(1) 2020 Submitted pp; Published mm/dd

Variational Policy Gradient Method for Reinforcement Learning with General Utilities

Junyu Zhang ZHAN4393@UMN.EDU
Department of Industrial and Systems Engineering
University of Minnesota
Minneapolis, Minnesota, 55455

Alec Koppel ALEC.E.KOPPEL.CIV@MAIL.MIL
Computational and Information Sciences Directorate
US Army Research Laboratory
Adelphi, MD 20783

Amrit Singh Bedi AMRIT0714@GMAIL.COM
Computational and Information Sciences Directorate
US Army Research Laboratory
Adelphi, MD, USA 20783

Csaba Szepesvari SZEPESVA@UALBERTA.CA
Department of Computer Science
DeepMind/University of Alberta
Princeton, NJ 08544

Mengdi Wang MENGDIW@PRINCETON.EDU
Department of Electrical Engineering
Center for Statistics and Machine Learning
Princeton University/Deepmind
Princeton, NJ 08544

Abstract

In recent years, reinforcement learning (RL) systems with general goals beyond a cumulative sum of rewards have gained traction, such as in constrained problems, exploration, and acting upon prior experiences. In this paper, we consider policy optimization in Markov Decision Problems, where the objective is a general concave utility function of the state-action occupancy measure, which subsumes several of the aforementioned examples as special cases. Such generality invalidates the Bellman equation. As this means that dynamic programming no longer works, we focus on direct policy search. Analogously to the Policy Gradient Theorem Sutton et al. (2000) available for RL with cumulative rewards, we derive a new Variational Policy Gradient Theorem for RL with general utilities, which establishes that the parametrized policy gradient may be obtained as the solution of a stochastic saddle point problem involving the Fenchel dual of the utility function. We develop a variational Monte Carlo gradient estimation algorithm to compute the policy gradient based on sample paths. We prove that the variational policy gradient scheme converges globally to the optimal policy for the general objective, though the optimization problem is nonconvex. We also establish its rate of convergence of the order $O(1/t)$ by exploiting the hidden convexity of the problem, and proves that it converges exponentially when the problem admits hidden strong convexity. Our analysis applies to the standard RL problem with cumulative rewards as a special case, in which case our result improves the available convergence rate.

1. Introduction

The standard formulation of reinforcement learning (RL) is concerned with finding a policy that maximizes the expected sum of rewards along the sample paths generated by the policy. The additive nature of the objective function creates an attractive algebraic structure which most efficient RL algorithms exploit. However, the cumulative reward objective is not the only one that has attracted attention. In fact, many alternative objectives made appearances already in the early literature on stochastic optimal control and operations research. Examples include various kinds of risk-sensitive objectives Kallenberg (1994); Borkar and Meyn (2002); Yu et al. (2009); Mannor and Tsitsiklis (2011), objectives to maximize the entropy of the state visitation distribution Hazan et al. (2018), the incorporation of constraints Derman and Klein (1965); Altman (1999); Achiam et al. (2017), and learning to "mimic" a demonstration Schaal (1997); Argall et al. (2009).

In this paper, we consider RL with general utility functions, and we aim to develop a principled methodology and theory for policy optimization in such problems. We focus on utility functions that are concave functionals of the state-action occupancy measure, which contains many, although not all, of the aforementioned examples as special cases.The general (or non-standard Kallenberg (1994)) utility is a strict generalization of cumulative reward, which itself can be viewed as a linear functional of the state-action occupancy measure, and as such, is a concave function of the occupancy measures.

When moving beyond cumulative rewards, we quickly run into technical challenges because of the lack of additive structure. Without additivity of rewards, the problem becomes non-Markovian in the cost-to-go Takács (1966); Whitehead and Lin (1995). Consequently, the Bellman equation fails to hold and dynamic programming (DP) breaks down. Therefore, stochastic methods based upon DP such as temporal difference Sutton (1988) and Q-learning Watkins and Dayan (1992); Ross (2014) are inapplicable. The value function, the core quantity for RL, is not even well defined for general utilities, thus invalidating the foundation of value-function based approach to RL.

Due to these challenges, we consider direct policy search methods for the solution of RL problems defined by general utility functions. We consider the most elementary policy-based method, namely the Policy Gradient (PG) method Williams (1992). The idea of policy gradient methods is that to represent policies through some policy parameterization and then move the parameters of a policy in the direction of the gradient of the objective function. When (as typical) only a noisy estimate of the gradient is available, we arrive at a stochastic approximation method Robbins and Monro (1951); Kiefer et al. (1952). In the classical cumulative reward objectives, the gradient can be written as the product of the action-value function and the gradient of the logarithm of the policy, or policy score function Sutton et al. (2000). State-of-the-art RL algorithms for the cumulative reward setting combine this result with other ideas, such as limiting the changes to the policies S Kakade (2002); Schulman et al. (2015, 2017), variance reduction Kakade (2002); Papini et al. (2018); Xu et al. (2019), or exploiting structural aspects of the policy parameterization Wang et al. (2019); Agarwal et al. (2019); Mei et al. (2020).

As mentioned, these approaches crucially rely on the standard PG Theorem Sutton et al. (2000), which is not available for general utilities. Compounding this challenge is the fact that the action-value function is not well-defined in this instance, either. Thus, how and whether the policy gradient can be effectively computed becomes a question. Further, due to the problem's nonconvexity, it is an open question whether an iterative policy improvement scheme converges to anything meaningful: In particular, while standard results for stochastic approximation would give convergence to stationary points Borkar (2009), it is unclear whether the stationary points give reasonable policies. Therefore, we ask the question:

Is policy search viable for general utilities,
when Bellman's equation, the value function, and dynamic programming all fail?

We will answer the question positively in this paper. Our contributions are three-folded:

- We derive a Variational Policy Gradient Theorem for RL with general utilities which establishes that the parametrized policy gradient is the solution to a stochastic saddle point problem.
- We show that the Variational Policy Gradient can be estimated by a primal-dual stochastic approximation method based on sample paths generated by following the current policy Arrow et al. (1958). We prove that the random error of the estimate decays at order $O(1/\sqrt{n})$ that also depends on properties of the utility, where n is the number of episodes .
- We consider the non-parameterized policy optimization problem which is nonconvex in the policy space. Despite the lack of convexity, we identify the problem's hidden convexity, which allows us to show that a variational policy gradient ascent scheme converges to the global optimal policy for general utilities, at a rate of $O(1/t)$, where t is the iteration index. In the special case of cumulative rewards, our result improves upon the best known convergence rate $O(1/\sqrt{t})$ for tabular policy gradient Agarwal et al. (2019), and matches the convergence rate of variants of the algorithm such as softmax policy gradient Mei et al. (2020) and natural policy gradient Agarwal et al. (2019). In the case where the utility is strongly concave in occupancy measures (e.g., utilities involving Kullback-Leiber divergence), we established the exponential convergence rate of the variational gradient scheme.

Related Work. Policy gradient methods have been extensive studied for RL with cumulative returns. There is a large body of work on variants of policy-based methods as well as theoretical convergence analysis for these methods. Due to space constraints, we defer a thorough review to Supplement A.

Notation. We let $\mathbb{R}$ denote the set of reals. We also let $\|\cdot\|$ denote the 2-norm, while for matrices we let it denote the spectral norm. For the p-norms ($1 \leq p \leq \infty$), we use $\|\cdot\|_p$. For any matrix B, $\|B\|_{\infty,2} := \max_{\|u\|_\infty \leq 1} \|Bu\|_2$. For a differentiable function f, we denote by ∇f its gradient. If f is nondifferentiable, we denote by $\hat{\partial} f$ the Fréchet superdifferential of f; see e.g. Drusvyatskiy and Paquette (2019).

2. Problem Formulation

Consider a Markov decision process (MDP) over the finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$. For each state $i \in \mathcal{S}$, a transition to state $j \in \mathcal{S}$ occurs when selecting action $a \in \mathcal{A}$ according to a conditional probability distribution $j \sim \mathcal{P}(\cdot|a, i)$, for which we define the short-hand notation $P_a(i, j)$. Let ξ be the initial state distribution of the MDP. We let S denote the number of states and A the number of actions. The goal is to prescribe actions based on previous states in order to maximize some long term objective. We call $\pi : \mathcal{S} \to P(\mathcal{A})$ a *policy* that maps states to distributions over actions, which we subsequently stipulate is stationary. In the standard (cumulative return) MDP, the objective is to maximize the expected cumulative sum of future rewards Puterman (2014), i.e.,

$$\max_{\pi} V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{s_t a_t} \;\middle|\; i_0 = s, a_t \sim \pi(\cdot|s_t), t = 0, 1, \ldots\right], \quad \forall s \in \mathcal{S}. \tag{2.1}$$

with reward $r_{s_t a_t} \in \mathbb{R}$ revealed by the environment when action a_t is chosen at state s_t.

In this paper we consider policy optimization for maximizing general objective functions that are not limited to cumulative rewards. In particular, we consider the problem

$$\max_{\pi} R(\pi) := F(\lambda^{\pi}) \tag{2.2}$$

where λ^{π} is known as the *cumulative discounted state-action occupancy measure*, or *flux* under policy π, and F is a general concave functional. Denote $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{L}$ as the set of policy and flux respectively,

then λ^π is given by the mapping $\Lambda : \Delta_{\mathcal{A}}^{\mathcal{S}} \mapsto \mathcal{L}$ as

$$\lambda_{sa}^\pi = \Lambda_{sa}(\pi) := \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\Big(s_t = s, a_t = a \;\Big|\; \pi, s_0 \sim \xi\Big) \quad \text{for} \quad \forall a \in \mathcal{A}, \forall s \in \mathcal{S}\,. \tag{2.3}$$

Similar to the LP formulation of a standard MDP, we can write (2.2) equivalently as an optimization problem in λ (see Zhang et al. (2020a)), giving rise to

$$\max_{\lambda} F(\lambda) \qquad \text{s.t.} \qquad \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi, \lambda \geq 0, \tag{2.4}$$

where $\lambda_a = [\lambda_{1a}, \cdots, \lambda_{Sa}]^\top \in \mathbb{R}^A$ is the a-th column of λ and ξ is the initial distribution over the state space $\mathcal{S}$. The constraints require that λ be the unnormalized state-action occupancy measure corresponding to *some* policy. In fact, it is well known that a policy π inducing λ can be extracted from λ using the mapping $\Pi : \mathcal{L} \mapsto \Delta_{\mathcal{A}}^{\mathcal{S}}$ as $\pi(a|s) = \Pi_{sa}(\lambda) := \frac{\lambda_{sa}}{\sum_{a'\in\mathcal{A}} \lambda_{sa'}}$ for all a, s.

Problem (2.2) contains the original MDP problem as a special case. To be specific, when $F(\lambda) = \langle r, \lambda\rangle$ with $r \in \mathbb{R}^{SA}$ as the reward function, then $F(\lambda) = \langle \lambda, r\rangle = \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r_{s_t a_t} \;\big|\; \pi, s_0 \sim \xi\big]$. This means that (2.4) is a generalization of (2.1), and reduces to the dual LP formulation of standard MDP for this (linear) choice of $F(\cdot)$ Kallenberg (1983). We focus on the case where F is concave, which makes (2.4) a concave (hence, convenient) maximization problem. Next we introduce a few examples that arise in practice for incentivizing safety, exploration, and imitation, respectively.

Example 2.1 (**MDP with Constraints or Barriers).** In discounted constrained MDPs the goal is to maximize the total expected discounted reward under a constraint where for some cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the total expected discounted cost incurred by the chosen policy is constrained from above. Letting r denote the reward function over $\mathcal{S} \times \mathcal{A}$, the underlying optimization problem becomes

$$\max_{\pi} v_r^\pi := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \qquad \text{s.t.} \qquad v_c^\pi := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \leq C. \tag{2.5}$$

As is well known, a relaxed formulation is

$$\max_{\lambda} F(\lambda) := \langle \lambda, r\rangle - \beta \cdot p(\langle \lambda, c\rangle - C) \qquad \text{s.t.} \qquad \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi, \lambda \geq 0. \tag{2.6}$$

where p is a penalty function (e.g., the log barrier function).

Example 2.2 (**Pure Exploration).** In the absence of a reward function, an agent may consider the problem of finding a policy whose stationary distribution has the largest "entropy", as this should facilitate maximizing the speed at which the agent explores its environment Hazan et al. (2018):

$$\max_{\pi} R(\pi) := \text{Entropy}(\bar{\lambda}^\pi), \tag{2.7}$$

where $\bar{\lambda}^\pi$ is the normalized state visitation measure given by $\bar{\lambda}_s^\pi = (1-\gamma)\sum_a \lambda_{sa}^\pi$ for all s. Various entropic measures are possible, but the simplest is the negative log-likelihood: $\text{Entropy}(\bar{\lambda}^\pi) = -\sum_s \bar{\lambda}_s^\pi \log[\bar{\lambda}_s^\pi]$. As is well known, this entropy is (strongly) concave.

Another example, when d state-action features $\boldsymbol{\phi}(s, a) \in \mathbb{R}^d$ are available, is to cover the entire feature space by maximizing the smallest eigenvalue of the covariance matrix:

$$\max_{\pi} R(\pi) := \sigma_{\min}\left(\mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^t \boldsymbol{\phi}(s_t, a_t)\boldsymbol{\phi}(s_t, a_t)^\top\right]\right). \tag{2.8}$$

In (2.8), observe that $\mathbb{E}^\pi[\sum_{t=1}^\infty \gamma^t \boldsymbol{\phi}(s_t,a_t)\boldsymbol{\phi}(s_t,a_t)^\top] = \sum_{sa} \lambda^\pi_{sa} \cdot \boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top$. By Rayleigh principle, it is again a concave function of λ.

