Appendices

Appendix A: Notation

Table EC.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 in state s.
Π	$\triangleq$	Set of all stationary policies.
$\mathcal{J}$	$\triangleq$	Set of bounded measurable 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 $\mathcal{M} \subset \mathcal{S}$
$g_\pi \in \mathcal{J}$	$\triangleq$	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. See (2).
$\ell(\pi) = (1-\gamma)\int J_\pi d\rho$	$\triangleq$	Expected discounted cost under initial state distribution ρ, 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. See (13).
$\|J\|_\infty$	$\triangleq$	Max-norm $\sup_s \|J(s)\|$.
$\|J\|_{1,\eta}$	$\triangleq$	Weighted 1-norm $\int \|J(s)\| \, d\eta$.
$\|A\|_p$ for $A \in \mathbb{R}^{n\times m}$	$\triangleq$	Matrix operator norm $\max\{\|Ax\|_p : \|x\|_p = 1\}$
$\lambda_{\min}(A)$ for $A = A^T \in \mathbb{R}^{n\times n}$	$\triangleq$	Minimum eigenvalue of A
$\lambda_{\max}(A)$ for $A = A^T \in \mathbb{R}^{n\times n}$	$\triangleq$	Maximum eigenvalue of A
$f \succeq g$, $f \succ g$ for functions f, g	$\triangleq$	Elementwise inequality, $f(x) \geq g(x)$ or $f(x) > g(x)\ \forall x$
$A \succeq B$, $A \succ B$ for $A, B \in \mathbb{R}^{n\times n}$	$\triangleq$	Indicates $A - B$ is positive (semi) definite.

Appendix B: Discussion on concurrent work of Agarwal et al. (2020).

Here we provide some comparisons to the approach of Agarwal et al. (2020), offering perspective on the strengths and shortcomings of our approach and highlighting some connections between our technical results and theirs.

The work of Agarwal et al. (2020) is mostly focused on analyzing specific algorithms, with special attention given to natural-gradient actor-critic algorithms. Explicit bounds on the convergence rates of different algorithms are also given. In comparison, we take a broad algorithm-independent view, focusing primarily on the landscape of the loss function $\ell(\cdot)$ to understand when local policy search is a sensible approach for reaching a (near) optimal policy. Our choice of focus yields several results and insights that are not covered by their work:

- In Example 1, we provide a simple illustration of how the multi-period nature of the decision problem can lead to bad local minima. With policy closure, we show how challenges raised in that example largely disappear, and the problem is reduced to studying the single period objective function in (11). This approach of reasoning about the landscape of long-horizon objective by analyzing structure in single period problems offers new insight and greatly simplifies our analysis.
- In addition to treating some examples covered by Agarwal et al. (2020), like tabular and linear MDPs, our analysis seamlessly covers problems with infinite action spaces, structured cost functions, and deterministic policies – for example linear quadratic control and optimal stopping.
- Our approach extends to finite horizon problems with non-stationary policy classes as shown in Theorem 3. We instantiate this theory for an important practical problem of finite horizon inventory control.
- In stating gradient dominance conditions for convergence rates, our approach to concentration coefficients implies tighter bounds for some examples as compared to the distribution mismatch coefficient in Agarwal et al. (2020).

Our focus on an algorithm-agnostic study of the landscape of $\ell(\cdot)$ comes at a cost. This is felt most clearly when studying the softmax parameterization which poses unique optimization challenges even in the seemingly simple single-state single-period problem with a finite number of actions. There is no optimal solution; instead optimal performance is attained only as certain components of the parameter vector tend to infinity. Although our theory applies easily to problems with entropy regularization (see e.g. Example 5), a more specialized analysis is required for the case without regularization. Agarwal et al. (2020) do this with a detailed study of specific algorithms and precisely characterize the convergence behavior for both the cases of with and without regularization. In addition, their work also shows how natural policy gradients can be particularly useful in alleviating conditioning issues arising with softmax policies and result in faster convergence.

The next subsection tries to clarify some connections between our theory and that appearing in Agarwal et al. (2020).

Connecting closure conditions to conditions on value function approximation. Compared to our work, a novel theory appears in Agarwal et al. (2020) on function approximation. Specifically, they analyze a natural gradient actor critic method working with a class of softmax linear policies and compatible function approximation (Sutton et al. 1999, Konda and Tsitsiklis 2000), i.e. deriving features for approximating the Q-function from score function of a stochastic policy. Although we do not treat this case in detail, we briefly remark on an interesting connection which shows that in one important setting (linearly parameterized Q-functions and the corresponding softmax linear policies), accuracy of value function approximation error implies approximate closure as shown in Condition 5.

This observation follows from a natural duality between a parametric class of policies and a parametric class of value functions. In Example 6, for instance, we leveraged the fact that if a parametric class of value functions $\{Q^\theta : \theta \in \Theta\}$ can approximate each Q_π, then the class of greedy policies induced by this class of value functions is closed under policy improvement. A similar observation holds for approximate closure as shown below in Example EC.1.

EXAMPLE EC.1. As in Example 3, 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. Suppose $g(s,a)$ and $P(s'|s,a)$ are linear in a, as in (14). Fix a feature mapping $\phi : \mathcal{S} \times \{1, \cdots, k\} \mapsto \mathbb{R}^d$. Consider the class of policies, $\Pi_\Theta = \{\pi_\theta : \theta \in \mathbb{R}^d\}$ where $\pi_\theta(s) = (\pi_\theta(s,1), \cdots \pi_\theta(s,k)) \in \Delta^{k-1}$ has components $\pi_\theta(s,i) \propto \exp\{\theta^\top \phi(s,i)\}$. $\square$

LEMMA EC.1 **(Accurate function approximation implies approximate closure)**. *In the setting of Example EC.1, suppose that for each* $\pi \in \Pi_\Theta$,

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{(s,i)\sim \eta_\pi \otimes \text{Unif}} \left[\left(Q_\pi(s,e_i) - w^\top \phi(s,i) \right)^2 \right] < \epsilon^2/k^2. \tag{EC.1}$$

Then Π_Θ *satisfies Condition 5.*

Proof of Lemma EC.1 Given π, let $\hat{w}$ satisfy $\mathbb{E}_{(s,i)\sim \eta_\pi \otimes \text{Unif}} \left[\left(Q_\pi(s,e_i) - w^\top \phi(s,i) \right)^2 \right] = \epsilon_0^2/k^2$ where $\epsilon_0 < \epsilon/2$. Set

$$\hat{Q}_\pi(s,a) = \sum_{i=1}^k a_i \hat{Q}_\pi(s,e_i) = \sum_{i=1}^k a_i (\hat{w}^\top \phi(s,i)).$$

to be the resulting approximate Q-function. For a scalar $c > 0$, consider the policy $\pi_{c\hat{w}}$ which assigns probability $\pi_{c\hat{w}}(s,i) \propto \exp\{c \cdot \hat{Q}_\pi(s,e_i)\}$ to base action i in state s. We show that $\pi_{c\hat{w}}$ approximately optimizes the policy iteration objective when c is large. Note,

$$\begin{aligned}
\mathcal{B}(\pi_{c\hat{w}} \,|\, \eta_\pi, J_\pi) &= \int \left(Q_\pi(s, \pi_{c\hat{w}(s)}) \right) \eta_\pi(ds) \\
&\leq \int \left(\hat{Q}_\pi(s, \pi_{c\hat{w}}(s)) \right) \eta_\pi(ds) + \int \max_{a \in \mathcal{A}} \left| \hat{Q}_\pi(s,a) - Q_\pi(s,a) \right| \eta_\pi(ds) \\
&\leq \int \left(\hat{Q}_\pi(s, \pi_{c\hat{w}}(s)) \right) \eta_\pi(ds) + \epsilon_0
\end{aligned}$$

where the final inequality follows from using the identity $\mathbb{E}[X^2] \geq (\mathbb{E}[|X|])^2$ for any random variable X and observing that

$$\begin{aligned}
\int \max_{a \in \mathcal{A}} \left| \hat{Q}_\pi(s,a) - Q_\pi(s,a) \right| \eta_\pi(ds) &\leq \int \sum_{i=1}^k \left| \hat{Q}_\pi(s,e_i) - Q_\pi(s,e_i) \right| \eta_\pi(ds) \\
&= k \int k^{-1} \sum_{i=1}^k \left| \hat{Q}_\pi(s,e_i) - Q_\pi(s,e_i) \right| \eta_\pi(ds) \\
&\leq k \sqrt{\int k^{-1} \sum_{i=1}^k \left(\hat{Q}_\pi(s,e_i) - Q_\pi(s,e_i) \right)^2 \eta_\pi(ds)} \leq \epsilon_0.
\end{aligned}$$

