider

$$\ell(\theta) = (1 - \gamma)\mathbb{E}_\rho\left[\sum_{t=0}^\infty \gamma^t s_t^\top (\theta^\top R\theta + C)s_t\right]$$

where s_t evolves according to linear dynamics, $s_t = (A + B\theta) s_{t-1}$. Define $\Sigma_\theta := (1-\gamma)^{-1}\mathbb{E}_\rho \left[\sum_{t=0}^\infty \gamma^t s_t s_t^\top \right]$. Clearly, $\Sigma_\theta \succeq \Sigma \succ 0$. Therefore, using that $\|\theta^\top R\theta + C\|_2 = \lambda_{\max}(\theta^\top R\theta + C)$,

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

Since $\|\theta^\top R\theta + C\|_2 \geq \|\theta^\top R\theta\|_2 - \|C\|_2$, this implies

$$\|\theta^\top R\theta\|_2 \leq \frac{\ell(\theta)}{\lambda_{\min}(\Sigma_\theta)} + \|C\|_2 \leq \frac{\ell(\theta)}{(1 - \gamma)\lambda_{\min}(\Sigma)} + \|C\|_2 \tag{31}$$

where we used that $\lambda_{\min}(\Sigma_\theta) > (1 - \gamma)\lambda_{\min}(\Sigma) > 0$. Since R is strictly positive definite, it is clear by (31) that $\lim_{\|\theta\|_2 \to \infty} \ell(\theta) = \infty$. By definition, sublevel sets of a coercive function are bounded, (see [for e.g. Peressini et al., 1988]) which completes our argument.

It is easy to show that $\ell(\cdot)$ is also smooth over sublevel sets. By definition, any twice differentiable function $f : \mathcal{X} \to \mathbb{R}$ is smooth on a subset $D \subseteq \mathcal{X}$ if $\nabla^2 f(x) \preceq LI$ for some constant $L < \infty$. As $\ell(\cdot)$ is twice continuously differentiable, $\|\nabla^2\ell(\theta)\|_2$ is a continuous function. Because any sublevel set C_α of $\ell(\cdot)$ is compact, the Extreme Value Theorem implies $\|\nabla^2\ell(\theta)\|_2$ is bounded on any sublevel set, i.e. $\sup_{\theta \in C_\alpha} \|\nabla^2\ell(\theta)\|_2 < \infty$. $\qquad\square$

D.2.3 Optimality of stationary points for LQ control: Proof of Lemma 5

Lemma 5. *For the LQ control problem formulated in Example 2, any stable linear policy θ satisfies $\nabla\ell(\theta) = 0$ if and only if $J_{\pi_\theta} = J^*$.*

Proof. Fix a stable linear policy π_θ and let $\pi_{\overline{\theta}}$ be the policy iteration update to π_θ. That is, $\overline{\theta}$ satisfies $T_{\pi_{\overline{\theta}}} J_{\pi_\theta} = TJ_{\pi_\theta}$. Set $\theta^\alpha = (1 - \alpha)\theta + \alpha\overline{\theta}$ for some $\alpha \in [0, 1]$. As both π_θ and $\pi_{\overline{\theta}}$ are linear policies, this implies, $\pi_{\theta^\alpha}(s) = (1 - \alpha)\theta s + \alpha\overline{\theta}s$. For every $s \in \mathbb{R}^n$,

$$\begin{aligned}
T_{\pi_{\theta^\alpha}} J_{\pi_\theta}(s) = Q_{\pi_\theta}(s, \pi_{\theta^\alpha}(s)) &= Q_{\pi_\theta}(s, (1 - \alpha)\theta s + \alpha\overline{\theta}s) \\
&\leq (1 - \alpha)Q_{\pi_\theta}(s, \theta s) + \alpha Q_{\pi_\theta}(s, \overline{\theta}s) \\
&= (1 - \alpha)T_{\pi_\theta} J_{\pi_\theta}(s) + \alpha T_{\pi_{\overline{\theta}}} J_{\pi_\theta}(s) \\
&= (1 - \alpha)J_{\pi_\theta}(s) + \alpha TJ_{\pi_\theta}(s) \\
&= J_{\pi_\theta}(s) - \alpha\left(J_{\pi_\theta}(s) - TJ_{\pi_\theta}(s) \right) \tag{32}
\end{aligned}$$