Example 2.3 (**Learning to mimic a demonstration).** When demonstrations are available, they may be employed to obtain information about a prior policy in the form of a state visitation distribution $\bar{\mu}$. Remaining close to this prior can be achieved by minimizing the Kullback-Liebler (KL) divergence between the state marginal distribution of λ and the prior $\bar{\mu}$ stated as

$$F(\lambda) = \mathrm{KL}\Big((1-\gamma)\sum_a \lambda_a || \bar{\mu}\Big) \tag{2.9}$$

which, when substituted into (2.4), yields a method for ensuring some baseline performance. We further note that in place of KL divergence, one can also use other convex distances such as Wasserstein, total variation, or Hellinger distances.

Additional instances may be found in Zhang et al. (2020a). With the setting clarified, we shift focus to developing an algorithmic solution to (2.4), that is, to solve for policy π.

3. Variational Policy Gradient Theorem

To handle the curse of dimensionality, we allow parametrization of the policy by $\pi = \pi_\theta$, where $\theta \in \Theta \subset \mathbb{R}^d$ is the parameter vector. In this way, we can narrow down the policy search problem to within a d-dimensional parameter space rather than the high-dimensional space of tabular policies. The policy optimization problem then becomes

$$\max_{\theta\in\Theta} \; R(\pi_\theta) := F(\lambda^{\pi_\theta}) \tag{3.1}$$

where F is the concave utility of the state-action occupancy measure $\lambda(\theta) := \lambda^{\pi_\theta}$, $\Theta \subset \mathbb{R}^d$ is a convex set. We seek to solve for the policy maximizing the utility as in (3.1) using gradient ascent over the parameter space Θ. Note that (3.1) is simply (2.2) with parameterization θ of policy π substituted. We denote by $\nabla_\theta R(\pi_\theta)$ the parameterized policy gradient of general utility.

First, recall the policy gradient theorem for RL with cumulative rewards Sutton et al. (2000). Let the reward function be r. Define $V(\theta; r) := \langle \lambda(\theta), r\rangle$, i.e., the total expected discounted reward under the reward function r and the policy π_θ. The Policy Gradient Theorem states that

$$\nabla_\theta V(\theta; r) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^\infty \gamma^t Q^{\pi_\theta}(s_t,a_t;r)\cdot \nabla_\theta \log \pi_\theta(a_t|s_t)\right], \tag{3.2}$$

where $Q^\pi(s,a;r) := \mathbb{E}^\pi\big[\sum_t \gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a, a_t \sim \pi(\cdot \mid s_t)\big]$. Unfortunately, this elegant result no longer holds when we consider a general function instead of cumulative rewards: The policy gradient theorem relies on the additivity of rewards, which is lost in our problem. For future reference, we denote $Q^\pi(s,a;z) := \mathbb{E}^\pi\big[\sum_t \gamma^t z_{s_t a_t} \mid s_0 = s, a_0 = a, a_t \sim \pi(\cdot \mid s_t)\big]$ where z is any "function" of the state-action pairs ($z \in \mathbb{R}^{SA}$). Moreover, $V(\theta; z)$ is defined similarly. These definitions are motivated by subsequent efforts to derive an expression for the gradient of (3.1).

3.1 Policy Gradient of $R(\pi_\theta)$

Now we derive the policy gradient of $R(\pi_\theta)$ with respect to θ. By the chain rule, the gradient of $F(\lambda(\theta)) := F(\lambda^{\pi_\theta})$, using the definition of $R(\pi_\theta)$, yields (assuming differentiability of F, λ):

$$\nabla_\theta R(\pi_\theta) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{\partial F(\lambda(\theta))}{\partial \lambda_{sa}} \cdot \nabla_\theta \lambda_{sa}(\theta). \tag{3.3}$$

To directly use the chain rule, one needs the partial derivatives $\frac{\partial F(\lambda(\theta))}{\partial \lambda_{sa}}$ and $\nabla_\theta \lambda_{sa}(\theta)$. Unfortunately, neither of them is easy to estimate. The partial gradient $\frac{\partial F(\lambda(\theta))}{\partial \lambda_{sa}}$ is a function of the current state-action occupancy measure λ^{π_θ}. One might attempt to estimate the measure λ^{π_θ} and then evaluate the gradient map $\frac{\partial F(\lambda(\theta))}{\partial \lambda_{sa}}$. However, estimates of distributions over large spaces tend to converge very slowly Tsybakov (2008).

As it turns out, a viable alternate route is to consider the Fenchel dual F^* of F. Recall that $F^*(z) = \inf_\lambda \langle \lambda, z\rangle - F(\lambda)$, where we use $\langle x, y\rangle := x^\top y$ (since F is concave, the dual is defined using inf, instead of sup). As is well known, for F concave, under mild regularity conditions, the bidual (dual of the dual) of F is equal to F. This forms the basis of our first result, which states that the steepest policy ascent direction of (3.1) is the solution to a stochastic saddle point problem. The proofs of this and subsequent results are given in the supplementary material.

Theorem 3.1 (**Variational Policy Gradient Theorem**). *Suppose F is concave and continuously differentiable in an open neighborhood of λ^{π_θ}. Denote $V(\theta; z)$ to be the cumulative value of policy π_θ when the reward function is z, and assume $\nabla_\theta V(\theta; z)$ always exists. Then we have*

$$\nabla_\theta R(\pi_\theta) = \lim_{\delta \to 0_+} \operatorname*{argmax}_x \inf_z \left\{ V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2 \right\}. \tag{3.4}$$

Therefore, to estimate $\nabla_\theta R(\pi_\theta)$ we require the cumulative return $V(\theta; z)$ of the function z, its associated "vanilla" policy gradient (3.2), and the gradient of the Fenchel dual of F at z. These ingredients are combined via (3.4) to obtain a valid policy gradient for general objectives. Next, we discuss how to estimate the gradient using sampled trajectories.

3.2 Estimating the Variational Policy Gradient

Theorem 3.1 implies that one can estimate $\nabla_\theta R(\pi_\theta)$ by solving a stochastic saddle point problem. Suppose we generate n i.i.d. episodes of length K following π_θ, denoted as $\zeta_i = \{s_k^{(i)}, a_k^{(i)}\}_{k=1}^K$. Then we can estimate $V(\theta; z)$ and $\nabla V(\theta; z)$ for any function z by

$$\tilde{V}(\theta; z) := \frac{1}{n}\sum_{i=1}^n V(\theta; z; \zeta_i) := \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \gamma^k \cdot z(s_k^{(i)}, a_k^{(i)}), \tag{3.5}$$

$$\nabla \tilde{V}(\theta; z) := \frac{1}{n}\sum_{i=1}^n \nabla_\theta V(\theta; z; \zeta_i) := \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \sum_{a\in\mathcal{A}} \gamma^k \cdot Q(s_k^{(i)}, a; z) \nabla_\theta \pi_\theta(a|s_k^{(i)}).$$

For a given value of K, the error introduced by "truncating" trajectories at length K is of order $\gamma^K/(1-\gamma)$, which quickly decays to zero for $\gamma < 1$. Plugging in the obtained estimates into (3.4) gives rise to the sample-average approximation to the policy gradient:

$$\hat{\nabla}_\theta R(\pi_\theta; \delta) := \operatorname*{argmax}_x \inf_{\|z\|_\infty \le \ell_F} \left\{ -F^*(z) + \tilde{V}(\theta; z) + \delta \nabla_\theta \tilde{V}(\theta; z)^\top x - \frac{\delta}{2}\|x\|^2 \right\}, \tag{3.6}$$

where ℓ_F is defined in the next theorem. Therefore, any algorithm that solves problem (3.6) will serve our purpose. A MC stochastic approximation scheme, i.e., Algorithm 1, is provided in Appendix B.1.

Theorem 3.2 (**Error bound of policy gradient estimates**). *Suppose the following holds:*
(i) *$\mathrm{dom} F = \mathbb{R}^{SA}$, there exists ℓ_F such that $\max\{\|\nabla F(\lambda)\|_\infty : \|\lambda\|_1 \le \frac{2}{1-\gamma}\} \le \ell_F$.*
(ii) *F is L_F-smooth under L_1 norm, i.e., $\|\nabla F(\lambda) - \nabla F(\lambda')\|_\infty \le L_F \|\lambda - \lambda'\|_1$.*
(iii) *F^* is (ℓ_{F^*})-Lipschitz with respect to the L_∞ norm in the set $\{z : \|z\|_\infty \le 2\ell_F,\ F^*(z) > -\infty\}$.*
(iv) *There exists C with $\|\nabla_\theta \pi(\cdot|s)\|_{\infty,2} \le C$, where $\nabla_\theta \pi(\cdot|s) = [\nabla_\theta \pi(1|s), \cdots, \nabla_\theta \pi(A|s)]$.*
Let $\hat{\nabla}_\theta R(\pi_\theta) := \lim_{\delta\to 0_+} \hat{\nabla}_\theta R(\pi_\theta; \delta)$. Then

$$\mathbb{E}[\|\hat{\nabla}_\theta R(\pi_\theta) - \nabla_\theta R(\pi_\theta)\|^2] \le \mathcal{O}\left(\frac{C^2(\ell_F^2 + L_F^2 \ell_{F^*}^2)}{n(1-\gamma)^4} + \frac{C^2 L_F^2}{n(1-\gamma)^6} \right) + \mathcal{O}(\gamma^K).$$

Remarks.
(1) Theorem 3.2 suggests an $O(1/\sqrt{n})$ error rate, proving that the variational policy gradient - though more complicated than the typical policy gradient that takes the form of a mean - can be efficiently estimated from finite data.
(2) Although the variable z is high dimensional, our error bound depends only on the properties of F.
(3) We assumed for simplicity that Q values are known. In practice, they can estimated by, e.g., an additional Monte Carlo rollout on the same sample path or temporal difference learning. As long as the estimator for $Q(s,a;z)$ is unbiased and upper bouded by $\mathcal{O}(\frac{\|z\|_\infty}{1-\gamma})$, the result will not change.
(4) For the case of cumulative rewards, we have $F(\lambda) = \langle r, \lambda\rangle$, so that $\ell_F = \|r\|_\infty$, $\ell_{F^*}{=}0, L_F{=}0$. Therefore $\mathbb{E}[\|\hat{\nabla}_\theta R(\pi_\theta) - \nabla_\theta R(\pi_\theta)\|^2] \le \mathcal{O}\left(\frac{C^2\|r\|_\infty^2}{n(1-\gamma)^4}\right)$.

Special cases of $\nabla_\theta R(\pi^\theta)$. We further explain how to obtain the variational policy gradient for several special cases of R, including constrained MDP, maximal exploration, and learning from demonstrations. See Appendix B.2 for more details.

4. Global Convergence of Policy Gradient Ascent

In this section, we analyze policy search for the problem (3.1), i.e., $\max_{\theta\in\Theta} R(\pi_\theta)$ via gradient ascent:

$$\theta^{k+1} = \operatorname*{argmax}_{\theta\in\Theta} R(\pi_{\theta^k}) + \big\langle \nabla_\theta R(\pi_{\theta^k}), \theta - \theta^k\big\rangle - \frac{1}{2\eta}\|\theta-\theta^k\|^2 = \mathrm{Proj}_\Theta\big\{\theta^k + \eta\nabla_\theta R(\pi_{\theta^k})\big\} \tag{4.1}$$

where $\mathrm{Proj}_\Theta\{\cdot\}$ denotes Euclidean projection onto Θ, and equivalence holds by the convexity of Θ.

4.1 No spurious first-order stationary solutions.

We study the geometry of the (possibly) nonconvex optimization problem (3.1). When F is a linear function of λ, and the parameterization is tabular or softmax, existing theory of cumulative-return RL problems have shown that every first-order stationary point of (3.1) is globally optimal – see Agarwal et al. (2019); Mei et al. (2020).

In what follows, we show that the problem (3.1) has no spurious extrema despite of its nonconvexity, for general utility functions and policy parametrization. Specifically, to generalize global optimality attributes of stationary points of (3.1) from (2.1), we exploit structural aspects of the relationship between occupancy measures and parameterized families of policies, namely, that these entities are related through a bijection. This bijection, when combined with the fact that (3.1) is concave in λ, and suitably restricting the parameterized family of policies, is what we subsequently describe as "hidden convexity." For these results to be valid, we require the following regularity conditions.

Assumption 4.1. *Suppose the following holds true:*
(i). *$\lambda(\cdot)$ forms a bijection between Θ and $\lambda(\Theta)$, where Θ and $\lambda(\Theta)$ are closed and convex.*
(ii). *The Jacobian matrix $\nabla_\theta\lambda(\theta)$ is Lipschitz continuous in Θ.*
(iii). *Denote $g(\cdot) := \lambda^{-1}(\cdot)$ as the inverse mapping of $\lambda(\cdot)$. Then there exists $\ell_\theta > 0$ s.t. $\|g(\lambda) - g(\lambda')| \le \ell_\theta|||\lambda - \lambda'|||$ for some norm $|||\cdot|||$ and for all $\lambda, \lambda' \in \lambda(\Theta)$.*

In particular, for the direct policy parametrization, also known as the "tabular" policy case, we have $\lambda(\theta) := \Lambda(\pi)$ where Λ is defined in (2.3). When ξ is positive-valued, Assumption 4.1 is true for the tabular policy case (as established in Appendix H).

Theorem 4.2 (Global optimality of stationary policies). *Suppose Assumption 4.1 holds, and F is a concave, and continuous function defined in an open neighbourhood containing $\lambda(\Theta)$. Let θ^* be a first-order stationary point of problem (3.1), i.e.,*

$$\exists u^* \in \hat{\partial}(F\circ\lambda)(\theta^*), \quad s.t. \quad \langle u^*, \theta - \theta^*\rangle \le 0 \qquad for \qquad \forall\theta\in\Theta. \tag{4.2}$$

Then θ^* is a globally optimal solution of problem (3.1).