This implies,

$$\begin{aligned}
\lim_{c\to\infty} \mathcal{B}(\pi_{c\hat{w}} \,|\, \eta_\pi, J_\pi) &\leq \lim_{c\to\infty} \int \left(\hat{Q}_\pi(s, \pi_{c\hat{w}}(s)) \right) \eta_\pi(ds) + \epsilon_0 \\
&= \int \lim_{c\to\infty} \left(\hat{Q}_\pi(s, \pi_{c\hat{w}}(s)) \right) \eta_\pi(ds) + \epsilon_0 \\
&= \int \min_{a \in \mathcal{A}} \left(\hat{Q}_\pi(s,a) \right) \eta_\pi(ds) + \epsilon_0 \\
&\leq \int \min_{a \in \mathcal{A}} \left(Q_\pi(s,a) \right) \eta_\pi(ds) + \int \max_{a \in \mathcal{A}} \left| \hat{Q}_\pi(s,a) - Q_\pi(s,a) \right| \eta_\pi(ds) + \epsilon_0 \\
&\leq \min_{\pi^+ \in \Pi} \mathcal{B}(\pi^+ \,|\, \eta_\pi, J_\pi) + 2\epsilon_0.
\end{aligned}$$

The first and second equality use the monotone convergence theorem together with the fact that, for each s, $\hat{Q}_\pi(s, \pi_{c\hat{w}}(s)) \uparrow \max_{a\in\mathcal{A}} \hat{Q}_\pi(s,a)$. Since $\epsilon_0 < \epsilon/2$ we can choose a $\hat{c} \in \mathbb{R}$ such that $\mathcal{B}(\pi_{\hat{c}\hat{w}} \mid \eta_\pi, J_\pi) \leq \min_{\pi^+\in\Pi} \mathcal{B}(\pi^+ \mid \eta_\pi, J_\pi) + \epsilon$. □

The theory in Agarwal et al. (2020) is a tour-de-force, containing many ideas which are not reflected in the result above. Let us remark on two of them:

1. Agarwal et al. (2020) treat generalizations of Example EC.1 where policies are still stochastic but are not log-linear. If $\pi_\theta(s,i) \propto \exp\{f_\theta(s,i)\}$, then the theory of compatible function approximation (Sutton et al. 1999) implies modeling $Q_\pi(s,e_i) \approx w^\top \nabla_\theta f_\theta(s,i)$ for some choice of weights $w \in \mathbb{R}^d$. When this approximation is accurate, natural gradient updates to the parameter vector will tilt the policies' action probabilities towards selecting actions with higher Q-values. Because it is local in nature, such a property appears to no longer imply closure of the policy class under policy improvement steps, which involves global properties of the policy class. The choice to focus on natural gradient methods helps Agarwal et al. (2020) to analyze this example, since there is an explicit formula for the parameter updates which has a very close connection to compatible approximation of the Q-function.
2. Agarwal et al. (2020) state their results in terms of a notion called *transfer error*. This measures implicitly depends on both issues of approximation, as in (EC.1), and issues of distribution shift, as in our term κ_ρ. It is a very nice insight that this term is what is really needed in the analysis.

Appendix C: 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 (Thrun 1992, Kakade and Langford 2002) or the "river swim" problem (Strehl and Littman 2008, Osband et al. 2013), 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.

Example EC.2. Consider the MDP shown in Figure EC.1. 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 vector13. 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

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

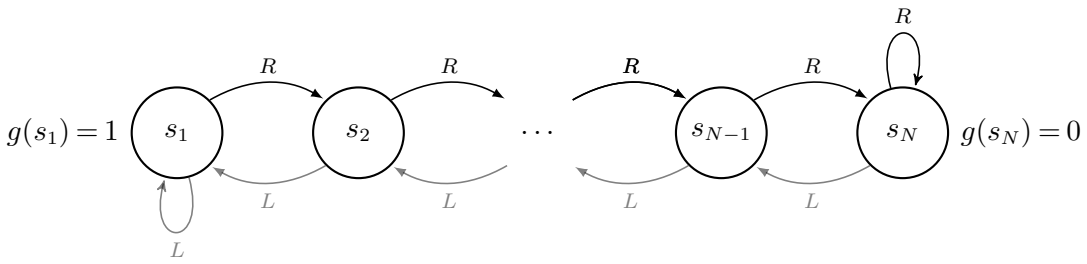

Figure EC.1 A simple chain MDP example to illustrate how policy gradient methods face suboptimal local minima in the absence of an exploratory initial distribution.

We argue that a suboptimal policy π that always moves left, i.e. $\pi(s_i) = 0 \; \forall \, i \in \{1, \ldots, N\}$, 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 explicit bound confirming that the policy gradient can be exponentially small in N is shown in Agarwal et al. (2020). □

Appendix D: Omitted proofs.

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

D.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.

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.

$$\|J_\pi - J^*\|_\infty \leq \frac{1}{1-\gamma} \|J_\pi - TJ_\pi\|_\infty \tag{EC.2}$$

where J^* is the optimal cost-to-go function. A simple argument (Bertsekas 1995) shows (EC.2).

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

Balance equation.

The best way to understand equation (EC.2) below is by analogy to an equivalent undiscounted problem: η_π is the steady state distribution in a problem in which the next state is drawn from the restart distribution $\rho(\cdot)$ with probability $1-\gamma$ and otherwise is drawn according to the transition kernel under the policy π. This result is only used explicitly in one place, but may provide helpful intuition throughout.

LEMMA EC.2. *The discounted state occupancy measure satisfies the balance equation,*

$$\eta_\pi(\mathcal{M}) = \int_{\mathcal{S}} \left[(1-\gamma)\rho(\mathcal{M}) + \gamma P(\mathcal{M}|s,\pi(s))\right] \eta_\pi(ds) \qquad \forall \mathcal{M} \subset \mathcal{S}.$$

Proof of Lemma EC.2. This can be directly verified by using the tower property of conditional expectation as follows:

$$\begin{aligned}
\eta_\pi(\mathcal{M}) &= (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\rho^\pi \left[\mathbb{1}\left(s_t \in \mathcal{M}\right)\right] \\
&= (1-\gamma)\rho(\mathcal{M}) + (1-\gamma)\mathbb{E}_\rho^\pi \left[\sum_{t=1}^{\infty} \gamma^t \mathbb{P}_\rho^\pi \left[s_t \in \mathcal{M} \mid s_{t-1}\right]\right] \\
&= (1-\gamma)\rho(\mathcal{M}) + (1-\gamma)\mathbb{E}_\rho^\pi \left[\sum_{t=1}^{\infty} \gamma^t P(\mathcal{M}|s_{t-1},\pi(s_{t-1}))\right] \\
&= (1-\gamma)\rho(\mathcal{M}) + \gamma(1-\gamma)\mathbb{E}_\rho^\pi \left[\sum_{t=0}^{\infty} \gamma^t P(\mathcal{M}|s_t,\pi(s_t))\right] \\
&= (1-\gamma)\rho(\mathcal{M}) + \gamma \int_{\mathcal{S}} P(\mathcal{M}|s,\pi(s))\eta_\pi(ds). \quad \square
\end{aligned}$$

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. $\rho\left(\{s \in \mathcal{S} : J_\pi(s) = J^*(s)\}\right) = 1$.*

Proof of Lemma 1. Recall $\ell(\pi) = (1-\gamma)\int J_\pi(s)\rho(ds)$. 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

$$\ell(\pi) - \ell(\pi^*) = (1-\gamma)\int \left(J_\pi(s) - J^*(s)\right)\rho(ds) \geq 0. \tag{EC.3}$$

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 (EC.3) if and only if $J_\pi - J^* = 0$ ρ-almost-surely. $\square$

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

$$\begin{aligned}
J_\pi(s_0) - J(s_0) &= T_\pi J(s_0) - J(s_0) + T_\pi J_\pi(s_0) - T_\pi J(s_0) \\
&= T_\pi J(s_0) - J(s_0) + \gamma P_\pi \left(J_\pi(s_0) - J(s_0)\right)
\end{aligned}$$

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

$$\mathbb{E}_\nu^\pi \left[J_\pi(s_0) - J(s_0)\right] = \mathbb{E}_\nu^\pi \left[\sum_{t=0}^{\infty} \gamma^t \left[T_\pi J(s_t) - J(s_t)\right]\right], \tag{EC.4}$$

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 (EC.4) and recalling that $\ell(\pi) = (1-\gamma)\mathbb{E}_\rho[J_\pi(s_0)]$ gives

$$\ell(\pi) - \ell(\bar{\pi}) = (1-\gamma)\mathbb{E}_\rho^\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(T_\pi J_{\bar{\pi}}(s_t) - J_{\bar{\pi}}(s_t)\right)\right] = \int \left[T_\pi J_{\bar{\pi}} - J_{\bar{\pi}}\right] d\eta_\pi. \tag{EC.5}$$

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 *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 of Lemma 6. 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 (EC.5), 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,

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

D.2. No suboptimal stationary points: Proof of Theorem 1.

For reader's convenience, we first restate Theorem 1.

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. To establish 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 EC.3. *Suppose Condition 2.A is satisfied. If θ is a stationary point of $\ell : \Theta \to \mathbb{R}$, then*

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

Proof of Lemma EC.3. 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,

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

On average Bellman equation:
Next, in Lemma EC.4, 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 EC.4 **(On average Bellman equation)**. *For any* $\pi \in \Pi$,

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

Proof of Lemma EC.4. 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).