where the first inequality uses that $a \mapsto Q_{\pi_\theta}(s, a)$ is convex, as noted in Section 4. As $J_{\pi_\theta} \succeq TJ_{\pi_\theta}$, we conclude from (32) that $J_{\pi_\theta} \succeq T_{\pi_{\theta^\alpha}} J_{\pi_\theta}$. Repeatedly applying the Bellman operator and using the monotonicity property gives,

$$J_{\pi_\theta} \succeq T_{\pi_{\theta^\alpha}} J_{\pi_\theta} \succeq T_{\pi_{\theta^\alpha}}^2 J_{\pi_\theta} \succeq \cdots \succeq \lim_{k \to \infty} T_{\pi_{\theta^\alpha}}^k J_{\pi_{\theta^\alpha}} = J_{\pi_{\theta^\alpha}}. \tag{33}$$

As, $J_{\pi_{\theta^\alpha}} \preceq J_{\pi_\theta}$, the interpolated policy π_{θ^α} is stable. Then from (32) and (33), we have

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

Multiplying each side by $(1 - \gamma)$, taking the expectation over s_0 drawn from the initial distribution ρ, and then taking $\alpha \to 0$ gives

$$\left.\frac{d}{d\alpha}\ell(\theta^\alpha)\right|_{\alpha=0} \leq (1 - \gamma)\mathbb{E}_\rho\left[TJ_{\pi_\theta}(s_0) - J_{\pi_\theta}(s_0) \right]. \tag{34}$$

We show that for a stable suboptimal policy π_θ, moving along the policy iteration update as in (34), is a feasible descent direction and therefore θ cannot be a stationary point. 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 as $J_{\pi_\theta} \succeq T J_{\pi_\theta}$. We argue that $\mathbb{E}_\rho [E(s_0)] < 0$ by using the fact that $\Sigma := \mathbb{E}_\rho \left[s_0 s_0^\top \right]$ is assumed to be positive definite.

To show this, note that as $J_{\pi_\theta} \in \mathcal{J}_q$ and $T J_{\pi_\theta} \in \mathcal{J}_q$, $E(s)$ is clearly a quadratic function. Since $E(s) \leq 0$, it can be written as $E(s) = s^\top K s$ for some symmetric $K \preceq 0$ with $K \neq 0$. Therefore, the matrix K has spectral decomposition $K = \sum_{i=1}^n \lambda_i q_i q_i^\top$, where $(q_1, \cdots, q_n)$ is an orthonormal basis of eigenvectors, each eigenvalue is non-positive, and at least one eigenvalue is strictly negative as $E(s) < 0$ for some state s. We get our result by noting,

$$\mathbb{E}_\rho [E(s_0)] = \sum_{i=1}^n \lambda_i \mathbb{E}_\rho \left[(q_i^\top s_0)^2 \right] = \sum_{i=1}^n \lambda_i q_i^\top \Sigma q_i < 0.$$

$\square$

D.3 Concentrability coefficient for LQ control

Lemma 12. *Consider the LQ control problem formulated in Example 2. Define $\Sigma_\rho := \mathbb{E}_\rho \left[s_0 s_0^\top \right]$ and let $\theta^* \in \mathbb{R}^{n \times k}$ denote the parameter of an optimal policy. Then,*

$$\|J - J^*\|_{1,\rho} \leq \frac{\kappa}{(1 - \gamma)} \|J - TJ\|_{1,\rho} \qquad \forall J \in \{J_{\pi_\theta} : \theta \in \Theta_S\}$$

when

$$\kappa = \frac{(1 - \gamma)}{1 - \gamma \|A + B\theta^*\|_2^2} \cdot \frac{\lambda_{\max} (\Sigma_\rho)}{\lambda_{\min} (\Sigma_\rho)}. \tag{20}$$