Theorem 4.2 provides conditions such that, despite of nonconvexity, local search methods can find the global optimal policies. Since we aim at general utilities, we naturally separated out the convex and non-convex maps in the composite objective and our conditions for optimality rely on the properties of these. In a recent paper, Bhandari and Russo (2020) proposed some sufficient conditions under which a result similar to Theorem 4.2 holds in the setting of the standard, cumulative total reward criterion. Their conditions are *(i)* the policy class is closed under (one-step, weighted) policy improvement and that *(ii)* all stationary points of the one-step policy improvement map are global optima of this map. It remains for future work to see the relationship between our conditions and these conditions: They appear to have rather different natures.

4.2 Convergence analysis

Now we analyze the convergence rate of the policy gradient scheme (4.1) for general utilities.

Assumption 4.3. *There exists $L > 0$ such that the policy gradient $\nabla_\theta R(\pi_\theta)$ is L-Lipschitz.*

The objective $R(\pi_\theta)$ is nonconvex in θ, so one might expect that gradient schemes converge to stationary solutions at a standard $\mathcal{O}(1/\sqrt{t})$ convergence rate Shapiro et al. (2014). Remarkably, the policy optimization problem admits a convex nature if we view it in the space of λ, as long as F is concave. By exploiting this hidden convexity, we establish an $\mathcal{O}(1/t)$ convergence rate for solving RL with general utilities. Further, we show that, when the utility F is strongly concave, the gradient ascent scheme converges to the globally optimal policy exponentially fast.

Theorem 4.4 (**Convergence rate of parameterized policy gradient iteration).** *Let Assumptions 4.1 and 4.3 hold. Denote $D_\lambda := \max_{\lambda,\lambda' \in \lambda(\Theta)} |||\lambda - \lambda'|||$ as defined in Assumption 4.1(iii). Then the policy gradient update* (4.1) *with $\eta = 1/L$ satisfies for all k*

$$R(\pi_{\theta^*}) - R(\pi_{\theta^k}) \leq \frac{4L\ell_\theta^2 D_\lambda^2}{k+1}.$$

Additionally, if $F(\cdot)$ is μ-strongly concave with respect to the $|||\cdot|||$ norm, we have

$$R(\pi_{\theta^*}) - R(\pi_{\theta^k}) \leq \Big(1 - \frac{1}{1 + L\ell_\theta^2/\mu}\Big)^k \left(R(\pi_{\theta^*}) - R(\pi_{\theta^0})\right).$$

The exponential convergence result of Theorem 4.4 implies that, when a regularizer like Kullback-Leiber divergence is used, policy gradient method converges much faster. In other words, policy search with general utilities can actually be easier than the typical, cumulative-return problem.

Finally, we study the case where policies are not parameterized, i.e., $\theta = \pi$. The next theorem establishes a tighter convergence rate than what Theorem 4.4 already implies.

Theorem 4.5 (**Convergence rate of tabular policy gradient iteration).** *Let $\theta = \pi$ and $\lambda(\theta) = \Lambda(\pi)$. Let Assumption 4.3 hold and assume that ξ is positive-valued. Then the iterates generated by* (4.1) *with $\eta = 1/L$ satisfy for all $k \geq 1$ that*

$$R(\pi^*) - R(\pi^k) \leq \frac{20L|\mathcal{S}|}{(1-\gamma)^2(k+1)} \cdot \left\| d_\xi^{\pi^*} / \xi \right\|_\infty^2.$$

The case of cumulative rewards. Let us consider the well-studied special case where F is a linear functional, i.e., $R(\pi) = V^\pi$ [cf. (2.1)] is the typical cumulative return. In this case, we have $L = \frac{2\gamma A}{(1-\gamma)^3}$ (Agarwal et al. (2019)). Now in order to obtain an ϵ-optimal policy $\bar{\pi}$ such that $V^{\pi^*} - V^{\bar{\pi}} \leq \epsilon$, the gradient ascent update requires $\mathcal{O}\Big(\frac{SA}{(1-\gamma)^5 \epsilon} \cdot \|d_\xi^{\pi^*}/\xi\|_\infty^2\Big)$ iterations according to Theorem 4.5. This bound is strictly smaller than the $\mathcal{O}\Big(\frac{SA}{(1-\gamma)^6 \epsilon^2} \|d_\xi^{\pi^*}/\xi\|_\infty^2\Big)$ iteration complexity proved by Agarwal et al. (2019) for tabular policy gradient. The improvement from $O(1/\epsilon^2)$ to $(1/\epsilon)$ comes from the fact that, although the policy optimization problem is nonconvex, our analysis exploits its hidden convexity in the space of λ.

5. Experiments

Now we shift to numerically validating our methods and theory on OpenAI Frozen Lake Brockman et al. (2016). Throughout, additional details may be found in Appendix C.

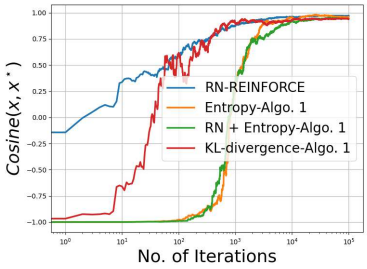
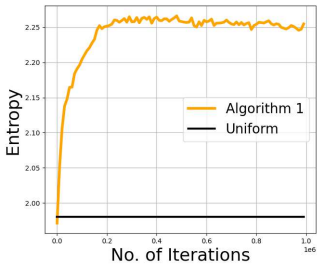
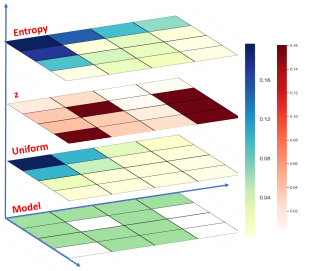


Figure 1: **PG estimation via Alg. 1** Cosine similarity between PG estimates $\hat{x}_t$ generated by Algorithm 1 after t samples and the ground truth $x^\star$, which consistently converges to near 1 across different instances (E.g. (2.1) - (2.3)) when t becomes large. For comparison, we also include the convergence of PG estimates from REINFORCE for cumulative returns.



(a) Entropy vs. # episodes



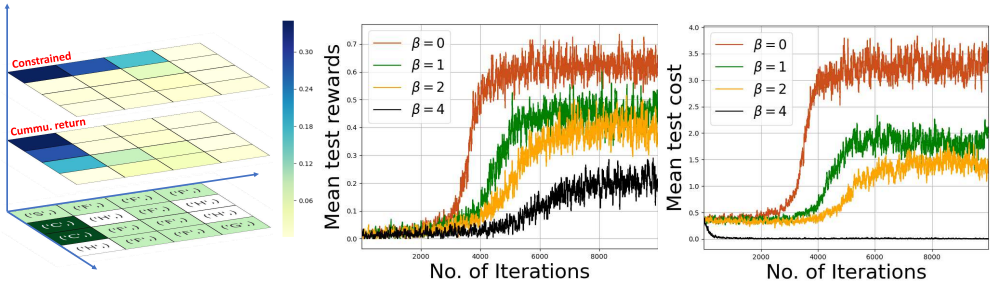
(b) World & occupancy dist. (Entropy)

Figure 2: **Results for maximum entropy exploration**: In Fig. 2(a), to quantify exploration, we present the entropy of flux λ over training index n for our approach, as compared with the entropy of a uniform random policy. Fig. 2(b)(bottom) visualizes the world model (holes in the lake have null entropy, as they terminate the episode), the lower middle layer displays the occupancy measure associated with a uniformly random policy, the upper-middle visualizes the pseudo-reward z^* defined by the Fenchel dual of the entropy (2.7) – see Appendix B.2. Lastly, on top we visualize the occupancy measure associated with the max entropy policy, which better covers the space than a uniformly random policy.

Policy Gradient (PG) Estimation. First we investigate the use of Theorem 3.1 and Algorithm 1 (Appendix B.1) for PG estimation, for several instances of the general utility. We also compare it with the gradient estimates computed by REINFORCE for cumulative returns. Specifically, in Figure 1 we illustrate the convergence of gradient estimates, measured using the cosine similarity between x_n (running estimate based on n episodes) and the true gradient x^* (which is evaluated using brute force Monte Carlo rollouts – see Appendix C.2). The cosine similarity converges to 1 across different instances, providing evidence that Algorithm 1 yields consistent gradient estimates for general utilities.

PG Ascent for Maximal Entropy Exploration. Next, we consider maximum entropy exploration (2.7) using algorithm (4.1), with softmax parametrization. First, we display the evolution of the entropy of the normalized occupancy measure over the number of episodes in Fig. 2(a). Then, we visualize the world model in Fig. 2(b)(bottom). Moreover, the lower middle is the occupancy measure associated with a uniformly random policy, the upper-middle layer visualizes the "pseudo-reward" z^* computed as the Fenchel dual of the entropy (2.7) – see Appendix B.2, which is null at the holes and positive otherwise. We use a different color to denote that its values are not likelihoods. The occupancy measure obtained by policy gradient ascent with gradient estimated by Algorithm 1 at the end of training is in Figure 2(b)(top) – observe the maximal entropy policy achieves significantly better coverage of the state space than the uniformly random policy.

PG Ascent for Avoiding Obstacles. Suppose our goal is to navigate the Frozen Lake and avoid obstacles. We consider imposing penalties to avoid costly states [cf. (2.6)] via a logarithmic barrier (B.3), and by applying variational PG ascent, we obtain an optimal policy whose resulting occupancy measure is depicted in Fig. 3(a)(top). For comparison, we consider optimizing the standard expected cumulative return (2.1), whose state occupancy measure is given in Fig. 3(a)(middle). Observe that imposing log penalties yields policies whose probability mass is concentrated away from obstacles (dark green). Further, we display in Fig. 3 the reward 3(b) and cost 3(c) accumulation during test



(a) World & occupancy dist. (CMDP) (b) Reward vs. # episodes (c) Cost vs. # episodes

Figure 3: **Results for avoiding obstacles.** Fig. 3(a)(bottom) depicts the world model of OpenAI Frozen Lake with augmentation to include costly states, e.g., obstacles: C represents costly states, F is the frozen lake, H is the hole, and G is the goal. We consider softmax policy parameterization, and visualize the occupancy measure associated with REINFORCE for the cumulative return (2.1) in the middle layer, and the relaxed **CMDP** (2.6) via a **logarithmic barrier** (B.3) at the top.The policy obtained via barriers avoids visiting costly states, in contrast to the middle. Fig. 3(b) and Fig. 3(c) show the reward/cost accumulated during test trajectories over training index for Algorithm 1. Observe that the reward/cost curves behave differently as the penalty parameter β varies: observe that without any constraint imposition (which implies $\beta = 0$ in red), one achieves the highest reward, but incurs the most costs, i.e., hits obstacles most often. Larger β imposes more penalty, and hence $\beta = 4$ incurs lowest cost and lowest reward. Other instances are also shown for $\beta = 1$ and $\beta = 2$.

trajectories as a function of the iteration index for the PG ascent (4.1) for the cumulative return (2.1) as compared with a logarithmic barrier imposed to solve (2.6) for different penalty parameters β.

6. Broader Impact

While RL has a great number of potential applications, our work is of foundational nature and as such, the application of the ideas in this paper can have both broad positive and negative impacts. However, this paper is purely theoretical, as we do not aim at any specific application, there is nothing we can say about the most likely broader impact of this work that would go beyond speculation.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 22–31. JMLR. org, 2017.

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

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

K.J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-Linear Programming*, volume II of *Stanford Mathematical Studies in the Social Sciences*. Stanford University Press, Stanford, December 1958.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv arXiv:1906.01786*, 2019.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. working paper, 2020. URL `https://djrusso.github.io/docs/policy_grad_optimality.pdf`.

Shalabh Bhatnagar, Richard Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Vivek S Borkar. *Stochastic approximation: A dynamical systems viewpoint*. Cambridge University Press, 2008.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 77. Wiley, 2009.

Vivek S Borkar and Sean P Meyn. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

Cyrus Derman and Morton Klein. Some remarks on finite horizon Markovian decision models. *Operations Research*, 13(2):272–278, April 1965. URL `https://doi.org/10.1287/opre.13.2.272`.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

Maryam Fazel, Rong Ge, Sham M 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.

Jerzy A. Filar, L. C. M. Kallenberg, and Huey-Miin Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989. URL `http://pubsonline.informs.org/doi/abs/10.1287/moor.14.1.147`.

Elad Hazan, Sham M Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. *arXiv preprint arXiv:1812.02690*, 2018.

Ying Huang and L. C. M. Kallenberg. On finding optimal policies for Markov decision chains: A unifying framework for mean-variance-tradeoffs. *Mathematics of Operations Research*, 19(2): 434–448, 1994. URL `http://pubsonline.informs.org/doi/abs/10.1287/moor.19.2.434`.

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

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(Jun):1865–1890, 2012.

L C M Kallenberg. *Linear Programming and Finite Markovian Control Problems*. CWI Mathematisch Centrum, 1983.

L. C. M. Kallenberg. Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory. *Zeitschrift für Operations Research*, 40(1):1–42, 1994.

Jack Kiefer, Jacob Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

Vijaymohan R Konda and Vivek S Borkar. Actor-critic–type learning algorithms for Markov Decision Processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, 1999.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Shie Mannor and John N Tsitsiklis. Mean-variance optimization in Markov decision processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 177–184, 2011.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*, 2020.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1806.05618*, 2018.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In *Advances in Neural Information Processing Systems*, pages 1394–1402, 2013.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in Lipschitz Markov Decision Processes. *Machine Learning*, 100(2-3):255–283, 2015.