Take π^+ to be the policy iteration update at π, i.e. $T_{\pi^+}J_\pi = TJ_\pi$. 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,

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

where (a) follows by using the performance difference lemma in (EC.5) 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. Since $J_\pi - TJ_\pi \succeq 0$, we must have $\rho(S_0^c) = 0$. But 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,

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

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 (EC.5). □

Completing proof of Theorem 1:
We now complete the 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,

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

where the first equality follows from Lemma EC.3 (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 $J_\pi \succeq TJ_\pi$, we have

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

The "on average Bellman equation" in Lemma EC.4 lets us conclude that $\ell(\theta) = \ell(\pi^*)$. □

D.3. Convergence rates: Proof of Theorem 2.

For reader's convenience, we first restate Theorem 2.

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

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. We first derive a consequence of the closure condition:

$$\begin{aligned}
\ell(\pi_\theta)-\min_{\pi\in\Pi}\ell(\pi) &= (1-\gamma)\int\left[J_{\pi_\theta}-J^*\right]d\rho \overset{(a)}{=} (1-\gamma)\|J_{\pi_\theta}-J^*\|_{1,\rho}\\
&\overset{(b)}{\le} \kappa_\rho\|J_{\pi_\theta}-TJ_{\pi_\theta}\|_{1,\rho}\\
&\overset{(c)}{\le} \frac{\kappa_\rho}{(1-\gamma)}\|J_{\pi_\theta}-TJ_{\pi_\theta}\|_{1,\eta_{\pi_\theta}}\\
&= \frac{\kappa_\rho}{(1-\gamma)}\int\left[J_{\pi_\theta}-TJ_{\pi_\theta}\right]d\eta_{\pi_\theta}\\
&\overset{(d)}{=} \frac{\kappa_\rho}{(1-\gamma)}\left(\int J_{\pi_\theta}d\eta_{\pi_\theta}-\min_{\pi\in\Pi_\Theta}\int\left[T_\pi J_{\pi_\theta}\right]d\eta_{\pi_\theta}\right)\\
&= \frac{\kappa_\rho}{(1-\gamma)}\left(\mathcal{B}(\theta\,|\,\eta_{\pi_\theta},J_{\pi_\theta})-\min_{\theta'\in\Theta}\mathcal{B}(\theta'\,|\,\eta_{\pi_\theta},J_{\pi_\theta})\right).
\end{aligned}$$

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 the definition in (2)) and (d) uses closure property of the policy class.

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

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

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

D.4. 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 $\ell(\pi_\theta)=\ell(\pi^*)$.*

Proof of Theorem 3. Before giving a detailed proof, we outline a brief sketch to give intuition. The proof proceeds by backward induction. We first show that at any stationary point θ of $\ell(\cdot)$, $J_{\pi_\theta}(s)=J^*(s)$ holds ρ-almost surely for all $s\in\mathcal{S}_H$. We then argue that the same statement holds for $s\in\mathcal{S}_h$ for all $h<H$ as well. Our result then follows by invoking Lemma 1.

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 notation14., 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. $\pi_\theta(s)=\pi_{\theta_h}(s)$ 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,

$$\theta \in \arg\min_{\overline{\theta}\in\Theta} \mathcal{B}(\overline{\theta} \mid \eta, J_\pi) \iff \theta_h \in \arg\min_{\overline{\theta}_h\in\Theta_h} \mathcal{B}_h(\overline{\theta}_h \mid \eta, J_\pi) \ \ \forall h. \tag{EC.6}$$

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 θ is a stationary point of $\min_{\overline{\theta}\in\Theta} \ell(\overline{\theta})$ if and only if θ_h is a stationary point of $\min_{\overline{\theta}_h\in\Theta_h} \mathcal{B}_h(\overline{\theta}_h \mid \eta_{\pi_\theta}, J_{\pi_\theta})$ $\forall h$. That is,

$$\langle \nabla\ell(\theta), \theta' - \theta\rangle \geq 0 \quad \forall \theta' \in \Theta \iff \left[\frac{\partial}{\partial \overline{\theta}_h} \mathcal{B}_h(\overline{\theta}_h \mid \eta_{\pi_\theta}, J_{\pi_\theta})\Big|_{\overline{\theta}_h=\theta_h}\right](\theta'_h - \theta_h) \geq 0 \quad \forall \theta'_h \in \Theta_h, \forall h \tag{EC.7}$$

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 $\pi \in \Pi_\Theta$. In the reverse direction, $\eta_\pi \succeq (1-\gamma)\rho$ by definition, which implies $\rho \ll \eta_\pi$ for each $\pi \in \Pi_\Theta$. Therefore, we can throughout claim that certain events hold "almost surely" without reference to whether the base measure is ρ or η_π.

Let θ be a stationary point of $\ell(\cdot)$. We proceed by backward induction, showing $J_{\pi_\theta}(s) = J^*(s)$ almost surely for $s \in \mathcal{S}_h$. As a base case, 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,

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

where we use throughout that $Q_{\pi_\theta} \succeq Q^*$ or $J_{\pi_\theta} \succeq J^*$. Inequality (a) is justified by the following balance equation that holds for the discounted state occupancy measure: for any $h \in \{1, \cdots, H\}$ and and $\mathcal{M} \subset \mathcal{S}_{h+1}$,

$$\eta_\pi(\mathcal{M}) = (1-\gamma)\rho(\mathcal{M}) + \gamma \int_{\mathcal{S}_h} P(\mathcal{M}|s, \pi(s))\eta_\pi(ds).$$

This follows from the general balance equation of Lemma EC.2, which states $\eta_\pi(\mathcal{M}) = (1-\gamma)\rho(\mathcal{M}) + \int P(\mathcal{M}|s,\pi(s))\eta_\pi(ds)$ holds for all $\mathcal{M} \subset \mathcal{S}$, and the fact that $P(\mathcal{S}_{h+1}|s,\pi(s)) = 0$ for $s \notin \mathcal{S}_h$ due to Condition 3. Inequality (b) uses the induction hypothesis that $J_{\pi_\theta}(s) = J^*(s)$ for $s \in \mathcal{S}_{h+1}$ almost surely.

Having shown $\mathcal{B}_h(\overline{\theta}_h|\eta_{\pi_\theta}, J_{\pi_\theta}) = \mathcal{B}_h(\overline{\theta}_h|\eta_{\pi_\theta}, J^*)$ for all $\overline{\theta}_h \in \Theta_h$, we now complete the induction step. As θ is a stationary point of $\ell(\cdot)$, the characterization of stationary points in (EC.7) implies that θ_h is a stationary point of the optimization problem $\min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h(\overline{\theta}_h|\eta_{\pi_\theta}, J_{\pi_\theta})$, or equivalently, of $\min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h(\overline{\theta}_h|\eta_{\pi_\theta}, J^*)$. But in Condition 4 we assumed that $\overline{\theta} \mapsto \mathcal{B}(\overline{\theta} \mid \eta_{\pi_\theta}, J^*)$ has no suboptimal stationary points. By the separability structure highlighted in (EC.6), this implies that $\overline{\theta}_h \mapsto \mathcal{B}_h(\overline{\theta}_h \mid \eta_{\pi_\theta}, J^*)$ also has no suboptimal stationary points, so we have shown $\theta_h \in \arg\min_{\overline{\theta}_h \in \Theta_h} \mathcal{B}_h\left(\overline{\theta}_h \mid \eta_{\pi_\theta}, J^*\right)$. Putting it all together, we get,

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

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 $J_{\pi_\theta} \succeq J^*$, we conclude that $J_{\pi_\theta}(s) = J^*(s)$ for $s \in \mathcal{S}_h$ almost surely, completing the induction step. The statement in Theorem 3 follows by invoking Lemma 1. □

D.5. Concentrability coefficients.

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

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

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

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

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

Proof of Theorem 4. 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 (EC.5), we get

$$(1-\gamma)\int (J_\pi - J^*)\, d\rho = (1-\gamma)\|J_\pi - J^*\|_{1,\rho} = \ell(\pi) - \ell(\pi^*) = \int (J_\pi - T_{\pi^*} J_\pi)\, d\eta_{\pi^*}$$

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

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 (EC.2), if T is contraction with modulus γ in $\|\cdot\|$, then $\|J - J^*\| \leq \frac{1}{(1-\gamma)}\|J - TJ^*\|$. Our result follows by noting,

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

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

D.6. Closure under approximate policy improvement: Proof of Theorem 5.

For reader's convenience, we first restate Theorem 5.