Proof. For any $\pi_\theta \in \Theta_S$, by definition $J_{\pi_\theta} \in \mathcal{J}_q$ and $T J_{\pi_\theta} \in \mathcal{J}_q$, and therefore $J_{\pi_\theta} - T J_{\pi_\theta} \in \mathcal{J}_q$. Then, using that $J_{\pi_\theta} \succeq T J_{\pi_\theta}$, we get

$$\|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\rho} = \mathbb{E}_\rho [J_{\pi_\theta}(s_0) - T J_{\pi_\theta}(s_0)] = \mathbb{E}_\rho \left[s_0^\top K s_0 \right] = \mathrm{Trace}(K \Sigma_\rho). \tag{35}$$

for some $K \in \mathbb{R}^{n \times n}, K \succ 0$. This simplifies the right hand side in the definition of κ_ρ. To simplify the left hand side, we derive the following performance difference lemma specialized to the LQ control setting.

$$(J_{\pi_\theta} - J_{\pi_{\theta^*}})(s_0) = \sum_{t=0}^\infty \gamma^t \left(J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_\theta} \right) (s_t^*) \tag{36}$$

where $s_t^* = (A + B\theta^*)^t s_0$ is the state at time t if π_{θ^*} is applied from initial state s_0. To see this, recall that by definition

$$(T_{\pi_\theta} J)(s) = (\theta s)^\top R(\theta s) + s^\top C s + \gamma J(As + B\theta s). \tag{37}$$

From this, and the Bellman equation, we have

$$J_{\pi_\theta} - J_{\pi_{\theta^*}} = T_{\pi_\theta} J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_{\theta^*}} = \left(T_{\pi_\theta} J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_\theta} \right) + \left(T_{\pi_{\theta^*}} J_{\pi_\theta} - T_{\pi_{\theta^*}} J_{\pi_{\theta^*}} \right)$$

Applying this from a particular state s_0, and using (37) gives

$$J_{\pi_\theta}(s_0) - J_{\pi_{\theta^*}}(s_0) = \left(T_{\pi_\theta} J_{\pi_\theta}(s_0) - T_{\pi_{\theta^*}} J_{\pi_\theta}(s_0)\right) + \gamma \left(J_{\pi_\theta}((A + B\theta^*)s_0) - J_{\pi_{\theta^*}}((A + B\theta^*)s_0)\right)$$
$$= \left(T_{\pi_\theta} J_{\pi_\theta}(s_0) - T_{\pi_{\theta^*}} J_{\pi_\theta}(s_0)\right) + \gamma \left(J_{\pi_\theta}(s_1^*) - J_{\pi_{\theta^*}}(s_1^*)\right).$$

The result in (36) follows by iterating over this recursion. Note that as $T J_{\pi_\theta} \preceq \inf_\pi T_\pi J_{\pi_\theta} \preceq T_{\pi_{\theta^*}} J_{\pi_\theta}$, from (36) we also get

$$(J_{\pi_\theta} - J_{\pi_{\theta^*}})(s_0) \leq \sum_{t=0}^{\infty} \gamma^t \left(J_{\pi_\theta} - T J_{\pi_\theta}\right)(s_t^*) = \sum_{t=0}^{\infty} \gamma^t (s_t^*)^\top K s_t^* = \mathrm{Trace}\left(K \sum_{t=0}^{\infty} \gamma^t s_t^* (s_t^*)^\top\right)$$

Taking expectations and using $J_{\pi_\theta} \succeq J_{\pi_{\theta^*}}$ gives

$$\|J_{\pi_\theta} - J_{\pi_{\theta^*}}\|_{1,\rho} = \mathbb{E}_\rho \left[J_{\pi_\theta}(s_0) - J_{\pi_{\theta^*}}(s_0)\right] \leq \frac{1}{(1-\gamma)} \mathrm{Trace}\left(K \Sigma_{\eta_{\theta^*}}\right), \tag{38}$$

where we define

$$\Sigma_{\eta_{\theta^*}} = (1-\gamma)\mathbb{E}_\rho \left[\sum_{t=0}^{\infty} \gamma^t s_t^* (s_t^*)^\top\right] = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \left[(A + B\theta^*)^t\right] \Sigma_\rho \left[(A + B\theta^*)^t\right]^\top.$$