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

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Sheldon M Ross. *Introduction to stochastic dynamic programming*. Academic press, 2014.

J Langford S Kakade. Approximately optimal approximate reinforcement learning. In *ICML*, pages 267–274, 2002.

Stefan Schaal. Learning from demonstration. In *Advances in neural information processing systems*, pages 1040–1046, 1997.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

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

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

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.

L Takács. Non-Markovian processes. In *Stochastic Process: Problems and Solutions*, pages 46–62. Springer, 1966.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Steven D Whitehead and Long-Ji Lin. Reinforcement learning of non-Markov decision processes. *Artificial intelligence*, 73(1-2):271–306, 1995.

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

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

Y.-L. Yu, Y. Li, D. Schuurmans, and Cs. Szepesvári. A general projection property for distribution families. In *Advances in Neural Information Processing Systems*, 2009.

Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. Cautious reinforcement learning via distributional risk in the dual domain. *arXiv preprint arXiv:2002.12475*, 2020a.

Junyu Zhang, Mingyi Hong, Mengdi Wang, and shuzhong Zhang. Generalization bounds for stochastic saddle point problems. *arXiv preprint arXiv:2006.02067*, 2020b.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019.

Supplementary Material for "Variational Policy Gradient Method for Reinforcement Learning with General Utilities"

Appendix A. Related Work

We provide a more extension discussion for the context of this work. Firstly, when closed-form expressions for the optimizer of a function are unavailable, solving optimization problems requires iterative schemes such as gradient ascent Nocedal and Wright (2006). Their convergence to global extrema is predicated on concavity and the tractability of computing ascent directions. When the objective takes the form of an expected value of a function parameterized by a random variable, stochastic approximations are required Robbins and Monro (1951); Kiefer et al. (1952). The PG Theorem mentioned above gives a specific form for obtaining ascent directions with respect to a parameterized family of stationary policies via trajectories in a Markov decision process, when the objective is the expected cumulative return Sutton et al. (2000), which gives rise to the REINFORCE algorithm.

The convergence of policy search for the expected cumulative return has been studied extensively in recent years. Under general parameterizations the problem becomes nonconvex. Hence, early work focused on asymptotic convergence to stationarity Pirotta et al. (2015) by invoking dynamical systems Borkar (2008). In actor-critic Konda and Borkar (1999); Konda and Tsitsiklis (2000), one replaces the Monte Carlo rollout of the Q function with a temporal difference estimator Sutton (1988), and its asymptotic stability follows similar logic Bhatnagar et al. (2009). Another line of work focused on only on per-step value increase, i.e., policy improvement bounds Pirotta et al. (2013, 2015). Recent interest has been on structural results that yield convergence to global optimality: when state transitions are linear Fazel et al. (2018); Bu et al. (2019)), the policy parameterization is direct (tabular) Bhandari and Russo (2019); Agarwal et al. (2019), function approximation error can be quantified S Kakade (2002); Liu et al. (2019). Clever step-size rules have also been designed to ensure convergence to second-order stationary points under general settings Zhang et al. (2019).

These results, however, are restricted to the expected cumulative return, a linear functional of the state-action occupancy measure, and hence do not apply to general concave functionals of the form considered in this work. Early works in operations research consider nonstandard utilities Huang and Kallenberg (1994), motivated by certain variance-penalizations which may also be written as concave functionals of occupancy measures Filar et al. (1989). Similar in spirit to this work is Kallenberg (1994), as it also puts occupancy measures at the center of its conceptual development. These works develop dynamic programming approaches for tabular settings, and hence are not scalable to problems with large spaces. More recently, maximizing the entropy of the state visitation distribution has been considered Hazan et al. (2018), a special case of the concave utilities we study. Moreover, the authors develop a model-based iteratively policy update, which requires explicit knowledge of the transition probability matrix. By contrast, in this work we prioritize model-free approaches for possibly large spaces via the fusion of direct policy search and parameterization over a family of policies.

Appendix B. Supplementary materials of Section 3

B.1 A Monte Carlo Algorithm for solving (3.6)

Note that any algorithm that solves problem (3.6) will serve our purpose. Therefore, we provide a Monte Carlo method that alternates between stochastic primal and dual updates as an example, stated in Algorithm 1, in which the projection operator onto the set $\{z : \|z\|_\infty \leq \ell_F\}$ is denoted as

$\text{Proj}_{\ell_F}\{z\}$. For any z, $z' = \text{Proj}_{\ell_F}\{z\}$ is defined as

$$z'_i = \begin{cases} -\ell_F, & \text{if } z_i \in (-\infty, -\ell_F), \\ z_i, & \text{if } z_i \in [-\ell_F, \ell_F], \\ \ell_F, & \text{if } z_i \in (\ell_F, +\infty). \end{cases}$$

It is worth noting that we have we omit the term $\delta \nabla \tilde{V}(\theta; z)^\top x$ when computing the gradient w.r.t.

Algorithm 1 Monte Carlo Variational Policy Gradient Estimation

Require: a differentiable policy parametrization $\pi_{\boldsymbol{\theta}}$, stepsizes $\alpha_t, \beta_t > 0$, initial points $\boldsymbol{x} = 0$, $\mathbf{z} = 0$. A constant ℓ_F.

policy parameter $\boldsymbol{\theta} \in \mathbb{R}^d$

Generate episodes $\zeta_i = \{(s_k, a_k)\}$ from $i = 1, \cdots, n$ following $\pi_\theta(a|s)$

For $t = 0, 1, 2, ...$ until some stopping criterion is met:

Sample (s_k, a_k) from the data set

Update

$$\mathbf{z}^{t+1} \leftarrow \text{Proj}_{\ell_F}\left\{\mathbf{z}^t - \frac{\alpha_t}{1-\gamma}\mathbf{1}_{s_k, a_k} + \alpha_t \nabla F^*(\mathbf{z}^t)\right\} \tag{B.1}$$

$$\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t + \beta_t \left[\sum_{a \in \mathcal{A}} Q^{\pi_\theta}(s_k, a; \mathbf{z}^t) \cdot \nabla_\theta \pi_\theta(a|s_k) - \boldsymbol{x}^t\right] \tag{B.2}$$

Output: the last iterate $\boldsymbol{x}$

z in (B.1). Note that for the iterates x^t are all well bounded, then $\delta \nabla \tilde{V}(\theta; z^t)^\top x^t = \mathcal{O}(\delta)$, which is negligible when $\delta \to 0$.

B.2 Special cases of policy gradient computation

We give several examples of the policy gradient for special cases of the general utility in (3.1).

Linear utility The simplest, where $F(\lambda) = \langle \lambda, r \rangle$ [cf. (2.1)], we have $F^*(z) = 0$ if $z = c \cdot r$ for some scalar c and $F^*(z) = \infty$ otherwise. In this case $z^* = r$ and Theorem 3.1 recovers the known policy gradient theorem for the risk-neutral MDP (2.1), that is $\nabla_\theta R(\pi_\theta) = \nabla_\theta V(\theta; r)$.

Constrained MDPs By contrast, in Example 2.1, i.e., when a constraint $\mathbb{E}^\pi \left[\sum_{t=0}^\infty \gamma^t c(s_t, a_t)\right] \le C$ on the accumulation of costs $c(s_t, a_t)$ is present, and we may enforce it approximately with a log barrier by defining

$$R(\pi_\theta) = \langle r, \lambda(\theta) \rangle + \beta \log\left(C - \langle c, \lambda(\theta) \rangle\right) = V(\theta; r) + \beta \log\left(C - V(\theta; c)\right), \tag{B.3}$$

where β is a regularization parameter, in which case the policy gradient takes the form

$$\nabla R(\pi_\theta) = \nabla_\theta V(\theta; r) - \beta \frac{\nabla_\theta V(\theta; c)}{C - V(\theta; c)}.$$

Estimating the policy gradient R of constrained MDP consists of estimating two policy gradients $\nabla_\theta V(\theta; c)$ and $\nabla_\theta V(\theta; r)$ and accumulated reward $V(\theta; c)$.

Minimum eigenvalue For case (2.8), define $\Phi(\lambda^{\pi_\theta}) = \sum_{s,a} \lambda^{\pi_\theta}_{sa} \cdot \phi(s, a)\phi(s, a)^\top$. Then $\Phi(\lambda^{\pi_\theta})$ is symmetric and positive semidefinite, since $\lambda^{\pi_\theta} \ge 0$. By using Rayleigh principle, we have

$$R(\pi_\theta) = \sigma_{min}\left(\Phi(\lambda^{\pi_\theta})\right) = \min_{\|u\|=1} u^\top \Phi(\lambda^{\pi_\theta}) u = \min_{\|u\|=1} \sum_{s,a} \lambda^{\pi_\theta}_{sa} |\phi(s, a)^\top u|^2. \tag{B.4}$$

which is the minimum of a family of linear function in λ. Let $v^{(1)}, ..., v^{(k)}$ be a group of orthonormal bases of the eigenspace of $\Phi(\lambda^{\pi_\theta})$ corresponding to the minimum eigenvalue. Then define k vectors as $r^{(i)}(s,a) = |\phi(s,a)^\top v^{(i)}|^2, \forall s, a,\ i = 1, ..., k$. Then the Fréchet superdifferential of R at θ is

$$\hat{\partial}_\theta R(\pi_\theta) = \left\{ \nabla_\theta V(\theta; r) : r \in \text{conv}(r^{(1)}, ..., r^{(k)}) \right\},$$

where $\text{conv}(\cdot)$ denotes the convex hull of a group of vectors. When the multiplicity of the minimum eigenvalue is 1, then $R(\cdot)$ is differentiable at this point and $\hat{\partial}_\theta R(\cdot) = \{\nabla_\theta R(\cdot)\}$.

Entropy maximization For the entropy (2.7), its Fenchel dual takes the form

$$F^*(z) = -\sum_{sa} \exp\big\{ -\frac{z_{sa}}{1-\gamma} - 1 \big\}.$$

Learning to mimic a distribution For the KL divergence to a prior μ in (2.9), we have

$$F^*(z) = \begin{cases} -\sum_s \mu_s \exp\big\{ -\frac{z_{s1}}{1-\gamma} - 1 \big\} & \text{if} \quad z_{sa_1} = z_{sa_2} \quad \forall s \in \mathcal{S}, a_1, a_2 \in \mathcal{A}, \\ -\infty & \text{otherwise.} \end{cases}$$

Appendix C. Additional Details of Experiments

C.1 Details of Environment

OpenAI Frozen Lake is a finite-state action problem. The standard state consists of $\{S, F, H, G\}$, to which we add an additional state C which is visualized in Fig. 3(a). At each step, an agent selects an action $a \in \mathcal{A}$, which consists of one of four directions (up, down, left, right), which may be enumerated as $\{1, \ldots, 4\}$. The reward is null at all Frozen F spaces, the start S location, and the Holes H in the lake. If the agent enters a hole, the episode terminates, and hence null reward is accumulated for this trajectory. The only positive reward is 1 and may be obtained when reaching the goal state G. Our augmentation is that costly states C have been added, which incur reward -0.4 to represent, for instance, obstacles. We note that only for the cumulative return and its constrained variants, or other utilities that are defined in terms of the problem's inherent reward do these quantities matter. That is, for the entropy maximization problem, there is no reward associated with any state in the usual sense. The MDP transition model is unknown and defined by the OpenAI environment, a simulation oracle that provides state-action-reward triples.

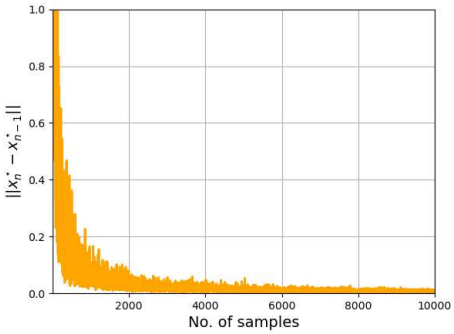
Throughout all experiments, for simplicity, we considered a softmax policy parameterization. For this parameterization, the policy takes the form $\pi_\theta(s \mid a) = e^{\theta_{sa}}/(\sum_{a'} e^{\theta_{sa'}})$ for $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. For the Frozen lake environment in this paper, we have $|\mathcal{S}| = 16$ and $|\mathcal{A}| = 4$.

C.2 Computing the True Policy Gradient

For comparison, we compute the true policy gradient by using a baseline approach based on the chain rule and a variant of REINFORCE Sutton et al. (2000): the second factor on the right-hand side of (3.3) is exactly computed using REINFORCE $\nabla_\theta \lambda_{sa}(\theta)$, whereas the first, $\frac{\partial F(\lambda(\theta))}{\partial \lambda_{sa}}$, is computed using an additional Monte Carlo rollout. We denote as x^* the result of this procedure and use it as ground truth. In Figure 4(a) we display the evolution of its norm difference $\|\hat{x}^\star_n - \hat{x}^\star_{n-1}\|$ as the sample size n increases. That it approaches null with the sample size implies that this brute force Monte Carlo variant of REINFORCE is convergent, and hence is a reasonable benchmark comparator.

C.3 Details about Maximum Entropy Exploration

For this problem instance, i.e., (2.7) from Example 2.2, we also consider the state space defined by Frozen Lake, but note that the reward as defined by the environment is now a moot point. This is



(a) Convergence of $x^\star$

Figure 4: Fig. 4(a) displays the convergence of a generalization of REINFORCE-based gradient estimator for (3.3) in terms of its difference $\|\hat{x}_n^\star - \hat{x}_{n-1}^\star\|$ as the number of processed trajectories n increases, which converges to null, certifying $\hat{x}_n^\star$ as a baseline.

because each state contributes positive entropy, with the exception of the holes in the lake, which terminate the episode. We visualize this setup at the bottom layer of Fig. 2(b). The lower middle layer visualizes the occupancy measure associated with a uniform policy. Moreover, the upper middle layer visualizes the "pseudo-reward" z for each point in the state space. This quantity is computed in terms of the Fenchel dual of the entropy – see Appendix B.2, and the occupancy measure associated with the output of Algorithm 1 at the end of training is visualized at the top layer. To obtain this result, we run it for 10^5 total episodes, and for each episode we evaluate the entropy using (2.7). We consider a constant step-size $\alpha = 0.01, \beta = 0.1$, and $\eta = 0.001$ throughout this experiment.