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

$$\begin{aligned}
\epsilon \geq \min_{\pi^+ \in \Pi_\Theta} \mathcal{B}(\pi^+ | \eta_{\pi_\theta}, J_{\pi_\theta}) - \min_{\pi' \in \Pi} \mathcal{B}(\pi' | \eta_{\pi_\theta}, J_{\pi_\theta}) &= \min_{\pi^+ \in \Pi_\Theta} \left(\int \left[T_{\pi^+} J_{\pi_\theta}\right] d\eta_{\pi_\theta} \right) - \int \left[TJ_{\pi_\theta}\right] d\eta_{\pi_\theta} \\
&= \int \left[J_{\pi_\theta} - TJ_{\pi_\theta}\right] d\eta_{\pi_\theta} \\
&= \|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\eta_{\pi_\theta}},
\end{aligned}$$

where the second equality follows from Lemma EC.3 and the final equality uses that $J_{\pi_\theta} \succeq TJ_{\pi_\theta}$ for any $\pi_\theta \in \Pi_\Theta$. Then, we have

$$\begin{aligned}
\ell(\pi_\theta) - \min_\pi \ell(\pi) = (1-\gamma) \int \left[J_{\pi_\theta} - J^*\right] d\rho &= (1-\gamma)\|J_{\pi_\theta} - J^*\|_{1,\rho} \\
&\leq \kappa_\rho \|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\rho} \\
&\leq \frac{\kappa_\rho}{(1-\gamma)} \|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\eta_{\pi_\theta}} \\
&\leq \frac{\kappa_\rho \cdot \epsilon}{(1-\gamma)}.
\end{aligned}$$

The first inequality follows from the definition of κ_ρ, the second uses that $\eta_{\pi_\theta} \succeq (1-\gamma)\rho$ and the final inequality follows from Condition 5. □

Appendix E: Example details.

E.1. Finite state and action MDPs with natural parameterization.

We restate and prove the convergence rate result for tabular MDPs given in Lemma 7. We use a result in (Agarwal et al. 2020) which shows that for the policy gradient objective with natural parameterization, $\nabla\ell$ is Lipschitz continuous with constant $L = \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^2}$.

LEMMA 7 *Consider the finite state action MDP with natural parameterization as formulated in Example 3. Assume $\gamma \geq 1/3$ and per-period costs are normalized with $\|g\|_\infty \leq 1$. For projected gradient descent, $\pi^{(t+1)} = \text{Proj}_\Pi(\pi^t - \alpha\nabla\ell(\pi^t)))$ with $\alpha = \frac{(1-\gamma)^2}{2\gamma|\mathcal{A}|}$,*

$$\ell(\pi^T) - \ell^* \leq \sqrt{\frac{8\gamma|\mathcal{S}||\mathcal{A}|\kappa_\rho^2}{(1-\gamma)^4}\frac{(\ell(\pi^0)-\ell^*)}{T}},$$

where $\ell^* = \inf_{\pi\in\Pi}\ell(\pi)$.

Proof of Lemma 7. We assumed that the per-step costs are normalized, i.e. $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|g(s,a)| \leq 1$. As the weighted PI objective $\pi' \mapsto \mathcal{B}(\pi'|\eta_\pi, J_\pi)$ in (15) is (1,0)-gradient dominated, Theorem 2 implies that $\ell(\cdot)$ is $(\frac{\kappa_\rho}{(1-\gamma)}, 0)$-gradient dominated. For finite state-actions MDPs, Theorem 4 implies $\kappa_\rho \leq \sup_{s\in\mathcal{S}}\frac{\eta_{\pi^*}(s)}{\rho(s)}$. Application of Lemma 3 also requires smoothness conditions. For this, we appeal to Lemma D3 in (Agarwal et al. 2020) which shows that $\nabla\ell$ is Lipschitz continuous with constant $L = \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^2}$. Note that there is a difference of a factor of $(1-\gamma)$ in the denominator as compared to the result in Lemma D3 of Agarwal et al. (2020), due to our definition of the policy gradient objective. While we consider the discounted average cost, $\ell(\pi) = (1-\gamma)\sum_{s\in\mathcal{S}} J_\pi(s)\rho(s)$, Agarwal et al. (2020) consider the undiscounted objective, $\ell(\pi) = \sum_{s\in\mathcal{S}} J_\pi(s)\rho(s)$. Finally, it is easy to compute the constants R and k in part (1) of Lemma 3 for tabular MDPs with natural parameterization as:

$$R = \sup_{\pi,\pi'\in\Pi}\|\pi - \pi'\|_2 \leq \sqrt{2|\mathcal{S}|}, \tag{EC.8}$$

$$k = \sup_{\pi\in\Pi}\|\nabla\ell(\pi)\|_2 \leq \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)} \tag{EC.9}$$

where (EC.8) follows from the fact that $\sum_{j=1}^k(\pi(s,e_j) - \pi'(s,e_j))^2 \leq 2$ for any $s\in\mathcal{S}$ and $\pi\in\Pi$. For (EC.9), we use that $\frac{\partial\ell(\pi)}{\partial\pi(s,e_j)} = \eta_\pi(s)Q_\pi(s,e_j)$, where $|Q_\pi(s,e_j)| \leq \frac{1}{(1-\gamma)}$ since we assume per-step costs are normalized and $0 \leq \eta_\pi(s) \leq 1$ by our definition of the discounted state occupancy measure. As $\frac{1}{k} > \frac{1}{L}$ for $\gamma \geq 1/3$, the claim follows by letting $\alpha = \frac{1}{L}$ and $c = \frac{\kappa_\rho}{(1-\gamma)}$ in the statement of Lemma 3 part (1). □

E.2. Regularized finite state and action MDPs with natural parameterization.

LEMMA 8. *Let $\ell_\lambda(\pi)$ denote the average cost function for the problem described in Example 4 with a given regularization parameter $\lambda \geq 0$. Then, if $\pi_\lambda^* \in \arg\min_{\pi\in\Pi}\ell_\lambda(\pi)$,*

$$\ell_0(\pi_\lambda^*) \leq \min_{\pi\in\Pi}\ell_\lambda(\pi) + \lambda\left(1 + \log\left(1 + \frac{c}{\lambda}\right)\right) \qquad \textit{where} \quad c = 2(\max_{s,i}|g_{s,i}|)/(1-\gamma).$$

Proof of Lemma 8. To make the dependence on λ explicit, we consider the MDP $M_\lambda = (\mathcal{S}, \mathcal{A}, g_\lambda, P, \gamma, \rho)$ where

$$g_\lambda(s,a) = g_s^\top a + \lambda\cdot D_{\text{KL}}(U||a) = \sum_{i=1}^k g_0(s,e_i)a_i + \lambda\cdot D_{\text{KL}}(U||a)$$

For the regularized problem, we write the appropriate terms as $T_\lambda^\pi, T_\lambda, \ell_\lambda(\pi), Q_\lambda^\pi$, Q_λ^*, J_λ^*, J_λ^π. For ease of notation, we write $D(\cdot||\cdot)$ to denote $D_{\mathrm{KL}}(\cdot||\cdot)$. We prove this result in a sequence of steps.

Step 1: Cost-function decomposition: Using the relationship between the per-step cost function and the average cost function, we have that for any $\pi \in \Pi$,

$$\ell_\lambda(\pi) = \sum_{s\in\mathcal{S}} \eta_\pi(s) g_\lambda(s,\pi(s)) = \sum_{s\in\mathcal{S}} \eta_\pi(s)\left(g_0(s,\pi(s)) + \lambda \cdot D(U||\pi(s))\right) = \ell_0(\pi) + \lambda \sum_{s\in\mathcal{S}} \eta_\pi(s) D(U||\pi(s)). \tag{EC.10}$$

Step 2: We construct a policy π_λ with bounded sub-optimality gap for the unregularized objective: Essentially, we show that there exists a π_λ, with

$$\ell_0(\pi_\lambda) \le \min_{\pi\in\Pi} \ell_0(\pi) + \lambda. \tag{EC.11}$$

To do this, define π_λ as the solution to the following regularized optimization problem,

$$\pi_\lambda(s) = \arg\min_{a^\top e = 1} \sum_{i=1}^k Q_0^*(s,e_i) a_i + \lambda \cdot D(U||a). \tag{EC.12}$$

where Q_0^* denotes the optimal state-action cost-to-go function of the unregularized MDP. We can interpret $D(U||a) = -\frac{1}{k}\sum_{i=1}^k \log(a_i) - \log(k)$ as a log-barrier penalty function for the probability simplex with effective regularization parameter λ/k. The log-barrier regularization plays a key role in the theory of interior point methods (Boyd and Vandenberghe 2004, Chaper 11). In particular, we use a result (Boyd and Vandenberghe 2004, page 566) which shows that optimal solutions to log-barrier regularized problems are near-optimal for un-regularized ones. For our construction in (EC.12), this implies

$$\sum_{i=1}^k Q_0^*(s,e_i)\pi_\lambda(i|s) \le \min_{a\in\Delta^{k-1}} \sum_{i=1}^k Q_0^*(s,e_i) a_i + \lambda$$

where $\pi_\lambda(s) = (\pi_\lambda(1|s), \cdots, \pi_\lambda(k|s))$. In terms of Bellman operators, this means

$$T_0^{\pi_\lambda} J_0^* \le T_0 J_0^* + \lambda e = J_0^* + \lambda e \tag{EC.13}$$

where e is a vector of 1's. Repeatedly applying $T_0^{\pi_\lambda}$ to each side of this expression and using the monotonicity and constant shift properties of the Bellman operator (Bertsekas 1995) shows $J_0^{\pi_\lambda} \preceq J_0^* + \frac{\lambda}{1-\gamma} e$. Combining this with definition of the average cost, $\ell_0(\pi) = (1-\gamma)\sum_s \rho(s) J_0^\pi(s)$, shows our result in (EC.11).