Combining (35) and (38) gives

$$\|J_{\pi_\theta} - J_{\pi_{\theta^*}}\|_{1,\rho} \leq \|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\rho} \cdot \frac{\mathrm{Trace}\left(K \Sigma_{\eta_{\theta^*}}\right)}{(1-\gamma)\mathrm{Trace}\left(K \Sigma_\rho\right)} \leq \|J_{\pi_\theta} - T J_{\pi_\theta}\|_{1,\rho} \cdot \frac{\lambda_{\max}\left(\Sigma_{\eta_{\theta^*}}\right)}{(1-\gamma)\lambda_{\min}\left(\Sigma_\rho\right)}$$

where the last step uses Lemma 15. We get our desired result by noting that,

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

where the final inequality uses that

$$\|\Sigma_{\eta_{\theta^*}}\|_2 \leq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \|A + B\theta^*\|_2^{2t} \|\Sigma_\rho\|_2 = \frac{(1-\gamma)}{1-\gamma\|A + B\theta^*\|_2^2} \cdot \|\Sigma_\rho\|_2$$

and the fact that $\|\Sigma\|_2 = \lambda_{\max}(\Sigma)$ for a symmetric positive definite matrix Σ. $\qquad \square$

D.4 Optimal Stopping

We now consider the optimal stopping problem as described in Example 5, continuing with the notation and assumptions introduced there. Recall that in our formulation, for each context $x \in \mathcal{X}$, the offer distribution is assumed to have a density, $q_x(\cdot)$ supported over $\mathcal{Y}$, i.e. $q_x(y) > 0$ for all $y \in \mathcal{Y}$. We also assume the initial distribution which is supported over the set of continuation states, factorizes as $\rho(x, y) = \nu(x)q_x(y)$, where $\nu(x) > 0$ for all $x \in \mathcal{X}$.

Notation. We simplify notation to write $\eta_\theta := \eta_{\pi_\theta}$, $T_\theta := T_{\pi_\theta}$, $J_\theta := J_{\pi_\theta}$ and $P_\theta := P_{\pi_\theta}$ for any $\theta \in \Theta$. We let η'_θ denote the marginal distribution over $\mathcal{X} \cup \{\tau\}$ under η_θ, and note that

$$\eta_\theta\Big(\{x\}, (y_1, y_2)\Big) = \eta'_\theta(x) \int_{y_1}^{y_2} q_x(y) dy. \tag{39}$$

We find it convenient to directly work with η'_θ and $q_x(y)$. We denote $P'_\theta \in \mathbb{R}^{(|\mathcal{X}|+1) \times (|\mathcal{X}|+1)}$ to be the transition matrix over $\mathcal{X} \cup \{\tau\}$ under π_θ, defined as

$$P'_\theta(x'|x) = p(x'|x) \int_{\mathcal{Y}} \mathbb{1}(y < \theta_x) q_x(y) dy, \quad P'_\theta(\tau|x) = \int_{\mathcal{Y}} \mathbb{1}(y \geq \theta_x) q_x(y) dy, \quad P'_\theta(\tau|\tau) = 1, \tag{40}$$

for all $x', x \in \mathcal{X}$. We now verify all the conditions needed for our main results.