C.4 Details about the Constrained Markov Decision Process

In this subsection, we elaborate upon the implementation of Example 2.1, specifically, (2.6) and its approximation using a logarithmic barrier as detailed in (B.3). We consider the problem of navigating through the FrozenLake environment as shown in Fig. 3(a)(bottom): we seek to reach the goal state G (reward $= 1$) from the starting location S (reward $= 0$), navigating along F frozen spaces (reward $= 0$), while avoiding locations marked C (reward $= -0.2$) that denote costly states (obstacles) and H holes.

We consider two approaches to the problem: first, we focus on optimizing the standard expected cumulative return (2.1), whose associated state occupancy measure is given in Fig. 3(a)(middle); second, we consider imposing constraints to avoid costly states [cf. (2.6)] via a logarithmic barrier (B.3), whose resulting occupancy measure is depicted in Fig. 3(a)(top). Bluer/yellower colors denote higher/lower likelihoods, respectively. We observe that imposing constraints yields policies whose probability mass is concentrated away from constraints and instead along paths from the start to the goal. Thus, Algorithm 1 combined with a policy search scheme (4.1) may be used to solve CMDPs.

This trend is corroborated in Fig. 3, which depicts the reward 3(b) and cost 3(c) accumulation during test trajectories as a function of training index for Algorithm 1 for the cumulative return (2.1) as compared with a logarithmic barrier imposed to solve CMDP (2.6) for different penalty parameters β. We may observe that without imposing any constraint ($\beta = 0$ in red), one achieves the highest reward, but incurs the most costs, i.e., hits obstacles most often, a form of "reckless boldness." Larger β means higher penalty for the constraints, and hence $\beta = 4$ incurs lower cost and lower reward. We further added the curves for $\beta = 1$ and $\beta = 2$ for comparison.

For all results reported in Fig. 3, we run the algorithm for $10K$ total training steps in the form of episodes. For each episode, we run a number of evaluation (test) trajectories in order to determine

their merit, both in terms of reward and cost accumulation. Put more simply, we evaluate the performance averaged over a few test trajectories as a function of episode number and report its average over last 20 episodes to show the trend. This is to illuminate policy improvement in its various forms (reward/cost accumulation) during training. Moreover, the algorithm is run with constant step-size $\eta = 0.1$ throughout this experiment.

Appendix D. Proof of Theorem 3.1

Proof. First note that for any $z \in \mathbb{R}^{SA}, x \in \mathbb{R}^d$, we have

$$\begin{aligned} V(\theta; z) &= \langle z, \lambda(\theta) \rangle \,, \\ \nabla_\theta V(\theta; z)^\top x &= \langle z, \nabla_\theta \lambda(\theta) x \rangle \,, \end{aligned} \tag{D.1}$$

where $\nabla_\theta \lambda(\theta)$ is the $SA \times d$ Jacobian matrix, the first identity holds by definition, and the second holds by directly differetiating the first indentity and product it with x.

Consider the saddle point problem in (3.4) for fixed $0 < \delta < 1$. Let G be any constant such that $\|\nabla F(\lambda(\theta))\|_\infty < G$. Define

$$(x^*(\delta), z^*(\delta)) \quad := \operatorname{argmax}_x \operatorname{argmin}_{\|z\|_\infty \le G} \left\{ V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) - \tfrac{\delta}{2}\|x\|^2 \right\}. \tag{D.2}$$

Note in (D.2) we added the auxiliary constraint set $\{z : \|z\|_\infty \le G\}$, and later we will show that this constraint is inactive for all δ sufficiently small. We will also show that $(x^*(\delta), z^*(\delta))$ are bounded for all δ sufficiently small.

By the first-order stationarity condition, we have

$$x^*(\delta) = \nabla_\theta V(\theta; z^*(\delta)).$$

Note that $\nabla_\theta V(\theta; \cdot)$ is a linear function of z, thus there exists $B > 0$ such that $\|\nabla_\theta V(\theta; z)\| \le B$ for all $z \in \{\|z\|_\infty \le G\}$. And consequently $\|x^*(\delta)\| \le B$ for all $\delta > 0$.

For all $x \in \{\|x\| \le 2B\}$, we have

$$\lim_{\delta \to 0_+} \lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x = \lambda(\theta).$$

Therefore, there exists some small $\delta_0 > 0$, such that for all $\delta < \delta_0$, the vector $\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x$ belongs to the neighborhood on which F is differentiable and

$$\|\nabla F(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x)\|_\infty < G, \quad \forall \;\; x \in \{x : \|x\| \le 2B\}.$$

In this case, we consider the unconstrained solution, for $\|x\| \le 2B$, defined by

$$z^*(x; \delta) := \operatorname*{argmin}_z V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) = \nabla F\big(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x\big),$$

and observe that the unconstrained solution satisfies $\|z^*(x; \delta)\|_\infty < G$, and consequently the constraint $\|z\|_\infty \le G$ is not active. Therefore, for $\delta < \delta_0$, we can equivalently rewrite (D.2) as

$$\begin{aligned} x^*(\delta) \quad &:= \quad \operatorname*{argmax}_{\|x\| \le 2B} \min_z \left\{ V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2 \right\} \\ &= \quad \operatorname*{argmax}_{\|x\| \le 2B} F\left(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x\right) - \frac{\delta}{2}\|x\|^2 \,, \end{aligned} \tag{D.3}$$

Recall that we showed $\|x^*(\delta)\| \le B$, therefore the constraint $\|x\| \le 2B$ is also inactive and removable. Therefore $x^*(\delta)$ is equivalent to the unconstrained min-max solution, for all δ sufficiently small, and

Fenchel duality together with the first-order stationarity condition implies

$$\begin{aligned} x^*(\delta) &= \operatorname*{argmax}_x \inf_z \left\{ V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2 \right\} \\ &= \nabla_\theta \lambda(\theta)^\top \nabla F\big(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta)\big). \end{aligned}$$

By using the fact that ∇F is continuous at $\lambda(\theta)$ and $x^*(\delta)$ is bounded, by letting $\delta \to 0$ on both sides, we get

$$\begin{aligned} \lim_{\delta \to 0_+} x^*(\delta) &= \lim_{\delta \to 0_+} \nabla_\theta \lambda(\theta)^\top \nabla F\big(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta)\big) \\ &= \nabla_\theta \lambda(\theta)^\top \nabla F\big(\lambda(\theta)\big) \\ &= \nabla R(\theta)\,, \end{aligned}$$

where the last equality uses the chain rule.

□

Appendix E. Proof of Theorem 3.2

Proof. First, let us denote the expression in (3.4) for fixed $0 < \delta < 1$ as

$$(x^*(\delta), z^*(\delta)) = \operatorname*{argmax}_x \operatorname*{argmin}_{\|z\|_\infty \le \ell_F} V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2, \tag{E.1}$$

and its approximation with empirically estimated value functions and their gradients in (3.5) as

$$(\hat{x}(\delta), \hat{z}(\delta)) = \operatorname*{argmax}_x \operatorname*{argmin}_{\|z\|_\infty \le \ell_F} \tilde{V}(\theta; z) + \delta \nabla_\theta \tilde{V}(\theta; z)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2. \tag{E.2}$$

Then we decompose the entity $\mathbb{E}\left[\left\|\hat{\nabla}_\theta R(\pi_\theta) - \nabla_\theta R(\pi_\theta)\right\|^2\right]$ into three terms by adding and subtracting (i) $x^*(\delta)$ and (ii) $\hat{x}(\delta)$, which we then establish depends on the difference between (iii) $\hat{z}(\delta)$ and $z^*(\delta)$. Taken together with computing the limit of the right-hand side as $\delta \to 0$ we obtain the result. Each of these steps is analyzed independently, whose estimation errors are derived in the following lemma.

Lemma E.1. *Consider $(x^*(\delta), z^*(\delta))$ and $(\hat{x}(\delta), \hat{z}(\delta))$ as defined in* (E.1)-(E.2)*, respectively. Under the technical conditions stated in Theorem 3.2, their respective estimation errors satisfy:*

(i) $\left\|x^*(\delta) - \nabla_\theta R(\pi_\theta)\right\|^2 = \mathcal{O}(\delta^2)$.

(ii) $\mathbb{E}\left[\left\|x^*(\delta) - \hat{x}(\delta)\right\|^2\right] \le \frac{2C^2\|z^*(\delta)\|_\infty^2}{(1-\gamma)^4} \cdot \left(\frac{\gamma^{2K}}{(1-\gamma)^2} + \frac{1}{n}\right) + \frac{2C^2}{(1-\gamma)^4} \cdot \mathbb{E}\left[\|z^*(\delta) - \hat{z}(\delta)\|_\infty^2\right]$.

(iii) $\mathbb{E}\left[\|\hat{z}(\delta) - z^*(\delta)\|_\infty^2\right] \le \mathcal{O}\left(\frac{L_F^2}{n(1-\gamma)^2} + \frac{L_F^2 \ell_{F^*}^2}{n} + \frac{L_F^2\delta^2 + L_F\delta}{n}\right)$.

Combining the three steps and the fact that $\|z^*(\delta)\|_\infty \le \ell_F$ yields

$$\mathbb{E}\left[\|\hat{x}(\delta) - \nabla_\theta R(\theta)\|^2\right] \le \mathcal{O}\left(\frac{C^2(\ell_F^2 + L_F^2\ell_{F^*}^2)}{n(1-\gamma)^4} + \frac{C^2 L_F^2}{n(1-\gamma)^6}\right) + \mathcal{O}(\delta^2 + \delta/n + \gamma^K).$$

Let $\delta \to 0$, we get

$$\mathbb{E}\left[\left\|\hat{\nabla}_\theta R(\pi_\theta) - \nabla_\theta R(\pi_\theta)\right\|^2\right] \le \mathcal{O}\left(\frac{C^2(\ell_F^2 + L_F^2\ell_{F^*}^2)}{n(1-\gamma)^4} + \frac{C^2 L_F^2}{n(1-\gamma)^6}\right) + \mathcal{O}(\gamma^K).$$

Lemma E.1(i) - (iii) is proved in the next subsection. For the ease of notation, we will simply denote x^* and $\hat{x}$ instead of $x^*(\delta)$ and $\hat{x}(\delta)$. Similarly, we denote z^* and $\hat{z}$ instead of $z^*(\delta)$ and $\hat{z}(\delta)$. □

E.1 Preliminary Technicalities

Linearity property. The functions Q, V and $\nabla_\theta V$ are linear in the reward function. Namely, for any $\alpha, \alpha' \in \mathbb{R}$ and $r, r' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,

$$\alpha \nabla_\theta V(\theta; r) + \alpha' \nabla_\theta V(\theta; r') = \nabla_\theta V(\theta; \alpha r + \alpha' r').$$

Similar identities holds for $Q^{\pi_\theta}(s, a; \cdot)$ and $V(\theta; \cdot)$. For the stochastic estimators $\nabla_\theta \tilde{V}(\theta; r; \zeta)$, it is straightforward to check that the linearity property is still true.

Upperbounding Q and V. Given an arbitrary reward function r, the upper bounds of Q and V functions are

$$|Q^{\pi_\theta}(s, a; r)| \leq \frac{\|r\|_\infty}{1-\gamma} \qquad \text{and} \qquad |V(\theta; r)| \leq \frac{\|r\|_\infty}{1-\gamma}.$$

Uniform upperbounds for estimators. Given any sample path $\zeta = \{(s_k, a_k)\}_{k=0}^K$, the estimators $\tilde{V}(\theta; z; \zeta)$ and $\nabla_\theta \tilde{V}(\theta; z; \zeta)$ are upper bounded by

$$\tilde{V}(\theta; z; \zeta)| \leq \frac{\|z\|_\infty}{1-\gamma} \qquad \text{and} \qquad \|\nabla_\theta \tilde{V}(\theta; z; \zeta)\| \leq \frac{C\|z\|_\infty}{(1-\gamma)^2}. \tag{E.3}$$

Consequently, as the sample averages of $\tilde{V}(\theta; z; \zeta_i)$ and $\nabla_\theta \tilde{V}(\theta; z; \zeta_i)$, we also have

$$|\tilde{V}(\theta; z)| \leq \frac{\|z\|_\infty}{1-\gamma} \qquad \text{and} \qquad \|\nabla_\theta \tilde{V}(\theta; z)\| \leq \frac{C\|z\|_\infty}{(1-\gamma)^2} \tag{E.4}$$

for any set of sample paths $\{\zeta_i\}_{i=1}^n$.

Proof. For $\tilde{V}(\theta; z; \zeta)$, for any z,

$$|\tilde{V}(\theta; z; \zeta)| = \left| \sum_{k=0}^K \gamma^k \cdot z(s_k, a_k) \right| \leq \sum_{k=0}^K \gamma^k \|z\|_\infty \leq \frac{\|z\|_\infty}{1-\gamma}$$

For $\nabla_\theta \tilde{V}(\theta; z; \zeta)$, for any z,

$$\begin{aligned}
\|\nabla_\theta \tilde{V}(\theta; z; \zeta)\| &= \left\| \sum_{k=1}^K \sum_{a \in \mathcal{A}} \gamma^k \cdot Q(s_k, a; z) \nabla_\theta \pi_\theta(a|s_k) \right\| \\
&\leq \sum_{k=1}^K \gamma^k \cdot \left\| \sum_{a \in \mathcal{A}} Q(s_k, a; z) \nabla_\theta \pi_\theta(a|s_k) \right\| \\
&\leq \sum_{k=1}^K \gamma^k \cdot \max_{\|u\|_\infty \leq \frac{\|z\|_\infty}{1-\gamma}} \|\pi_\theta(\cdot|s_k) u\| \\
&\leq \frac{C\|z\|_\infty}{(1-\gamma)^2}.
\end{aligned}$$

□

E.2 Proof of Lemma E.1(i).