Step 3: Bound $D(U||\pi_\lambda(s))$: Combining the result in (EC.11) with the cost decomposition in (EC.10) implies,

$$\ell_\lambda(\pi_\lambda) \le \min_{\pi\in\Pi} \ell_0(\pi) + \lambda + \lambda \sum_{s\in\mathcal{S}} \eta_\pi(s) D(U||\pi_\lambda(s)) \tag{EC.14}$$

We now show an upper bound on $D(U||\pi_\lambda(s))$ which shows our result. For simplicity, focus on a single fixed state s and take

$$a^* = \pi_\lambda(s) = \arg\min_{a^\top e = 1} \sum_{i=1}^k Q_0^*(s,e_i) a_i - \frac{\lambda}{k}\sum_{i=1}^k \log(a_i).$$

Without loss of generality, assume actions are ordered such that $Q_0^*(s,e_1) \ge Q_0^*(s,e_j)$ for all j. In this case we must have that $a_1^* \le a_j^*$ for each j, i.e. $a_1^* \le 1/k$. From the first order optimality conditions, we have

$$Q_0^*(s,e_1) - \frac{(\lambda/k)}{a_1^*} = Q_0^*(s,e_j) - \frac{(\lambda/k)}{a_j^*} \quad \forall j \ge 2.$$

This implies,

$$Q_0^*(s,e_1) - \frac{(\lambda/k)}{a_1^*} = \frac{1}{1-a_1^*}\left[\sum_{j=2}^{k}\left(Q_0^*(s,e_1) - \frac{(\lambda/k)}{a_1^*}\right)a_j^*\right] = \frac{1}{1-a_1^*}\left[\sum_{j=2}^{k}\left(Q_0^*(s,e_j) - \frac{(\lambda/k)}{a_j^*}\right)a_j^*\right]$$
$$= \frac{1}{1-a_1^*}\sum_{j=2}^{k} Q_0^*(s,e_j)a_j^* - \frac{\lambda\cdot(k-1)/k}{1-a_1^*}.$$

Rearranging terms gives

$$\frac{\lambda/k}{a_1^*} - \frac{\lambda\cdot(k-1)/k}{1-a_1^*} = Q_0^*(s,e_1) - \frac{1}{1-a_1^*}\sum_{j=2}^{k} Q_0^*(s,e_j)a_j^* := \Delta.$$

Here, Δ is roughly interpreted as the excess cost of action 1 over the weighted average of other actions. We use the uniform bound $\Delta \le c$. Multiplying each side by k/λ and using that $a_1^* \le 1/k \implies 1-a_1^* \ge (k-1)/k$,

$$\frac{1}{a_1^*} = \frac{k}{\lambda}\cdot\Delta + \frac{k-1}{1-a_1^*} \le \frac{k}{\lambda}\cdot\Delta + k \le k\left(1+\frac{c}{\lambda}\right).$$

Since $D(U||a^*) = \frac{1}{k}\sum_{j=1}^{k}\log\left(\frac{1}{a_j^*}\right) - \log(k) \le \log(1/a_1^*) - \log(k) \le \log\left(1+\frac{c}{\lambda}\right)$, we have that

$$\lambda\cdot D(U||\pi_\lambda(s)) \le \lambda\log\left(1+\frac{c}{\lambda}\right).$$

Since this holds for every s, combining it with (EC.14) gives,

$$\min_{\pi\in\Pi}\ell_\lambda(\pi) \le \ell_\lambda(\pi_\lambda) \le \min_{\pi\in\Pi}\ell_0(\pi) + \lambda + \lambda\log\left(1+\frac{c}{\lambda}\right).$$

Our final result follows by noting that for any policy π, $\ell_0(\pi) \le \ell_\lambda(\pi)$. In particular, if $\pi_\lambda^* \in \arg\min_{\pi\in\Pi}\ell_\lambda(\pi)$, then $\ell_0(\pi_\lambda^*) \le \min_{\pi\in\Pi}\ell_\lambda(\pi)$. □

E.3. Regularized finite state and action MDPs with nonlinear parameterization.

We prove Lemma 9 as discussed in Example 5. Given a feasible descent direction D, which by definition means $\pi_\theta + \alpha D \in \Pi$ is a feasible policy for sufficiently small α, our goal is to verify that there exists N solving the linear system $\left[\frac{\partial\pi_\theta}{\partial\theta}\right]N = D$. For simplicity, imagine there is only a single state, so that a policy is described by the vector $(\pi_\theta(1),\cdots,\pi_\theta(k))$. If there were multiple states, the same argument could be repeated for each block of parameters corresponding to each distinct state.

Since we have effectively fixed a choice of $\theta_1 = 0$ in Example 5, $\left[\frac{\partial\pi_\theta}{\partial\theta}\right]N = D$ is a system of k equations with $k-1$ variables, denoted $N = (N_2,\cdots,N_k)$. We first temporarily ignore the first linear equality and show we can solve the remaining $k-1$ equations: $\frac{\partial\pi_\theta(i)}{\partial\theta_2}N_2 + \cdots + \frac{\partial\pi_\theta(i)}{\partial\theta_k}N_k = D_i$ for each $i \ge 2$. To see this, consider the $(k-1)\times(k-1)$ sub-matrix of the Jacobian that comes from dropping the first row:

$$\left[\frac{\partial\pi_\theta(i)}{\partial\theta_j}\right]_{i,j\ge 2} = \begin{bmatrix} \pi_\theta(2)(1-\pi_\theta(2)) & -\pi_\theta(2)\pi_\theta(3) & \cdots & -\pi_\theta(2)\pi_\theta(k) \\ -\pi_\theta(2)\pi_\theta(3) & \pi_\theta(3)(1-\pi_\theta(3)) & \cdots & -\pi_\theta(3)\pi_\theta(k) \\ \vdots & \vdots & \vdots & \vdots \\ -\pi_\theta(2)\pi_\theta(k) & -\pi_\theta(3)\pi_\theta(k) & \cdots & \pi_\theta(k)(1-\pi_\theta(k)) \end{bmatrix}.$$

This matrix is diagonally dominant and therefore non-singular, implying a solution to these $k-1$ linear equations exist.

We now show that the first linear equation is redundant and follows from the other $k-1$. In particular, we have

$$\sum_{j=2}^{k} \frac{\partial \pi_\theta(1)}{\partial \theta_j} N_j = \sum_{j=2}^{k} \left(-\sum_{i=2}^{k} \frac{\partial \pi_\theta(i)}{\partial \theta_j} N_j \right) = \sum_{i=2}^{k} \left(-\sum_{j=2}^{k} \frac{\partial \pi_\theta(i)}{\partial \theta_j} N_j \right) = -\sum_{i=2}^{k} D_i = D_1,$$

where the first equality uses that $\pi_\theta(1) + \cdots + \pi_\theta(k) = 1$ and the final equality uses that $D_1 + \cdots + D_k = 0$ for any feasible descent direction.

E.4. LQ control.

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 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 EC.5 **(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 $\pi \in \{\pi_\theta : \theta \in \Theta_S\}$, the following 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 found15. in standard textbooks (e.g. Bertsekas 1995). In addition, we use the following standard property of the trace operator repeatedly in our analysis. This can be found in (Fang et al. 1994), for example. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a symmetric matrix respectively.

LEMMA EC.6. *For any two symmetric and positive semi-definite symmetric matrices $A, B \in \mathbb{R}^{n\times n}$,*

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

Finally, we use several times the following observation. Recall that the second moment matrix of the initial state, denoted $\Sigma_\rho = \mathbb{E}_\rho[s_0 s_0^\top]$ is assumed to be finite and positive definite.

LEMMA EC.7. *For a square matrix M with Frobenius norm $\|M\|_F = \sqrt{\mathrm{Trace}(M^\top M)}$,*

$$\lambda_{\min}(\Sigma_\rho)\|M\|_F^2 \le \mathbb{E}_\rho\left[\|Ms_0\|_2^2\right] \le \lambda_{\max}(\Sigma_\rho)\|M\|_F^2.$$

Proof of Lemma EC.7. Using Lemma EC.6 gives

$$\begin{aligned}\mathbb{E}_\rho\left[\|Ms_0\|_2^2\right] = \mathbb{E}_\rho\left[s_0^\top M^\top M s_0\right] = \mathbb{E}\left[\mathrm{Trace}\left(s_0^\top M^\top M s_0\right)\right] &= \mathrm{Trace}\left(M^\top M \Sigma_\rho\right) \\ &\ge \lambda_{\min}(\Sigma_\rho)\,\mathrm{Trace}(M^\top M) \\ &= \lambda_{\min}(\Sigma_\rho)\|M\|_F^2.\end{aligned}$$

An identical argument yields a corresponding upper bound. □

Stability in LQ control.