Condition 1: Closure under policy improvement. 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 $s = (x, y) \in \mathcal{S}_{\mathrm{C}}$:

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

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

Condition 0 and twice continuous differentiability. For any $\theta, \overline{\theta} \in \Theta$, we show that $\mathcal{B}(\overline{\theta}|\eta_\theta, J_\theta)$ is a twice continuously differentiable function of $\overline{\theta}$. The fundamental theorem of calculus implies,

$$\frac{\partial}{\partial \overline{\theta}_x} \mathcal{B}(\overline{\theta}|\eta_\theta, J_\theta) = \frac{\partial}{\partial \overline{\theta}_x} \sum_{\tilde{x}} \eta'_\pi(\tilde{x}) \int_{\mathcal{Y}} Q_{\pi_\theta}\big((\tilde{x}, y), \pi_{\overline{\theta}}(\tilde{x}, y)\big) q_{\tilde{x}}(y) dy \tag{41}$$

$$= \eta'_\pi(x) \frac{\partial}{\partial \overline{\theta}_x} \int_{\mathcal{Y}} Q_{\pi_\theta}\big((x, y), \pi_{\overline{\theta}}(x, y)\big) q_x(y) dy$$

$$= \eta'_\pi(x) \frac{\partial}{\partial \overline{\theta}_x} \int_{\mathcal{Y}} \big[\mathbb{1}(y \geq \overline{\theta}_x) y + \mathbb{1}(y < \overline{\theta}_x) c_{\pi_\theta}(x) \big] q_x(y) dy$$

$$= \big(c_{\pi_\theta}(x) - \theta_x \big) \eta'_\pi(x) q_x(\theta_x). \tag{42}$$

Using the assumption that $q_x(\cdot)$ is itself continuously differentiable, (41) implies that $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta}|\eta_\theta, J_\theta)$ is twice continuously differentiable in $\overline{\theta}$. Now consider $\mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta)$. As the continuation value from the terminal state τ is zero, we can write

$$\mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta) = \sum_{x \in \mathcal{X}} \eta'_{\overline{\theta}}(x) \int_{\mathcal{Y}} J_\theta(x, y) q_x(y) dy$$

We show $\eta'_{\overline{\theta}} \in \mathbb{R}^{|\mathcal{X}|+1}$ is twice continuously differentiable, which establishes the same for $\overline{\theta} \mapsto \mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta)$. Note that for finite states, we can simplify (2) to write

$$\eta'_{\overline{\theta}} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \rho(P'_{\overline{\theta}})^t = (1-\gamma)\rho(I - \gamma P'_{\overline{\theta}})^{-1}$$

48

Per our definition in (40), its easy to see that $P'_{\overline{\theta}}$ is twice continuously differentiable in $\overline{\theta}$, as $q_x(\cdot)$ is assumed to be continuously differentiable. Also, $(I - \gamma P'_{\overline{\theta}})$ is invertible as $P'_{\overline{\theta}}$ is a stochastic matrix. Therefore, by inverse function theorem, $\eta'_{\overline{\theta}}(\cdot)$ is twice continuously differentiable.

Condition 2.A: No suboptimal stationary points for the weighted PI objective. We now show that $\theta \mapsto \mathcal{B}(\theta|\eta_\pi, J_\pi)$ has no suboptimal stationary points for each $\pi \in \Pi_\Theta$. As we formulate the optimal stopping example as a maximization problem, any stationary point θ satisfies,

$$\max_{\theta'_x \in \mathcal{Y}} \frac{\partial}{\partial \theta_x} \mathcal{B}(\theta|\eta, J_\pi) \cdot (\theta'_x - \theta_x) \le 0 \iff \max_{\theta'_x \in \mathcal{Y}} (c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x) \le 0, \quad (43)$$

where we have used the derivative calculation in (41). We argue that $c_\pi(x) < y_{\max}$ for any $x \in \mathcal{X}$, as $q_x(\cdot)$ is supported over $\mathcal{Y}$ and therefore $c_\pi(x) = \gamma \int_{\mathcal{Y}} J_\pi(x, y) q_x(y) dy < y_{\max}$. Using this with (43), one can show that if θ is a stationary point, then

$$\theta_x = \begin{cases} c_\pi(x) & \text{if } c_\pi(x) \in (y_{\min}, y_{\max}), \\ y_{\min} & \text{if } y_{\min} \ge c_\pi(x) \end{cases} \quad (44)$$