Consider the problem (E.1). First let us ignore the requirement that $\|z\|_\infty \le \ell_F$. For this series of unconstrained problem, Theorem 3.1 suggests that

$$\lim_{\delta \to 0_+} x^*(\delta) = \nabla_\theta R(\pi_\theta).$$

Consequently, $\lim_{\delta\to 0_+} \lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta) = \lambda(\theta)$. Because $\|\lambda(\theta)\|_1 = (1-\gamma)^{-1}$, $\exists \delta_0 > 0$ s.t. when $\delta < \delta_0$ we have

$$\|\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta)\|_1 \le \frac{2}{1-\gamma}.$$

According to condition (i) of this theorem, we have

$$\|\nabla F(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta))\|_\infty \le \ell_F.$$

It is worth noting that $z^*(\delta) = \nabla F(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta))$ is also the solution to the unconstrained version of (E.1). Therefore we have $\|z\|_\infty \le \ell_F$, so that we can add this to the constraint without changing the optimal solutions. By the intermediate result in the proof of Theorem 3.1, we have

$$x^*(\delta) = \nabla_\theta \lambda(\theta)^\top \nabla F\big(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta)\big).$$

Consequently, by the Lipschitz continuity of ∇F, we have

$$\begin{aligned}
\Big\| x^*(\delta) - \nabla_\theta R(\theta) \Big\|^2 &= \Big\| \nabla_\theta \lambda(\theta)^\top \nabla F\big(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta)\big) - \nabla_\theta \lambda(\theta)^\top \nabla F\big(\lambda(\theta)\big) \Big\|^2 \\
&\le \|\nabla_\theta \lambda(\theta)^\top\|_{\infty,2} \cdot \Big\| \nabla F\big(\lambda(\theta) + \delta \nabla_\theta \lambda(\theta) x^*(\delta)\big) - \nabla F\big(\lambda(\theta)\big) \Big\|_\infty^2 \\
&\le L_F \|\nabla_\theta \lambda(\theta)^\top\|_{\infty,2}^2 \cdot \Big\| \delta \nabla_\theta \lambda(\theta) x^*(\delta) \Big\|_1^2 \\
&= \mathcal{O}(\delta^2).
\end{aligned}$$

as stated in Lemma E.1(i). In the last step, we used the fact that $x^*(\delta)$ is bounded because $x^*(\delta) \to \nabla_\theta R(\pi_\theta)$. □

E.3 Proof of Lemma E.1(ii).

By the first order stationarity condition of the problems (E.1)-(E.2), we know

$$x^* = \nabla_\theta V(\theta; z^*) \qquad \text{and} \qquad \hat{x} = \nabla_\theta \tilde{V}(\theta; \hat{z}).$$

Consider the norm-difference between the preceding quantities:

$$\mathbb{E}\left[\big\| x^* - \hat{x} \big\|^2 \right] \le 2\mathbb{E}\left[\|\nabla_\theta V(\theta; z^*) - \nabla_\theta \tilde{V}(\theta; z^*)\|^2 \right] + 2\mathbb{E}\left[\|\nabla_\theta \tilde{V}(\theta; z^*) - \nabla_\theta \tilde{V}(\theta; \hat{z})\|^2 \right]. \tag{E.5}$$

To bound the term $\mathbb{E}\left[\|\nabla_\theta V(\theta; z^*) - \nabla_\theta \tilde{V}(\theta; z^*)\|^2 \right]$, recall the definition (3.5):

$$\nabla \tilde{V}(\theta; z) := \frac{1}{n} \sum_{i=1} \nabla_\theta V(\theta; z; \zeta_i) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \gamma^k Q(s_k^{(i)}, a; z) \nabla_\theta \pi_\theta(a | s_k^{(i)}).$$

Consider the first term on the right-hand side of (E.5). Add and subtract $\mathbb{E}\left[\nabla_\theta \tilde{V}(\theta; z^*) \right]$ and use the fact that $\mathbb{E}\left[\nabla_\theta \tilde{V}(\theta; z^*) \right] = \nabla_\theta V(\theta; z^*)$, i.e., the bias-variance decomposition identity, to write

$$\begin{aligned}
&\mathbb{E}\left[\Big\| \nabla_\theta V(\theta; z^*) - \nabla_\theta \tilde{V}(\theta; z^*) \Big\|^2 \right] \\
&= \Big\| \nabla_\theta V(\theta; z^*) - \mathbb{E}\left[\nabla_\theta \tilde{V}(\theta; z^*) \right] \Big\|^2 + \mathbb{E}\left[\Big\| \nabla_\theta \tilde{V}(\theta; z^*) - \mathbb{E}\left[\nabla_\theta \tilde{V}(\theta; z^*) \right] \Big\|^2 \right].
\end{aligned} \tag{E.6}$$

For the first (squared bias) term on the right-hand side of (E.6), denote $d^\pi_{\xi,K}(s) = (1-\gamma)\sum_{t=0}^K \gamma^t \mathbf{Prob}(s_t = s|\pi, s_0 \sim \xi)$. Then it is straightforward that $\sum_s |d^\pi_{\xi,K}(s) - d^\pi_\xi(s)| \leq \frac{\gamma^K}{1-\gamma}$. As a result, we know

$$\begin{aligned}
&\left\|\nabla_\theta V(\theta; z^*) - \mathbb{E}\left[\nabla_\theta \tilde{V}(\theta; z^*)\right]\right\|^2 \qquad \text{(E.7)}\\
&= \frac{1}{(1-\gamma)^2}\left\|\sum_s \left(d^\pi_\xi(s) - d^\pi_{\xi,K}(s)\right)\sum_a Q^{\pi_\theta}(s,a;z^*)\nabla_\theta \pi_\theta(a|s)\right\|^2\\
&= \frac{1}{(1-\gamma)^2}\left(\sum_s |d^\pi_\xi(s) - d^\pi_{\xi,K}(s)| \cdot \left\|\sum_a Q^{\pi_\theta}(s,a;z^*)\nabla_\theta \pi_\theta(a|s)\right\|\right)^2\\
&= \frac{1}{(1-\gamma)^2}\left(\sum_s |d^\pi_\xi(s) - d^\pi_{\xi,K}(s)| \cdot \left\|\sum_a Q^{\pi_\theta}(s,a;z^*)\nabla_\theta \pi_\theta(a|s)\right\|\right)^2\\
&\leq \frac{1}{(1-\gamma)^2}\left(\sum_s |d^\pi_\xi(s) - d^\pi_{\xi,K}(s)| \cdot \max_{\|u\|_\infty \leq \frac{\|z^*\|_\infty}{1-\gamma}} \|\nabla_\theta \pi(\cdot|s)u\|\right)^2\\
&\leq \frac{\|\nabla_\theta \pi(\cdot|s)\|^2_{\infty,2}\cdot \|z^*\|^2_\infty}{(1-\gamma)^4}\left(\sum_s |d^\pi_\xi(s) - d^\pi_{\xi,K}(s)|\right)^2\\
&\leq \frac{C^2\|z^*\|^2_\infty}{(1-\gamma)^6}\gamma^{2K}.
\end{aligned}$$

Next, we consider the second (variance) term on the right-hand side of (E.6). By substituting (3.5) in for $\nabla_\theta \tilde{V}(\theta; z^*)$ to rewrite it in terms of trajectories ζ_i, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla_\theta \tilde{V}(\theta; z^*) - \mathbb{E}\left[\nabla_\theta \tilde{V}(\theta; z^*)\right]\right\|^2\right] &= \frac{1}{n}\mathbb{E}\left[\left\|\nabla_\theta \tilde{V}(\theta; z^*; \zeta_i) - \mathbb{E}\left[\nabla_\theta \tilde{V}(\theta; z^*; \zeta_i)\right]\right\|^2\right]\\
&\leq \frac{1}{n}\mathbb{E}\left[\left\|\nabla_\theta \tilde{V}(\theta; z^*; \zeta_i)\right\|^2\right]\\
&\leq \frac{C^2\|z^*\|^2_\infty}{n(1-\gamma)^4}.
\end{aligned}$$

The first inequality comes from crudely upper-bounding the bias by the estimator itself. The last equality uses (E.3).

Now, returning focus to the second term in the bound (E.5), by the linearity of the stochastic estimators with respect to the differential and (E.4), we have

$$\left\|\nabla_\theta \tilde{V}(\theta; z^*) - \nabla_\theta \tilde{V}(\theta; \hat{z})\right\|^2 = \left\|\nabla_\theta \tilde{V}(\theta; z^* - \hat{z})\right\|^2 \leq \frac{C^2\|z^* - \hat{z}\|^2_\infty}{(1-\gamma)^4}.$$

Taking the expectation after squaring both sides yields

$$\mathbb{E}\left[\left\|\nabla_\theta \tilde{V}(\theta; z^*) - \nabla_\theta \tilde{V}(\theta; \hat{z})\right\|^2\right] \leq \frac{C^2}{(1-\gamma)^4}\mathbb{E}\left[\|z^* - \hat{z}\|^2_\infty\right]. \qquad \text{(E.8)}$$

Combining inequalities (E.5), (E.6), (E.7), (E.8), (E.8) yields

$$\mathbb{E}\left[\|x^* - \hat{x}\|^2\right] \leq \frac{2C^2\|z^*\|^2_\infty}{(1-\gamma)^4}\cdot\left(\frac{\gamma^{2K}}{(1-\gamma)^2} + \frac{1}{n}\right) + \frac{2C^2}{(1-\gamma)^4}\cdot\mathbb{E}\left[\|z^* - \hat{z}\|^2_\infty\right].$$

which is as stated in Lemma E.1(ii). □

E.4 Proof of Lemma E.1(iii).

In this section we will apply the generalization bound for stochastic saddle points from Zhang et al. (2020b) to bound the term $\mathbb{E}[\|\hat{z} - z^*\|_\infty^2]$. To achieve this, we need a compact feasible region for x. Note that for problems (E.1) and (E.2), the solutions x^* and $\hat{x}$ has the form

$$x^* = \nabla_\theta V(\theta; z^*) \qquad \text{and} \qquad \hat{x} = \nabla_\theta \tilde{V}(\theta; \hat{z}).$$

Due to (E.4) and the constraint that $\|z\|_\infty \le \ell_F$, we have $\|x^*\| \le \frac{C\|z^*\|_\infty}{(1-\gamma)^2} \le \frac{C\ell_F}{(1-\gamma)^2}$ and thus $\|\hat{x}\| \le \frac{C\ell_F}{(1-\gamma)^2}$ with probability 1. Therefore, adding a constraint that $\|x\| \le \frac{C\ell_F}{(1-\gamma)^2}$ will not change the solutions of problems (E.1) and (E.2). Formally speaking, we will then apply the theory of Zhang et al. (2020b) to the following pair of constrained problems:

$$(x^*, z^*) = \operatorname*{argmax}_{x\in\mathcal{X}} \operatorname*{argmin}_{z\in\mathcal{Z}} V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2, \tag{E.9}$$

and

$$(\hat{x}, \hat{z}) = \operatorname*{argmax}_{x\in\mathcal{X}} \operatorname*{argmin}_{z\in\mathcal{Z}} \tilde{V}(\theta; z) + \delta \nabla_\theta \tilde{V}(\theta; z)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2. \tag{E.10}$$

with $\mathcal{X} = \{x : \|x\| \le \frac{C\ell_F}{(1-\gamma)^2}\}$ and $\mathcal{Z} = \{z : \|z\|_\infty \le \ell_F\}$. The problems (E.1) and (E.9) share the same solution, and problems (E.2) and (E.10) share the same solution.

Finally, similar to the proof of (E.7), for any $x \in \mathcal{X}$ and $z \in \mathcal{Z}$

$$V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - \mathbb{E}\left[\tilde{V}(\theta; z; \zeta_i) + \delta \nabla_\theta \tilde{V}(\theta; z; \zeta_i)^\top x\right] = \mathcal{O}\left(\frac{\gamma^K}{1-\gamma}\right).$$

For the simplicity of discussion, let us assume that K is large enough so that we can ignore the $\mathcal{O}\left(\frac{\gamma^K}{1-\gamma}\right)$ bias. Therefore problem (E.10) can be viewed as an empirical version of the problem (E.9) with negligible bias. To apply the theory of Zhang et al. (2020b), define

$$\Psi_\zeta(x, z) := \tilde{V}(\theta; z; \zeta) + \delta \nabla_\theta \tilde{V}(\theta; z; \zeta)^\top x - F^*(z) - \frac{\delta}{2}\|x\|^2.$$

Then for any sample path ζ, Ψ_ζ satisfies the following set of properties:

- $\Psi_\zeta(\cdot, z)$ is μ_x-strongly concave under L_2 norm. And $\Psi_\zeta(x, \cdot)$ is μ_z-strongly convex under the L_∞ norm. In other words, for $\forall x, x' \in \mathcal{X}$ and $z, z' \in \mathcal{Z}$,

$$\begin{cases} \Psi_\zeta(x', z) \ge \Psi_\zeta(x, z) + \langle u, x' - x\rangle + \frac{\mu_x}{2}\|x' - x\|^2, & u \in \partial_x \Psi_\zeta(x, z), \\ \Psi_\zeta(x, z') \le \Psi_\zeta(x, z) + \langle v, z' - z\rangle - \frac{\mu_z}{2}\|z' - z\|_\infty^2, & v \in \partial_z \Psi_\zeta(x, z). \end{cases}$$

  In our case, it is clear that $\mu_x = \delta$. Due to Theorem 3 of Kakade et al. (2012), $\mu_z = L_F^{-1}$.