The following lemma is a straightforward adaption of classical understanding of stability in linear dynamical systems to our setting, which considers the cumulative discounted cost incurred from a random initial state. To ensure reproducibility, a proof of the following result is given in the supplementary technical report (Bhandari and Russo 2021).

LEMMA EC.8. *In the LQ control problem formulated in Example 2, $\ell(\theta) < \infty$ if and only if $\theta \in \Theta_S$.*

Smoothness properties for LQ control.

Recall the cost cost function in LQ control is

$$J_{\pi_\theta}(s_0) = \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t s_t^\top (\theta^\top R\theta + C) s_t \quad \text{where} \quad s_t = [A+B\theta]^t s_0.$$

Since s_t is differentiable in θ, differentiability properties for finite T follow almost immediately. For stable policies, $(\sqrt{\gamma})^t s_t \to 0$ at a geometric rate, so later terms in the sum have a negligible contribution to the total cost. It is then natural that smoothness properties hold for stable policies, which is shown carefully by Rautert and Sachs (1997) (though derivative calculations seem to date back even to Kalman et al. (1960)). The next lemma states this finding in a form that is sufficient to apply results on the convergence of first-order algorithms like Lemma 2. The idea is that, beginning with a stable linear policy θ_0, iterates produced by first order methods with appropriate step-sizes are assured to stay in the sublevel set $C_{\ell(\theta_0)}$, which only contains 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 $C_\alpha := \left\{\theta \in \mathbb{R}^{n\times k} : \ell(\theta) \le \alpha\right\}$ is a compact subset of Θ_S and, if it is nonempty, $\sup_{\theta\in C_\alpha} \|\nabla^2 \ell(\theta)\|_2 < \infty$.*

Proof of Lemma 4. That any sublevel set only contains stable policies, i.e. $C_\alpha \subseteq \Theta_{\mathrm{S}}$, is an immediate consequence of Lemma EC.8. With a little bit of algebraic simplification (see Rautert and Sachs (1997) in continuous time and Bu et al. (2019) in discrete time), one can show that $\ell(\theta)$ is twice continuously differentiable for any $\theta \in \Theta_{\mathrm{S}}$ and hence over sublevel sets (as $C_\alpha \subseteq \Theta_{\mathrm{S}}$).

We show that sublevel sets are compact by showing that they are closed and bounded. As $\ell(\cdot)$ is continuous over Θ_{S}, by definition its sublevel sets are closed. We show $\ell(\theta)$ is a coercive function, meaning $\lim_{\|\theta\|\to\infty} \ell(\theta) = \infty$. (By the equivalence of norms in finite dimensional spaces, the definition does not depend on the choice of matrix norm.) By definition, sublevel sets of a coercive function are bounded, (see for example (Peressini et al. 1988)) so this completes our argument. To show $\ell(\cdot)$ is coercive, we lower bound it using Lemma EC.7 as

$$\begin{aligned}(1-\gamma)^{-1}\ell(\theta) = \mathbb{E}_\rho\left[\sum_{t=0}^{\infty} \gamma^t s_t^\top (\theta^\top R\theta + C) s_t\right] &\ge \mathbb{E}_\rho\left[(\theta s_0)^\top R(\theta s_0)\right] \ge \lambda_{\min}(R)\mathbb{E}_\rho\left[\|\theta s_0\|_2^2\right] \\ &\ge \lambda_{\min}(R)\lambda_{\min}(\Sigma_\rho)\|\theta\|_F^2,\end{aligned}$$

which clearly tends to infinity as $\|\theta\|_F^2 \to \infty$. Recall that $\Sigma_\rho = \mathbb{E}_\rho[s_0 s_0^\top]$. To prove the final claim, observe that 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. $\max_{\theta\in C_\alpha} \|\nabla^2\ell(\theta)\|_2 < \infty$. □

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 of Lemma 5. If θ defines an optimal policy, then the first order necessary conditions imply $\nabla\ell(\theta) = 0$. Now consider a sub-optimal stable linear policy π_θ and let $\pi_{\overline{\theta}}$ be the policy iteration update to π_θ. That is, $\overline{\theta}$ satisfies $T_{\pi_{\overline{\theta}}} J_{\pi_\theta} = T J_{\pi_\theta}$. 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 $s \in \mathbb{R}^n$,

$$\begin{aligned}T_{\pi_{\theta^\alpha}} J_{\pi_\theta}(s) = Q_{\pi_\theta}(s, \pi_{\theta^\alpha}(s)) &= Q_{\pi_\theta}(s, (1-\alpha)\theta s + \alpha\overline{\theta} s) \\ &\le (1-\alpha) Q_{\pi_\theta}(s, \theta s) + \alpha Q_{\pi_\theta}(s, \overline{\theta} s) \\ &= (1-\alpha) T_{\pi_\theta} J_{\pi_\theta}(s) + \alpha T_{\pi_{\overline{\theta}}} J_{\pi_\theta}(s) \\ &= (1-\alpha) J_{\pi_\theta}(s) + \alpha T J_{\pi_\theta}(s) \\ &= J_{\pi_\theta}(s) - \alpha\left(J_{\pi_\theta}(s) - T J_{\pi_\theta}(s)\right) \qquad \text{(EC.15)}\end{aligned}$$

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 (EC.15) that $J_{\pi_\theta} \succeq T_{\pi_{\theta^\alpha}} J_{\pi_\theta}$. Repeatedly applying the Bellman operator and using the monotonicity property gives,

$$J_{\pi_\theta} \succeq T_{\pi_{\theta^\alpha}} J_{\pi_\theta} \succeq T^2_{\pi_{\theta^\alpha}} J_{\pi_\theta} \succeq \cdots \succeq \lim_{k\to\infty} T^k_{\pi_{\theta^\alpha}} J_{\pi_{\theta^\alpha}} = J_{\pi_{\theta^\alpha}}. \tag{EC.16}$$

As, $J_{\pi_{\theta^\alpha}} \preceq J_{\pi_\theta}$, the interpolated policy π_{θ^α} is stable. Then from (EC.15) and (EC.16), we have

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

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

$$\frac{d}{d\alpha}\ell(\theta^\alpha)\Big|_{\alpha=0} \leq (1-\gamma)\,\mathbb{E}_\rho\left[TJ_{\pi_\theta}(s_0) - J_{\pi_\theta}(s_0)\right] = -(1-\gamma)\mathbb{E}_\rho\left[E(s_0)\right] \tag{EC.17}$$

where we have denoted the error in Bellman's equation by $E(s) \triangleq J_{\pi_\theta}(s) - TJ_{\pi_\theta}(s)$.

We know $E(s) \geq 0$ for all s because $J_{\pi_\theta} \succeq TJ_{\pi_\theta}$ and the inequality is strict at some s because the policy π_θ is sub-optimal by assumption and therefore $J_{\pi_\theta} \neq TJ_{\pi_\theta}$. We argue that $\mathbb{E}_\rho[E(s_0)] > 0$, showing that the right hand side of (EC.17) is negative and hence θ cannot be a stationary point. To show this, we first observe that, since $J_{\pi_\theta} \in \mathcal{J}_q$ and $TJ_{\pi_\theta} \in \mathcal{J}_q$ are both quadratic functions (See Lemma EC.5), $E(s)$ is a quadratic function. We can then write $E(s) = s^\top K s$ for some symmetric matrix K satisfying $K \succeq 0$ and $K \neq 0$. Applying Lemma EC.7 gives,

$$\mathbb{E}_\rho\left[E(s_0)\right] = \mathbb{E}_\rho\left[s_0^\top K s_0\right] = \mathbb{E}_\rho\left[\|K^{1/2}s_0\|_2^2\right] \geq \lambda_{\min}(\Sigma_\rho)\|K^{1/2}\|_F^2 = \lambda_{\min}(\Sigma_\rho)\cdot \text{Trace}(K) > 0. \quad \square$$

Concentrability coefficient for LQ control.

We first recall the lemma statement from Section 7.

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

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

where

$$\kappa = \frac{\lambda_{\max}\left(\Sigma_{\eta_{\pi_{\theta^*}}}\right)}{\lambda_{\min}\left(\Sigma_\rho\right)}.$$

The proof leverages the performance difference lemma, described in Subsection D.1 and equation (EC.5) in particular. Here, to make precise the result used in the proof, we state in Lemma EC.9 below, an analogue of that formula in linear quadratic control, which restricts to stable policies to rule out divergent sums. The proof is essentially identical to that leading to (EC.5) and is therefore omitted. Observe that, since the dynamics are deterministic with states evolving according to $s_{t+1} = (A + B\overline{\theta})s_t$, all randomness is due to the initial state s_0 drawn from ρ.