Now, consider the policy iteration update,

$$\theta = \arg\max_{\theta' \in \Theta} \mathcal{B}(\theta'|\eta_\pi, J_\pi) = \arg\max_{\theta' \in \Theta} \eta'_\pi(x) \int_{\mathcal{Y}} \left[\mathbb{1}(y \ge \theta'_x) y + \mathbb{1}(y < \theta'_x) c_\pi(x) \right] q_x(y) dy$$

As argued above, for any state $(x, y) \in \mathcal{S}_C$, the improved policy should accept only if $y \ge c_\pi(x)$. Hence, $\theta_x = \max\{c_\pi(x), y_{\min}\}$. Comparing with (44), it is easy to see that any stationary point of $\theta \mapsto \mathcal{B}(\theta|\eta_\pi, J_\pi)$ solves the weighted policy iteration objective.

Condition 2.B: Gradient dominance. We first recall the claim.

Lemma 9 (Gradient dominance for optimal stopping). *Consider the optimal stopping problem formulated in Example 5. For any $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta|\eta_\pi, J_\pi)$ is $(\beta, 0)$–gradient-dominated where $\beta = \max_{x \in \mathcal{X}, y \in \mathcal{Y}} q_x(y) / \min_{x \in \mathcal{X}, y \in \mathcal{Y}} q_x(y)$.*

Proof. Fix any $\pi \in \Pi_\Theta$. As we formulate the optimal stopping as a reward maximization problem, following our notion of gradient dominance in Definition 2, we want to show that

$$\max_{\theta' \in \Theta} \langle \nabla_\theta \mathcal{B}(\theta|\eta_\pi, J_\pi), \theta' - \theta \rangle \ge \frac{1}{\beta} \left(\mathcal{B}(\theta^+|\eta_\pi, J_\pi) - \mathcal{B}(\theta|\eta_\pi, J_\pi) \right) \quad (45)$$

where θ^+ is the parameter of a policy iteration update to π, i.e. $\theta^+ = \arg\max_{\theta \in \Theta} \mathcal{B}(\theta|\eta_\pi, J_\pi)$. We first lower bound the left hand side of (45) as,

$$\max_{\theta' \in \Theta} \langle \nabla_\theta \mathcal{B}(\theta|\eta_\pi, J_\pi), \theta' - \theta \rangle = \sum_{x \in \mathcal{X}} \max_{\theta'_x \in \mathcal{Y}} \frac{\partial \mathcal{B}(\theta|\eta_\pi, J_\pi)}{\partial \theta_x} \cdot (\theta'_x - \theta_x)$$

$$= \sum_{x \in \mathcal{X}} \max_{\theta'_x \in \mathcal{Y}} (c_\pi(x) - \theta_x) \cdot \eta'_\pi(x) q_x(\theta_x) \cdot (\theta'_x - \theta_x)$$

$$\ge \left(\min_{x' \in \mathcal{X}, y' \in \mathcal{Y}} q_{x'}(y') \right) \sum_{x \in \mathcal{X}} \eta'_\pi(x) \left\{ \max_{\theta'_x \in \mathcal{Y}} \left[(c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x) \right] \right\}.$$

$$(46)$$

49

where second equality uses the derivative calculation in (41). We now upper bound the right hand side of (45). Using that $T_{\pi_{\theta+}} J_\pi = T J_\pi$, we have

$$\mathcal{B}(\theta^+|\eta_\pi, J_\pi) - \mathcal{B}(\theta|\eta_\pi, J_\pi) = \sum_{x \in \mathcal{X}} \eta'_\pi(x) \int_\mathcal{Y} \left(T J_\pi(x,y) - T_{\pi_\theta} J_\pi(x,y) \right) q_x(y) dy$$

$$\leq \left(\max_{x' \in \mathcal{X}, y' \in \mathcal{Y}} q_{x'}(y') \right) \sum_{x \in \mathcal{X}} \eta'_\pi(x) \int_\mathcal{Y} \left(T J_\pi(x,y) - T_{\pi_\theta} J_\pi(x,y) \right) dy. \tag{47}$$