- The feasible regions $\mathcal{X}$ and $\mathcal{Z}$ are compact convex sets. For every ζ, there exist constants $\ell_x(\xi, z)$ and $\ell_z(\xi, x)$ s.t.

$$\begin{cases} |\Psi_\zeta(x', z) - \Psi_\zeta(x, z)| \le \ell_x(\zeta, z)\|x' - x\|, & \forall x, x' \in \mathcal{X} \text{ and } y \in \mathcal{Y}, \\ |\Psi_\zeta(x, z') - \Psi_\zeta(x, z)| \le \ell_z(\zeta, x)\|z' - z\|_\infty, & \forall z, z' \in \mathcal{Z} \text{ and } x \in \mathcal{X}. \end{cases}$$

  In our case, we gave $\ell_z(\zeta, x) = \sup\{\|u\|_1 : z \in \mathcal{Z}, u \in \partial_z \Psi_\zeta(x, z)\} = \frac{1}{1-\gamma} + \ell_{F^*} + \mathcal{O}(\delta)$ and $\ell_x(\zeta, z) = \sup_{x\in\mathcal{X}} \|\nabla_x \Psi_\zeta(x, z)\| = \mathcal{O}(\delta)$. Consequently,

$$\begin{cases} (\ell_x^w)^2 := \sup_{z\in\mathcal{Z}} \mathbb{E}\big[\ell_x^2(\zeta, z)\big] = \mathcal{O}(\delta^2), \\ (\ell_z^w)^2 := \sup_{x\in\mathcal{X}} \mathbb{E}\big[\ell_z^2(\zeta, x)\big] = \mathcal{O}(\ell_{F^*}^2 + \frac{1}{(1-\gamma)^2} + \delta^2). \end{cases}$$

With the above two properties, Theorem 1 of Zhang et al. (2020b) indicates that

$$\frac{\mu_z}{2}\mathbb{E}\left[\|\hat{z}-z^*\|_\infty^2\right] \leq \frac{2\sqrt{2}}{n}\cdot\left(\frac{(\ell_x^w)^2}{\mu_x}+\frac{(\ell_z^w)^2}{\mu_z}\right).$$

With the detailed parameters substituted in the above inequality, we have

$$\mathbb{E}\left[\|\hat{z}-z^*\|_\infty^2\right] \leq \mathcal{O}\left(\frac{L_F^2}{n(1-\gamma)^2}+\frac{L_F^2\ell_{F^*}^2}{n}+\frac{L_F^2\delta^2+L_F\delta}{n}\right)$$

as stated in Lemma E.1(iii). □

Appendix F. Proof of Theorem 4.2

Proof. Let θ^* be a first-order stationary solution of (3.1). When F is concave and locally Lipschitz continuous in a neighbourhood containing $\lambda(\Theta)$, we can compute the Fréchet superdifferential of $F\circ\lambda$ at θ^* by the chain rule, see Drusvyatskiy and Paquette (2019). That is

$$\hat{\partial}(F\circ\lambda)(\theta^*) = \left[\nabla_\theta\lambda(\theta^*)\right]^\top \partial F(\lambda^*)$$

where $\partial F(\lambda^*)$ denotes the set of supergradients of the concave function F at λ^*. Then there exists $w^*\in\partial F(\lambda^*)\in\mathbb{R}^{SA}$ such that $u^* := [\nabla_\theta\lambda(\theta^*)]^\top w^* \in \hat{\partial}(F\circ\lambda)(\theta^*)$ as in (4.2). It follows from (4.2) that

$$\langle w^*, \nabla_\theta\lambda(\theta^*)(\theta-\theta^*)\rangle \leq 0, \quad \text{for} \quad \forall\theta\in\Theta. \tag{F.1}$$

For any $\lambda\in\lambda(\Theta)$, we let $\theta := g(\lambda)$ such that $\lambda=\lambda(\theta)$. Therefore, by adding and subtracting $\nabla_\theta\lambda(\theta^*)\theta$ inside the inner product we have

$$\begin{aligned}\langle w^*, \lambda-\lambda^*\rangle &= \langle w^*, \lambda(\theta)-\lambda(\theta^*)\rangle \\ &= \langle w^*, \nabla_\theta\lambda(\theta^*)(\theta-\theta^*)\rangle + \langle w^*, \lambda(\theta)-\lambda(\theta^*)-\nabla_\theta\lambda(\theta^*)(\theta-\theta^*)\rangle \\ &\leq 0 + \|w^*\|\|\lambda(\theta)-\lambda(\theta^*)-\nabla_\theta\lambda(\theta^*)(\theta-\theta^*)\|.\end{aligned} \tag{F.2}$$

where in the last inequality we group terms and apply Cauchy-Schwartz. Note that the Jacobian matrix $\nabla_\theta\lambda(\theta)$ is Lipschitz continuous. Denote the Lipschitz constant by L_λ, i.e., $\|\nabla_\theta\lambda(\theta)-\nabla_\theta\lambda(\theta')\|\leq L_\lambda\|\theta-\theta'\|$ for all $\theta,\theta'\in\Theta$. Then,

$$\|\lambda(\theta)-\lambda(\theta^*)-\nabla_\theta\lambda(\theta^*)(\theta-\theta^*)\| \leq \frac{L_\lambda}{2}\|\theta-\theta^*\|^2.$$

By Assumption 4.1, we know

$$\|\theta-\theta^*\|^2 = \|g(\lambda)-g(\lambda^*)\|^2 \leq \ell_\theta^2\|\lambda-\lambda^*\|^2.$$

Substituting the above inequalities into (F.2) yields

$$\langle w^*, \lambda-\lambda^*\rangle \leq \frac{L_\lambda\ell_\theta^2}{2}\|w^*\|\|\lambda-\lambda^*\|^2 \qquad \forall\lambda\in\lambda(\Theta). \tag{F.3}$$

Note that (F.3) holds for arbitrary $\lambda\in\lambda(\Theta)$. Therefore, since $\lambda(\Theta)$ is assumed to be convex (Assumption 4.1(i)), we can also substitute λ with $(1-\alpha)\lambda^*+\alpha\lambda, \alpha\in[0,1]$ into the above equation, which yields

$$\alpha\langle w^*, \lambda-\lambda^*\rangle \leq \frac{L_\lambda\ell_\theta^2\alpha^2}{2}\|w^*\|\|\lambda-\lambda^*\|^2 \qquad \forall\lambda\in\mathcal{L}, \forall\alpha\in[0,1].$$

Divide both sides of the preceding expression by α and take $\alpha \to 0+$ gives

$$\langle w^*, \lambda - \lambda^* \rangle \leq \lim_{\alpha \to 0+} \frac{L_\lambda \ell_\theta^2 \alpha}{2} \|w^*\| \|\lambda - \lambda^*\|^2 = 0 \qquad \forall \lambda \in \lambda(\Theta).$$

Recall that the following problem is concave in λ:

$$\max_{\lambda} \quad F(\lambda) \qquad \text{s.t.} \qquad \lambda \in \lambda(\Theta),$$

therefore we conclude that λ^* is the global optimal solution. Then $\theta^* = g(\lambda^*)$ is the globally optimal solution of the nonconvex optimization problem (3.1). □

Appendix G. Proof of Theorem 4.4

G.1 Proof of sublinear convergence

Proof. First, the Lipschitz continuity in Assumption 4.3 indicates that

$$\left| F(\lambda(\theta)) - F(\lambda(\theta^k)) - \langle \nabla_\theta F(\lambda(\theta^k)), \theta - \theta^k \rangle \right| \leq \frac{L}{2} \|\theta - \theta^k\|^2.$$

Consequently, for any $\theta \in \Theta$ we have the ascent property:

$$F(\lambda(\theta)) \geq F(\lambda(\theta^k)) + \langle \nabla_\theta F(\lambda(\theta^k)), \theta - \theta^k \rangle - \frac{L}{2} \|\theta - \theta^k\|^2 \geq F(\lambda(\theta)) - L\|\theta - \theta^k\|^2. \tag{G.1}$$

The optimality condition in the policy update rule (4.1) then yields

$$\begin{aligned} F(\lambda(\theta^{k+1})) &\geq F(\lambda(\theta^k)) + \langle \nabla_\theta F(\lambda(\theta^k)), \theta^{k+1} - \theta^k \rangle - \frac{L}{2} \|\theta^{k+1} - \theta^k\|^2 \\ &= \max_{\theta \in \Theta} F(\lambda(\theta^k)) + \langle \nabla_\theta F(\lambda(\theta^k)), \theta - \theta^k \rangle - \frac{L}{2} \|\theta - \theta^k\|^2 \\ &\overset{(a)}{\geq} \max_{\theta \in \Theta} F(\lambda(\theta)) - L\|\theta - \theta^k\|^2 \\ &\overset{(b)}{\geq} \max_{\alpha \in [0,1]} \left\{ F(\lambda(\theta_\alpha)) - L\|\theta_\alpha - \theta^k\|^2 : \theta_\alpha = g(\alpha \lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) \right\}. \end{aligned} \tag{G.2}$$

Here, step (a) is due to (G.1) and step (b) uses the convexity of $\lambda(\Theta)$. Now, we proceed to analyze the right-hand side of (G.2). First, by the concavity of F and the fact that $\lambda \circ g = id$, we have

$$F(\lambda(\theta_\alpha)) = F(\alpha \lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) \geq \alpha F(\lambda(\theta^*)) + (1-\alpha) F(\lambda(\theta^k)).$$

Moreover, by the Lipschitz continuity assumption of g, we have

$$\begin{aligned} \|\theta_\alpha - \theta^k\|^2 &= \|g(\alpha \lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) - g(\lambda(\theta^k))\|^2 \\ &\leq \alpha^2 \ell_\theta^2 \|\lambda(\theta^*) - \lambda(\theta^k)\|^2 \\ &\leq \alpha^2 \ell_\theta^2 D_\lambda^2. \end{aligned} \tag{G.3}$$

Substituting the above two inequalities into the right-hand side of (G.2), we get

$$\begin{aligned} &F(\lambda(\theta^*)) - F(\lambda(\theta^{k+1})) \\ &\quad \leq \min_{\alpha \in [0,1]} \left\{ F(\lambda(\theta^*)) - F(\lambda(\theta_\alpha)) + L\|\theta_\alpha - \theta^k\|^2 : \theta_\alpha = g(\alpha \lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) \right\} \\ &\quad \leq \min_{\alpha \in [0,1]} (1-\alpha)\big(F(\lambda(\theta^*)) - F(\lambda(\theta^k))\big) + \alpha^2 L \ell_\theta^2 D_\lambda^2 \,. \end{aligned} \tag{G.4}$$

Let $\alpha_k = \frac{F(\Lambda(\pi^*))-F(\Lambda(\pi^k))}{2L\ell_\theta^2 D_\lambda^2} \geq 0$, which is the minimizer of the RHS of (G.4) as long as it satisfies $\alpha_k \leq 1$.

Now, we claim the following: If $\alpha_k \geq 1$ then $\alpha_{k+1} < 1$. Further, if $\alpha_k < 1$ then $\alpha_{k+1} \leq \alpha_k$. The two claims together mean that $(\alpha_k)_k$ is decreasing and all α_k are in $[0, 1)$ except perhaps α_0.

To prove the first of the two claims, assume $\alpha_k \geq 1$. This implies that $F(\Lambda(\pi^*)) - F(\Lambda(\pi^k)) \geq 2L\ell_\theta^2 D_\lambda^2$. Hence, choosing $\alpha = 1$ in (G.4), we get

$$F(\lambda(\theta^*)) - F(\lambda(\theta^k)) \leq L\ell_\theta^2 D_\lambda^2$$

which implies that $\alpha_{k+1} \leq 1/2 < 1$.

To prove the second claim, we plug α_k into (G.4) to get

$$F(\lambda(\theta^*)) - F(\lambda(\theta^{k+1})) \leq \left(1 - \frac{F(\lambda(\theta^*)) - F(\lambda(\theta^k))}{4L\ell_\theta^2 D_\lambda^2}\right)(F(\lambda(\theta^*)) - F(\lambda(\theta^k))),$$

which shows that $\alpha_{k+1} \leq \alpha_k$ as required.

Now, by our preceding discussion, for $k = 1, 2, \ldots$ the previous recursion holds. Using the definition of α_k, we rewrite this in the equivalent form

$$\frac{\alpha_{k+1}}{2} \leq \left(1 - \frac{\alpha_k}{2}\right) \cdot \frac{\alpha_k}{2}.$$

By rearranging the preceding expressions and algebraic manipulations, we obtain

$$\frac{2}{\alpha_{k+1}} \geq \frac{1}{\left(1 - \frac{\alpha_k}{2}\right) \cdot \frac{\alpha_k}{2}} = \frac{2}{\alpha_k} + \frac{1}{1 - \frac{\alpha_k}{2}} \geq \frac{2}{\alpha_k} + 1.$$

For simplicity assume that $\alpha_0 < 1$ also holds. Then, $\frac{2}{\alpha_k} \geq \frac{2}{\alpha_0} + k$, and consequenlty

$$F(\lambda(\theta^*)) - F(\lambda(\theta^k)) \leq \frac{F(\lambda(\theta^*)) - F(\lambda(\theta^0))}{1 + \frac{F(\lambda(\theta^*))-F(\lambda(\theta^0))}{4L\ell_\theta^2 D_\lambda^2} \cdot k} \leq \frac{4L\ell_\theta^2 D_\lambda^2}{k}.$$

A similar analysis holds when $\alpha_0 > 1$. Combining these two gives that $F(\lambda(\pi^*)) - F(\lambda(\pi^k)) \leq \frac{4L\ell_\theta^2 D_\lambda^2}{k+1}$ no matter the value of α_0, which proves the result. □

G.2 Proof of exponential convergence

When the strong concavity of F is available, we further provide the exponential convergence result.