LEMMA EC.9 **(Performance difference lemma for LQ control)**. *Consider the LQ control problem formulated in Example 2. For any $\theta, \overline{\theta} \in \Theta_S$,*

$$\ell(\overline{\theta}) - \ell(\theta) = (1-\gamma)\mathbb{E}_\rho^{\pi_{\overline{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t\left(T_{\pi_{\overline{\theta}}}J_{\pi_\theta}(s_t) - J_{\pi_\theta}(s_t)\right)\right] = \int\left[T_{\pi_{\overline{\theta}}}J_{\pi_\theta} - J_{\pi_\theta}\right]d\eta_{\pi_{\overline{\theta}}}.$$

Proof of Lemma 12. For any $\pi_\theta \in \Theta_S$, by Lemma EC.5, $J_{\pi_\theta} \in \mathcal{J}_q$ and $TJ_{\pi_\theta} \in \mathcal{J}_q$. Therefore $J_{\pi_\theta} - TJ_{\pi_\theta} \in \mathcal{J}_q$, which means there exists symmetric $K \in \mathbb{R}^{n\times n}$ such that $J_{\pi_\theta}(s) - TJ_{\pi_\theta}(s) = s^\top K s$ for all s. Since $J_{\pi_\theta} \succeq TJ_{\pi_\theta}$, we have that $K \succeq 0$. We get

$$\|J_{\pi_\theta} - TJ_{\pi_\theta}\|_{1,\rho} = \mathbb{E}_\rho\left[J_{\pi_\theta}(s_0) - TJ_{\pi_\theta}(s_0)\right] = \mathbb{E}_\rho\left[s_0^\top K s_0\right] = \mathrm{Trace}(K\Sigma_\rho). \tag{EC.18}$$

This simplifies the right hand side in the definition of κ_ρ. To simplify the left hand side, we use the performance difference lemma above (with a choice $\overline{\theta} = \theta^*$),

$$\begin{aligned}(1-\gamma)\|J_{\pi_\theta} - J^*\|_{1,\rho} = -\left(\ell(\theta^*) - \ell(\theta)\right) = \int \left[J_{\pi_\theta} - T_{\pi_{\overline{\theta}}} J_{\pi_\theta}\right] d\eta_{\pi_{\theta^*}} &\le \int \left[J_{\pi_\theta} - TJ_{\pi_\theta}\right] d\eta_{\pi_{\theta^*}} \\ &= \mathbb{E}_{\eta_{\theta^*}}\left[s_0^\top K s_0\right] \\ &= \mathrm{Trace}\left(K\Sigma_{\eta_{\pi_{\theta^*}}}\right),\end{aligned}$$

where the inequality used that $TJ \preceq T_{\pi_{\theta^*}} J$ holds for all cost-to-go functions J. Combining this with (EC.18) and applying Lemma EC.6 gives,

$$\frac{\|J_{\pi_\theta} - J^*\|_{1,\rho}}{\|J_{\pi_\theta} - TJ\|_{1,\rho}} \le \frac{1}{1-\gamma} \cdot \frac{\mathrm{Trace}(K\Sigma_{\eta_{\pi_{\theta^*}}})}{\mathrm{Trace}(K\Sigma_\rho)} \le \frac{1}{1-\gamma} \cdot \frac{\lambda_{\max}\left(\Sigma_{\eta_{\pi_{\theta^*}}}\right)}{\lambda_{\min}(\Sigma_\rho)}. \quad \square$$

E.5. Optimal Stopping.

We now consider the optimal stopping problem as described in Example 7, 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, $q_x(\cdot)$, with continuous derivative and support $\{y \in \mathbb{R} : q_x(y) > 0\} = (y_{\min}, y_{\max})$ where $y_{\min} > 0$. We set $\mathcal{Y} = [y_{\min}, y_{\max}]$. We also assume the initial distribution places zero probability on trivial instances that begin in the terminal state (i.e $\rho(\tau) = 0$) and factorizes over continuation states as $\rho(x, dy) = \nu(x) q_x(y) dy$ where $\nu(x) > 0$ for all $x \in \mathcal{X}$.

Preliminaries and notation.

Technically, the state at time t consists of both the context x_t and the offer y_t. But since y_t depends only on x_t and not on the previous state, it is helpful to work directly with the state variable x_t and then separately take expectations over offers drawn from $q_{x_t}(\cdot)$. To this end, we develop some specialized notation. Let η'_π denote the marginal distribution over $\mathcal{X} \cup \{\tau\}$ under the discounted state occupancy measure η_π, and note that

$$\eta_\pi\Big(\{x\} \times (y_1, y_2)\Big) = \eta'_\pi(x) \int_{y_1}^{y_2} q_x(y) dy. \tag{EC.19}$$

We find it convenient to directly work with η'_π and $q_x(y)$. We now state several formulas in terms of η'_π and $q_x(y)$ that will be used throughout the analysis. For an action $a \in [0,1]$, indicating a probability of stopping,

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

where

$$c_\pi(x) := \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J_\pi(x',y') q_{x'}(y') dy' \in [0, \gamma y_{\max}] \tag{EC.20}$$

is called the "continuation value" from context x under policy π. That $c_\pi(x) \le \gamma y_{\max}$ is due the basic fact that $J_\pi(x', y') \le y_{\max}$, i.e. no policy can accrue expected reward exceeding the maximum possible offer. Similarly, the weighted policy iteration objective becomes,

$$\begin{aligned}\mathcal{B}(\theta \mid \eta_\pi, J_\pi) &= \sum_{x\in\mathcal{X}} \eta'_\pi(x) \int_{y_{\min}}^{y_{\max}} Q_\pi((x,y)\,,\, \mathbb{1}(y>\theta_x)) q_x(y) dy \\ &= \sum_{x\in\mathcal{X}} \eta'_\pi(x) \int_{y_{\min}}^{y_{\max}} \left[y\mathbb{1}(y>\theta_x) + c_\pi(x)\mathbb{1}(y<\theta_x)\right] q_x(y) dy. \end{aligned} \tag{EC.21}$$

Here we use that there is zero value-to-go from the terminal state τ to restrict the sum to continuation states. We now proceed to verify all the conditions needed to apply our general 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 $\pi \in \Pi_\Theta$, the policy iteration update for any state $s = (x,y) \in \mathcal{S}_{\mathrm{C}}$ is given by

$$\pi^+(x,y) = \arg\max_{a\in\{0,1\}} Q_\pi((x,y),a) = \mathbb{1}(y > c_\pi(x)).$$

This result can be seen immediately from the formula for $Q_\pi(\cdot)$ given above. Clearly the policy iteration update is another threshold policy. In particular, the policy iteration update to π is a new threshold policy π_{θ^+} with parameters

$$\theta_x^+ = \max\{y_{\min}\,,\, c_\pi(x)\}. \tag{EC.22}$$

for each $x \in \mathcal{X}$. This works because a policy with stopping threshold $y_{\min}$ accepts the next offer with probability one when $c_\pi(x) < y_{\min}$ and otherwise and we are assured that $c_\pi(x) \in [y_{\min}, \gamma y_{\max}]$ is a feasible choice for θ_x^+.

Condition 2.A: No suboptimal stationary points for the weighted PI objective.

This result here is implied by the gradient dominance result that follows, but it provides a warm-up for that analysis.

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,

$$\frac{\partial}{\partial\theta_x}\mathcal{B}(\theta|\eta, J_\pi)\cdot(\theta'_x - \theta_x) \le 0 \qquad \forall \theta'_x \in \mathcal{Y},$$

for every $x \in \mathcal{X}$. It is possible to separate the stationarity condition into a component wise inequality in this manner because the parameters space is the Cartesian product $\Theta = \mathcal{Y}^{|\mathcal{X}|}$.

From (EC.21), we have the following formula for the derivative:

$$\frac{\partial}{\partial\theta_x}\mathcal{B}(\theta \mid \eta_\pi, J_\pi) = (c_\pi(x) - \theta_x)\,\eta'_\pi(x) q_x(\theta_x) \tag{EC.23}$$

Therefore, θ is stationary point when $(c_\pi(x) - \theta_x)\cdot(\theta'_x - \theta_x) \le 0$ for all $\theta'_x \in [y_{\min}, y_{\max}]$. If $c_\pi(x) \in (y_{\min}, y_{\max})$ is in the interior of the feasible region, this implies $\theta_x = c_\pi(x)$. Otherwise (since it is impossible to have $c_\pi(x) \ge y_{\max}$), we have $c_\pi(x) \le y_{\min}$ and a stationary point must satisfy $\theta_x = y_{\min}$. We have found any stationary point θ satisfies, $\theta_x = \max\{y_{\min}, c_\pi(x)\}$, which matches the formula (EC.22) for the policy iteration update that maximizes $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$.

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

LEMMA 11. *For the optimal stopping problem in Example 7, 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 of Lemma 11. 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 $J, J' \in \mathcal{J}$,

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

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,

$$|TJ(x, y) - TJ'(x, y)| \leq \gamma \sum_{x'\in\mathcal{X}} p(x'|x) \int_{\mathcal{Y}} |J(x', y') - J'(x', y')|\, q_{x'}(y')dy'. \tag{EC.24}$$

As the right hand side in (EC.24) is independent of y, integrating (EC.24) with respect to $q_x(\cdot)$ gives

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

Therefore,

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

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

Further results in the supplementary technical report.
One expects that many smoothness results hold for this problem (since with a continuous offer distribution, infinitesimal changes to a stopping threshold should have an infinitesimal impact on performance), and we have already calculated one derivative in (EC.23). For completeness, we give a detailed verification of the differentiability properties in Condition 0 in the technical report Bhandari and Russo (2021).