We can simplify to write,

$$T_{\pi_\theta} J_\pi(x,y) = y \cdot \mathbb{1}(y \geq \theta_x) + c_{\pi_\theta}(x) \cdot \mathbb{1}(y < \theta_x) \tag{48}$$

$$T J_\pi(x,y) = y \cdot \mathbb{1}(y \geq \theta_x^+) + c_{\pi_\theta}(x) \cdot \mathbb{1}(y < \theta_x^+) \tag{49}$$

We argued above that $c_\pi(x) < y_{\max}$ and $\theta_x^+ = \max\{c_\pi(x), y_{\min}\}$. We consider two cases.

Case (1): Assume that $c_\pi(x) \in (y_{\min}, y_{\max})$ in which case $\theta_x^+ = c_\pi(x)$. Using equations (48) and (49), it is easy to check that

$$\int_\mathcal{Y} \left(T J_\pi(x,y) - T_{\pi_\theta} J_\pi(x,y) \right) dy \leq (c_\pi(x) - \theta_x)^2. \tag{50}$$

As $c_\pi(x) \in (y_{\min}, y_{\max})$, observing the lower bound in (46), we find that

$$\max_{\theta'_x \in \mathcal{Y}} (c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x) \geq (c_\pi(x) - \theta_x)^2 \tag{51}$$

We get our desired result by using (50) and (51) to compare Equations (47) and (46) respectively.

Case (2): Now assume that $c_\pi(x) \notin (y_{\min}, y_{\max})$ in which case $\theta_x^+ = y_{\min}$. Again, using (48) and (49), it is easy to check that

$$\int_\mathcal{Y} \left(T J_\pi - T_{\pi_\theta} J_\pi(x,y) \right) dy = \int_{\theta_x^+}^{\theta_x} (y - c_\pi(x)) \, dy \leq \int_{\theta_x^+}^{\theta_x} (\theta_x - c_\pi(x)) \, dy$$

$$= (\theta_x - c_\pi(x)) \left(\theta_x - \theta_x^+ \right). \tag{52}$$

Similarly, we can lower bound the right hand side of (46) as

$$\max_{\theta'_x \in \mathcal{Y}} (c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x) \geq (\theta_x - c_\pi(x)) \cdot (\theta_x - \theta_x^+). \tag{53}$$

We get out desired result by using (52) and (53) to compare Equations (47) and (46) respectively. $\square$

Smoothness for optimal stopping. We first recall the claim.

Lemma 10. *For the optimal stopping problem in Example 5, $\max_{\theta \in \Theta} \|\nabla^2 \ell(\theta)\| < \infty$.*

Proof. Using the policy gradient theorem as shown in Lemma 6 and the derivative calculations in (41), we have

$$\frac{\partial}{\partial \theta_x} \ell(\theta) = \frac{\partial}{\partial \bar{\theta}_x} \mathcal{B}(\bar{\theta} | \eta_\theta, J_\theta) \Big|_{\bar{\theta}=\theta} = \left(c_{\pi_\theta}(x) - \theta_x \right) \eta'_\theta(x) q_x(\theta_x).$$

We argued above when verifying Condition 0 that $\eta'_\theta(x)$ is continuously differentiable. By assumption, $q_x(\theta_x)$ is differentiable in θ_x. Therefore, $\nabla \ell(\theta)$ has a continuous derivative if $c_{\pi_\theta}(x)$ is continuously differentiable in θ. To show this, recall that by definition,

$$c_{\pi_\theta}(x) = \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J_\theta(x', y') q_x(y') dy = \gamma \sum_{x' \in \mathcal{X}} p(x'|x) J'_\theta(x') \qquad \forall x \in \mathcal{X} \qquad (54)$$

where we define $J'_\theta(x) := \int_{\mathcal{Y}} J_\theta(x', y') q_x(y') dy'$ to be the expected cost-to-go function from context x. Similarly, denote $g'_\theta(x) := \int_{\mathcal{Y}} \mathbb{1}(y \geq \theta_x) y q_x(y) dy$ to be the expected reward earned from context x. Clearly, g'_θ is continuously differentiable. Also, by definition,