Proof. We start from (G.2) whose proof requires no assumption on strong concavity of F, which is

$$F(\lambda(\theta^{k+1})) \geq \max_{\alpha\in[0,1]} \left\{F(\lambda(\theta_\alpha)) - L\|\theta_\alpha - \theta^k\|^2 : \theta_\alpha = g(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k))\right\}. \tag{G.5}$$

By the μ-strong concavity of F, we have

$$F(\lambda(\theta_\alpha)) = F(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) \geq \alpha F(\lambda(\theta^*)) + (1-\alpha)F(\lambda(\theta^k)) + \frac{\mu}{2}\alpha(1-\alpha)\|\lambda(\theta^*) - \lambda(\theta^k)\|^2.$$

By the Lipschitz continuity of g, we know that

$$\|\theta_\alpha - \theta^k\| = \|g(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) - g(\lambda(\theta^k))\| \leq \alpha\ell_\theta\|\lambda(\theta^*) - \lambda(\theta^k)\|$$

Substituting the above two inequalities into the right-hand side of (G.5), we get

$$\begin{aligned} & F(\lambda(\theta^*)) - F(\lambda(\theta^{k+1})) \\ \leq\ & \min_{\alpha\in[0,1]} \left\{F(\lambda(\theta^*)) - F(\lambda(\theta_\alpha)) + L\|\theta_\alpha - \theta^k\|^2 : \theta_\alpha = g(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k))\right\} \\ \leq\ & \min_{\alpha\in[0,1]} (1-\alpha)\big(F(\lambda(\theta^*)) - F(\lambda(\theta^k))\big) - \alpha\left(\frac{1-\alpha}{2}\mu - L\ell_\theta^2\alpha\right)\|\lambda(\theta^*) - \lambda(\theta^k)\|^2 \end{aligned} \tag{G.6}$$

Suppose we choose $\bar{\alpha} = \frac{1}{1+L\ell_\theta^2/\mu} < 1$ such that $\left(\frac{1-\bar{\alpha}}{2}\mu - L\ell_\theta^2\bar{\alpha}\right) = 0$. Then we have a contraction with modulus $1-\bar{\alpha}$ as

$$F(\lambda(\theta^*)) - F(\lambda(\theta^{k+1})) \leq (1-\bar{\alpha})F(\lambda(\theta^*)) - F(\lambda(\theta^k)).$$

Consequently, for any $k \geq 1$, we have

$$F(\lambda(\theta^*)) - F(\lambda(\theta^k)) \leq (1-\bar{\alpha})^k \left(F(\lambda(\theta^*)) - F(\lambda(\theta^0))\right).$$

which can be translated into iteration complexity by fixing ϵ and initialization θ^0, and solving for the minimal k such that $F(\lambda(\theta^*)) - F(\lambda(\theta^k)) \leq \epsilon$. Doing so is an algebraic exercise which results in

$$\mathcal{O}\left(\frac{1}{\bar{\alpha}}\log\left(\frac{F(\lambda(\theta^*)) - F(\lambda(\theta^0))}{\epsilon}\right)\right) = \mathcal{O}\left(\frac{L\ell_\theta^2}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$$

□

Appendix H. Validating Assumption 4.1 for tabular policy case

For the tabular policy case, the following Proposition holds true and hence the Assumption 4.1 is satisfied in this case.

Proposition H.1. *Suppose $\xi_s > 0$ for $\forall s \in \mathcal{S}$. Then the following hold:*

(i). The mappings Π and Λ form a pair of bijections between the convex sets $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{L}$;

(ii). $\exists L_\lambda > 0$ s.t. $\|\nabla\Lambda(\pi) - \nabla\Lambda(\pi')\| \leq L_\lambda\|\pi - \pi'\|, \forall \pi, \pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}}$;

(iii). For all $\lambda, \lambda' \in \mathcal{L}$, we have

$$\|\Pi(\lambda) - \Pi(\lambda')\|^2 \leq 2\sum_s \Big(\sum_a(\lambda'_{sa} - \lambda_{sa})^2 + (\sum_a \lambda'_{sa} - \lambda_{sa})^2\Big)/\big(\sum_a \lambda_{sa}\big)^2.$$

Consequently, $\|\Pi(\lambda) - \Pi(\lambda')\| \leq \frac{2}{\min_s \xi_s}\|\lambda - \lambda'\|_1$

Proof.

Proof of (i): The equations $\Pi \circ \Lambda = \mathrm{id}_{\mathcal{L}}$ and $\Lambda \circ \Pi = \mathrm{id}_{\Delta_{\mathcal{A}}^{\mathcal{S}}}$ are standard. See, e.g., Altman (1999) or Appendix A of Zhang et al. (2020a).

Proof of (ii): For the existence of the L_λ-Lipschitz constant of the gradient $\nabla\Lambda$, note that the t-th term of the infinite sum

$$\Lambda_{sa}(\pi) = \sum_{t=0}^{\infty}\gamma^t \cdot \mathbb{P}\Big(s_t = s, a_t = a \;\Big|\; \pi, s_0 \sim \xi\Big)$$

is a $(t+1)$-th order polynomial. Therefore, $\Lambda_{sa}(\pi)$ can actually be defined for any π even if $\pi \notin \Delta_{\mathcal{A}}^{\mathcal{S}}$, as long as this infinite series of polynomial of π converges absolutely. Note that for $\forall \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, since $0 \leq \mathbb{P}\big(s_t = s, a_t = a \;\big|\; \pi, s_0 \sim \xi\big) \leq 1$ this infinite series is absolutely convergent. Because we have $0 < \gamma < 1$, even if we slightly purterb the π within a neighbourhood of it (not necessarily in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ after purterbation), the infinite series is still absolutely convergent. This indicates that Λ_{sa} is infinitely continuously differentiable in an open neighbourhood containing $\Delta_{\mathcal{A}}^{\mathcal{S}}$, then due to the compactness of $\Delta_{\mathcal{A}}^{\mathcal{S}}$, we are able to argue that there exists a L_λ s.t. $\nabla\Lambda$ is L_λ-Lipschitz continuous within $\Delta_{\mathcal{A}}^{\mathcal{S}}$.

Proof of (iii): Now, we provide the calculation of the Lipschitz constant of Π. For the ease of notation, let us define $\mu_s = \sum_{a\in\mathcal{A}} \lambda_{sa}$ and $\mu'_s = \sum_{a\in\mathcal{A}} \lambda'_{sa}$. Then for $\forall \lambda, \lambda' \in \mathcal{L}$ and $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$, it holds that

$$\begin{aligned}\Pi_{sa}(\lambda) - \Pi_{sa}(\lambda') &= \frac{\lambda_{sa}}{\mu_s} - \frac{\lambda'_{sa}}{\mu'_s} \\ &= \left(\frac{\lambda_{sa}}{\mu_s} - \frac{\lambda'_{sa}}{\mu_s}\right) + \left(\frac{\lambda'_{sa}}{\mu_s} - \frac{\lambda'_{sa}}{\mu'_s}\right) \\ &= \frac{1}{\mu_s}(\lambda_{sa} - \lambda'_{sa}) + \frac{\mu'_s - \mu_s}{\mu_s \mu'_s}\lambda'_{sa}.\end{aligned}$$

Consequently, we can compute the norm difference of the preceding expression and apply the triangle inequality:

$$\begin{aligned}\|\Pi(\lambda) - \Pi(\lambda')\|^2 &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left(\Pi_{sa}(\lambda) - \Pi_{sa}(\lambda')\right)^2 \qquad \text{(H.1)} \\ &\le 2\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{1}{\mu_s^2}(\lambda_{sa} - \lambda'_{sa})^2 + 2\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{(\mu'_s - \mu_s)^2}{\mu_s^2(\mu'_s)^2}(\lambda'_{sa})^2 \\ &\le 2\sum_{s\in\mathcal{S}} \frac{1}{\mu_s^2}\left(\sum_{a\in\mathcal{A}}(\lambda_{sa} - \lambda'_{sa})^2 + (\mu'_s - \mu_s)^2\right),\end{aligned}$$

where the last inequality follows because $\|x\|_2^2 \le \|x\|_1^2$ holds for any vector x (here, $\|\cdot\|_p$ denotes the p-norm). Finally, note that $\mu_s \ge \xi_s > 0$, we have

$$\begin{aligned}\|\Pi(\lambda) - \Pi(\lambda')\|^2 &\le 2\sum_{s\in\mathcal{S}} \frac{1}{\mu_s^2}\left(\sum_{a\in\mathcal{A}}(\lambda_{sa} - \lambda'_{sa})^2 + (\mu'_s - \mu_s)^2\right) \\ &\le \frac{2}{\min_s \xi_s^2}\sum_{s\in\mathcal{S}}\left(\sum_{a\in\mathcal{A}}(\lambda_{sa} - \lambda'_{sa})^2 + \big(\sum_{a\in\mathcal{A}}|\lambda_{sa} - \lambda_{sa'}|\big)^2\right) \\ &\le \frac{4}{\min_s \xi_s^2}\|\lambda - \lambda'\|_1^2\end{aligned}$$

Take the square root of both sides completes the proof. □

Appendix I. Proof of Theorem 4.5

Proof. To prove this theorem, it suffices to observe that (G.2) is still true with $\theta = \pi$, $\lambda(\theta) = \Lambda(\pi)$ and $g(\lambda) = \Pi(\lambda)$. Therefore, (G.2) can be translated as

$$F(\Lambda(\pi^{k+1})) \ge \max_{\alpha\in[0,1]} \left\{F(\Lambda(\pi_\alpha)) - L\|\pi_\alpha - \pi^k\|^2 : \pi_\alpha = \Pi(\alpha\Lambda(\pi^*) + (1-\alpha)\Lambda(\pi^k))\right\}. \qquad \text{(I.1)}$$

By the concavity of F and the fact that $\Lambda \circ \Pi = id$, we have

$$F(\Lambda(\pi_\alpha)) = F(\alpha\Lambda(\pi^*) + (1-\alpha)\Lambda(\pi^k)) \ge \alpha F(\Lambda(\pi^*)) + (1-\alpha)F(\Lambda(\pi^k)). \qquad \text{(I.2)}$$

For the inequality (G.3), we can derive a tighter bound by the following argument:

$$
\begin{aligned}
\|\pi_\alpha - \pi^k\|^2 &= \|\Pi(\alpha\Lambda(\pi^*) + (1-\alpha)\Lambda(\pi^k)) - \Pi(\Lambda(\pi^k))\|^2 \qquad \text{(I.3)}\\
&\le \alpha^2 \sum_s \frac{1}{\big(\sum_a \lambda_{sa}\big)^2} \left(\sum_a (\lambda^*_{sa} - \lambda_{sa})^2 + \big(\sum_a \lambda^*_{sa} - \sum_a \lambda_{sa}\big)^2 \right) \\
&\le 4\alpha^2 \sum_s \frac{1}{\big(\sum_a \lambda_{sa}\big)^2} \left(\big(\sum_a \lambda^*_{sa}\big)^2 + \big(\sum_a \lambda_{sa}\big)^2 \right) \\
&= 4\alpha^2 \sum_s \frac{\big(d_\xi^{\pi^*}(s)\big)^2 + \big(d_\xi^{\pi^k}(s)\big)^2}{\big(d_\xi^{\pi^k}(s)\big)^2} \\
&= 4\alpha^2 |\mathcal{S}| + 4\alpha^2 \sum_s \left(\frac{d_\xi^{\pi^*}(s)}{d_\xi^{\pi^k}(s)} \right)^2 \\
&\le 4\alpha^2 |\mathcal{S}| + 4\alpha^2 |\mathcal{S}| \left\| \frac{d_\xi^{\pi^*}}{d_\xi^{\pi^k}} \right\|_\infty^2 \\
&\le 4\alpha^2 |\mathcal{S}| \cdot \left(1 + (1-\gamma)^{-2} \left\| d_\xi^{\pi^*}/\xi \right\|_\infty^2 \right) \\
&\le \frac{5\alpha^2 |\mathcal{S}|}{(1-\gamma)^2} \left\| d_\xi^{\pi^*}/\xi \right\|_\infty^2
\end{aligned}
$$

Denote $D := \frac{5|\mathcal{S}|}{(1-\gamma)^2} \left\| d_\xi^{\pi^*}/\xi \right\|_\infty^2$. Substituting the above two inequalities into the right-hand side of (I.1), we get

$$
\begin{aligned}
&F(\Lambda(\pi^*)) - F(\Lambda(\pi^{k+1})) \\
&\quad \le \min_{\alpha\in[0,1]} \big\{ F(\Lambda(\pi^*)) - F(\Lambda(\pi_\alpha)) + L\|\pi_\alpha - \pi^k\|^2 : \pi_\alpha = \Pi(\alpha\Lambda(\pi^*) + (1-\alpha)\Lambda(\pi^k)) \big\} \\
&\quad \le \min_{\alpha\in[0,1]} (1-\alpha)\big(F(\Lambda(\pi^*)) - F(\Lambda(\pi^k))\big) + LD\alpha^2 \,. \qquad \text{(I.4)}
\end{aligned}
$$

Note that (I.4) differs from (G.4) by replacing $\ell_\theta^2 D_\lambda^2$ with D. The latter proof of Theorem 4.5 is almost identical to that of Theorem 4.4 and hence we omit the proof. □