For a reader who is interested in this specific optimal stopping problem, we note that it possible to prove stronger results that yield *convergence rates* rather than just convergence to an optimal policy. For example, in the technical report, we prove the following gradient dominance and smoothness results, which are sufficient to guarantee the convergence rates in Lemma 3. Notice that 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 EC.10 **(Gradient dominance for optimal stopping)**. *Consider the optimal stopping problem formulated in Example 7. For any $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta|\eta_\pi, J_\pi)$ is $(\beta, 0)$–gradient-dominated where $\beta = \max_{x\in\mathcal{X}, y\in\mathcal{Y}} q_x(y) / \min_{x\in\mathcal{X}, y\in\mathcal{Y}} q_x(y)$.*

LEMMA EC.11. *For the optimal stopping problem in Example 7, $\max_{\theta\in\Theta} \|\nabla^2 \ell(\theta)\| < \infty$.*

E.6. Finite horizon inventory control.

Differentiability and derivative calculations.

In inventory control, an attractive approach for computing derivatives with respect to policy parameters is described in detail in Glasserman and Tayur (1995). One can compute derivatives of expected costs with respect to the base-stock levels by calculating the derivative for many simulated sample paths (i.e. realizations of the initial state and demands) and averaging the result. Making such an argument rigorous requires justifying the exchange of an integral and derivative. This is essentially treated in past work like Glasserman and Tayur (1995), but for reproducibility we carefully verify the partial differentiability requirements in Condition 0 in the technical report (Bhandari and Russo 2021). The proof also shows that under any base-stock policy, the distribution of the inventory levels $(x_0, x_1, \cdots)$ has a density, which is a basic consequence of the assumption that the demands follow a continuous distribution.

Verifying Condition 4.

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

LEMMA EC.12. *Consider the finite horizon inventory control problem in Example 8. Let J^* be the cost-to-go function corresponding to the optimal policy. Then, for any $\pi, \pi_\theta \in \Pi_\Theta$, the weighted policy iteration objective $\mathcal{B}(\theta|\eta_\pi, J^*)$ has no suboptimal stationary points.*

Proof of Lemma EC.12. Recall $Q^*(s,a) = Q_{\pi^*}(s,a)$ denotes the Q-function corresponding to an optimal policy. We follow a classical approach to rewriting costs as a function of the target inventory level $x+a$ rather than the state and action. We find

$$\begin{aligned} Q^*((x,h),a) &= c\cdot a + \mathbb{E}_w\left[b\max\{x+a-w,0\} + p\max\{-x-a+w,0\} + J^*((x+a-w,h+1))\right] \\ &= c\cdot x + G_h(x+a) \end{aligned}$$

where the expectation is taken over the demand distribution and $G_h(y) := \mathbb{E}_w[b\max\{y-w,0\} + p\max\{-y+w,0\} + J^*((y-w,h+1))]$. Here y is thought of as a target inventory level. The function $G_h(\cdot)$ is well known to be convex (see e.g Bertsekas 1995).

Recall that $\pi_\theta((x,h)) = \max\{\theta_h - x\,,\, 0\}$. We can then calculate the derivative of the weighted PI costs in terms of $G_h(\cdot)$ as

$$\begin{aligned} \frac{\partial}{\partial\theta_h}\mathcal{B}(\theta|\eta_\pi, J^*) &= \frac{\partial}{\partial\theta_h}\int Q^*\left((x,h), \pi_\theta(x,h)\right)\,\eta_\pi(dx,h) \\ &= \frac{\partial}{\partial\theta_h}\left[\int_{x<\theta_h} Q^*((x,h),\theta_h - x)\,\eta_\pi(dx,h) + \int_{x>\theta_h} Q^*((x,h),0)\,\eta_\pi(dx,h)\right] \\ &\overset{(a)}{=} \int_{x<\theta_h} \frac{\partial}{\partial\theta_h} Q^*((x,h),\theta_h - x)\,\eta_\pi(dx,h) \end{aligned}$$

$$= \int_{x<\theta_h} G_h'(\theta_h)\, \eta_\pi(dx,h)$$
$$= G_h'(\theta_h) \int_{x<\theta_h} \eta_\pi(dx,h)$$

where (a) uses the Leibniz rule. Since θ_h is constrained to lie above 0, and negative (i.e. back-ordered) inventory levels are possible under the initial distribution, we know $\int_{x<\theta_h} \eta_\pi(dx,h) > 0$. Let θ^* an optimal vector of base-stock levels, so $\theta_h^* \in \arg\min_y G_h(y)$. If θ_h is not a minimizer of $G_h(\cdot)$ (so it is a suboptimal base-stock level), then by convexity $G_h'(\theta_h)\cdot(\theta_h^* - \theta_h) < 0$, implying θ_h cannot be a stationary point of $\mathcal{B}(\cdot|\eta_\pi, J^*)$. □

E.7. Linear MDPs: proof of Lemma 10.

The proof sketch in the body of the paper already established Condition 1 and that

$$\frac{d}{d\alpha}\mathcal{B}(\pi_{\theta^\alpha} \mid \eta_\pi, J_\pi) = \int \frac{d}{d\alpha} Q_\pi(s, \pi_{\theta^\alpha}(s))\eta_\pi(ds) = \int \frac{d}{d\alpha} Q^{\theta^+}(s, \pi_{\theta^\alpha}(s))\eta_\pi(ds).$$

Our remaining goal is to show that $\frac{d}{d\alpha} Q^{\theta^+}(s, \pi_{\theta^\alpha}(s)) < 0$ for any state at which $Q^{\theta^\alpha}(s,\cdot) \neq Q^{\theta^+}(s,\cdot)$. To see this, note that one can rewrite the problem $\max_{a\in\mathcal{A}} Q_\pi(s,a)$ as $\max_{a_{1:k-1}\in\mathcal{X}} f(a_{1:k-1}, y)$ where f is as in Lemma EC.13, $a_{1:k-1}$ are components of the action except a_k (which is redundant, as $a_k = 1 - a_1 - \cdots - a_{k-1}$) and $y_j = e_j^\top \Phi(s)\theta - e_k^\top \Phi(s)\theta$. Using Lemma EC.13, one can establishes the claim.

LEMMA EC.13. *Define* $\mathcal{X} = \{x \in (0,1)^{k-1} : \sum_{i=1}^{k-1} x_i < 1\}$ *and* $f : \mathcal{X} \times \mathbb{R}^{k-1} \to \mathbb{R}$ *by*

$$f(x,\theta) = x^\top \theta + \lambda H(x)$$

where $H(x) = \sum_{i=1}^k \frac{1}{k} \log\left(\frac{1/k}{x_i}\right)$ *with* $x_k \equiv 1 - \sum_{i=1}^{k-1} x_i$. *For fixed* $\theta_0 \in \mathbb{R}^{k-1}$, *define*

$$h(\theta) = f(x^*(\theta), \theta_0) \quad \textit{where} \quad x^*(\theta) \equiv \arg\min_{x\in\mathcal{X}} f(x,\theta).$$

Then $h(\cdot)$ *is differentiable,* θ_0 *is its unique minimizer, and* $(\theta_0 - \theta)^\top \nabla h(\theta) = -\|\theta_0 - \theta\|^2_{(\nabla^2 H(x^*(\theta)))^{-1}}/\lambda < 0$.

Proof of Lemma EC.13. The function H satisfies $\nabla^2 H(x) \succ 0$. It is well known that $x^*(\theta)$ satisfies $x_i^*(\theta) = Ce^{-\theta_i/\lambda}$ where $C = 1 + \sum_{j=1}^k e^{-\theta_j/\lambda}$. Instead of using this formula, we use implicit differentiation to derive a formula for $\nabla x^*(\theta)$. Since $x^*(\theta)$ is an interior solution (i.e. $0 < \sum_{i=2}^k x^*(\theta)_i < 1$), the first order condition implies $\theta + \lambda \nabla H(x^*(\theta)) = 0$. Differentiating this expression again and re-arranging terms yields the formula $\nabla x^*(\theta) = -\frac{1}{\lambda}\left(\nabla^2 H(x^*(\theta))\right)^{-1} \in \mathbb{R}^{(k-1)\times(k-1)}$. Then

$$\nabla h(\theta) = (\nabla x^*(\theta))\theta_0 + \lambda(\nabla x^*(\theta))\nabla H(x^*(\theta)) = (\nabla x^*(\theta))(\theta_0 - \theta) = -\lambda^{-1}\left(\nabla^2 H(x^*(\theta))\right)^{-1}(\theta_0 - \theta)$$

where the second equality uses the first order optimality condition. Taking the dot product of both sides of the equation above with $\theta_0 - \theta$ yields the result. □

Endnotes

13. Note that unlike Example 1, this policy class is closed under policy improvement.

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

15. 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.