$$J'_\theta = \left(I - \gamma P'_\theta \right)^{-1} g'_\theta.$$

where the transition matrix P'_θ was defined in (40). While verifying Condition 0 above, we argued that P'_θ is continuously differentiable and $(I - \gamma P'_\theta)$ is invertible. Using the inverse function theorem, we get that J'_θ is continuously differentiable as well. Hence using (54) we find that c_{π_θ} is continuously differentiable in θ. Since Θ is compact, the Extreme Value theorem implies $\max_{\theta \in \Theta} \|\nabla^2 \ell(\theta)\|_2$ exists and is finite.

$\square$

Concentrability coefficient for optimal stopping. We first recall the claim.

Lemma 11. *For the optimal stopping problem in Example 5, consider a policy π_C that never stops, i.e. $\pi_C(s) = 1$ for all $s \in \mathcal{S}_C$. Let μ be a stationary distribution of the induced Markov process, meaning $\mu(\mathcal{M}) = \int P(\mathcal{M}|s', 1) \mu(ds')$ for any $\mathcal{M} \subset \mathcal{S}$. Then, choosing $\rho = \mu$ implies $\kappa_\rho \leq 1$.*

Proof. We show that the Bellman operator T is a contraction with modulus γ in $\|\cdot\|_{1,\mu}$. The proof then follows immediately using part (c) of Theorem 4.

For a policy that never stops, the stationary distribution over continuation states, $(x, y) \in \mathcal{S}_C$ factorizes as $\mu(x, y) = \mu'(x) q_x(y)$ where μ' is the marginal stationary distribution over context states $\mathcal{X}$ such that $\mu'(x') = \sum_{x \in \mathcal{X}} \mu'(x) p(x'|x)$. Then, for any bounded cost-to-go functions, J, J'

$$\|TJ - TJ'\|_{1,\mu} = \sum_{x \in \mathcal{X}} \mu'(x) \left| \int_{\mathcal{Y}} \left(TJ(x, y) - TJ'(x, y) \right) q_x(y) dy \right|$$

By definition,

$$TJ(x, y) = \max\{y, \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J(x', y') q_{x'}(y') dy'\}$$

Note that for any scalars (x_1, x_2, y), we have $|\max\{y, x_1\} - \max\{y, x_2\}| \leq |x_1 - x_2|$. Therefore,

$$\left| TJ(x,y) - TJ'(x,y) \right| \leq \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} \left| J(x',y') - J'(x',y') \right| q_{x'}(y') dy' \qquad (55)$$

As the right hand side in (55) is independent of y, integrating (55) with respect to $q_x(\cdot)$ gives

$$\int_{\mathcal{Y}} \left| TJ(x,y) - TJ'(x,y) \right| q_x(y) dy \leq \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} \left| J(x',y') - J'(x',y') \right| q_{x'}(y') dy'$$

Therefore,

$$
\begin{aligned}
\|TJ - TJ'\|_{1,\mu} &= \sum_{x \in \mathcal{X}} \mu'(x) \int_{\mathcal{Y}} \left| TJ(x,y) - TJ'(x,y) \right| q_x(y) dy \\
&\leq \gamma \sum_{x \in \mathcal{X}} \mu'(x) \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} \left| J(x',y') - J'(x',y') \right| q_{x'}(y') dy' \\
&\overset{(a)}{=} \gamma \sum_{x' \in \mathcal{X}} \mu'(x') \int_{\mathcal{Y}} \left| J(x',y') - J'(x',y') \right| q_{x'}(y') dy' \\
&= \gamma \|J - J'\|_{1,\mu}
\end{aligned}
$$

where (a) follows as μ' is the stationary distribution over $\mathcal{X}$. For $\rho = \mu$, we have $C, c = 1$ in part (c) of Theorem 4, implying that $\kappa_\rho \leq 1$. $\qquad \square$