Policy-based Primal-Dual Methods for Convex Constrained Markov Decision Processes

Donghao Ying
UC Berkeley
Berkeley, CA 94720
donghaoy@berkeley.edu

Mengzi Guo
UC Berkeley
Berkeley, CA 94720
mengzi_guo@berkeley.edu

Yuhao Ding
UC Berkeley
Berkeley, CA 94720
yuhao_ding@berkeley.edu

Javad Lavaei
UC Berkeley
Berkeley, CA 94720
lavaei@berkeley.edu

Zuo-Jun (Max) Shen
UC Berkeley
Berkeley, CA 94720
maxshen@berkeley.edu

Abstract

We study convex Constrained Markov Decision Processes (CMDPs) in which the objective is concave and the constraints are convex in the state-action visitation distribution. We propose a policy-based primal-dual algorithm that updates the primal variable via policy gradient ascent and updates the dual variable via projected sub-gradient descent. Despite the loss of additivity structure and the nonconvex nature, we establish the global convergence of the proposed algorithm by leveraging a hidden convexity in the problem under the general soft-max parameterization, and prove the $\mathcal{O}\left(T^{-1/3}\right)$ convergence rate in terms of both optimality gap and constraint violation. When the objective is strongly concave in the visitation distribution, we prove an improved convergence rate of $\mathcal{O}\left(T^{-1/2}\right)$. By introducing a pessimistic term to the constraint, we further show that a zero constraint violation can be achieved while preserving the same convergence rate for the optimality gap. This work is the first one in the literature that establishes non-asymptotic convergence guarantees for policy-based primal-dual methods for solving infinite-horizon discounted convex CMDPs.

1 Introduction

Reinforcement Learning (RL) aims to learn how to map situations to actions so as to maximize the expected cumulative reward. Mathematically, this objective can be rewritten as an inner product between the state visitation distribution induced by the policy and a policy-independent reward for each state-action pair. However, many decision-making problems of interests take a form beyond the cumulative reward, such as apprenticeship learning [1], diverse skill discovery [2], pure exploration [3], and state marginal matching [4], among others. Recently, [5, 6] abstract such problems as convex Markov Decision Processes (MDPs), which focus on finding a policy to maximize a concave function of the induced state-action visitation distribution.

However, in many safety-critical applications of convex MDP problems, e.g., in autonomous driving [7], cyber-security [8], and financial management [9], the agent is also subject to safety constraints. Nonetheless, the classical safe RL and CMDPs [10], which assume that the objective and constraints are linear in the state-action visitation distribution, are not directly applicable to more general convex CMDP problems where the objective and the constraints can respectively be concave and convex in the state-action visitation distribution.

Preprint. Under review.

In this paper, we focus on the optimization perspective of convex CMDP problems and aim to develop a principled methodology and theory for the direct policy search method. When moving beyond linear structures in the objective and the constraints, we quickly face several technical challenges. Firstly, the convex CMDP problem has a nonconcave objective and the nonconvex constraints even under the simplest direct policy parameterization. Thus, the existing tools from the convex constrained optimization literature are not applicable. Secondly, as the gradient of the objective/constraint with respect to the state-action visitation distribution becomes policy-dependent, evaluating the single-step improvement of the algorithm becomes harder without knowing the visitation distribution. Yet, evaluating the visitation distribution for a given policy can be inefficient [11]. Thirdly, the performance difference lemma [12], which is key to the analysis of the policy-based primal-dual method for the standard CMDP [13], is no longer helpful for general convex CMDPs.

In view of the aforementioned challenges, our main contributions to the policy search of convex CMDP problems are summarized in Table 1 and are provided below:

- Despite being nonconvex with respect to the policy and nonlinear with respect to the state-action visitation distribution, we prove that the strong duality still holds for convex CMDP problems under some mild conditions.
- We propose a simple but effective algorithm – Primal-Dual Projected Gradient method (PDPG) – for solving discounted infinite-horizon convex CMDPs. We employ policy gradient ascent to update the primal variable and projected sub-gradient descent to update the dual variable. Strong bounds on the optimality gap and the constraint violations are established for both the convex objective and the the strongly concave objective cases.
- Inspired by the idea of "optimistic pessimism in the face of uncertainty", we further propose a modified method, named PDPG-0, which can achieve a zero constraint violation while maintaining the same convergence rate as the PDPG method.

Algorithm	**Objective**	**Constraint**	**Optimality Gap**	**Constraint Violation**
PDPG	Concave	Convex	$\mathcal{O}\left(T^{-1/3}\right)$	$\mathcal{O}\left(T^{-1/3}\right)$
PDPG	Strongly concave	Convex	$\mathcal{O}\left(T^{-1/2}\right)$	$\mathcal{O}\left(T^{-1/2}\right)$
PDPG-0	Concave	Convex	$\mathcal{O}\left(T^{-1/3}\right)$	0
PDPG-0	Strongly concave	Convex	$\mathcal{O}\left(T^{-1/2}\right)$	0

Table 1: We summarize our results for the policy-based primal dual methods for general convex CMDPs. Here T is the total number of iterations.

1.1 Related work

Convex MDP Motivated by emerging applications in RL whose objectives are beyond cumulative rewards [14, 1, 15, 3, 16, 4], a series of recent works have focused on developing general approaches for convex MDPs. In particular, [5] develops a new policy gradient approach called variational policy gradient and establishes the global convergence of the gradient ascent method by exploiting the hidden convexity of the problem. The REINFORCE-based policy gradient and its variance-reduced version are studied in [17]. The paper [6] transforms the convex MDP problem to a saddle-point problem using Fenchel duality and proposes a meta-algorithm to solve the problem with standard RL techniques. The work [18] proves the equivalence between convex MDPs and mean-field games (MFGs) and shows that algorithms for MFGs can be used to solve convex MDPs. However, the above papers only consider the unconstrained RL problem, which may lead to undesired policies in safety-critical applications. Therefore, additional effort is required to deal with the rising safety constraints, and our work addresses this challenge.

CMDP Our work is also pertinent to policy-based CMDP algorithms [10, 19–23]. In particular, [13] develops a natural policy gradient-based primal-dual algorithm and shows that it enjoys an $\mathcal{O}(T^{-1/2})$ global convergence rate regarding both the optimality gap and the constraint violation under the soft-max parameterization. The work [24] considers a primal-based approach and achieves a similar global convergence rate. More recently, [25–27] introduce entropy regularization and obtain improved convergence rates with dual methods. Nonetheless, these works focus on cumulative rewards/utilities and do not directly generalize to a boarder class of safe RL problems, such as safe imitation learning [28] and safe exploration [3]. Beyond CMDPs with cumulative rewards/utilities, the concurrent work [29] also studies the convex CMDP problem, and their algorithm is based on the randomized linear programming method proposed by [30]. However, as their approach works directly in the space of state-action visitation distributions, it is thus not applicable to more general problems where the state-action spaces are large and a function approximation is needed. In comparison, our work addresses this issue by focusing on the policy-based primal-dual method and adopting a general soft-max policy parameterization.

1.2 Notations

For a finite set $\mathcal{S}$, let $\Delta(\mathcal{S})$ denote the probability simplex over $\mathcal{S}$, and let $|\mathcal{S}|$ denote its cardinality. When the variable s follows the distribution ρ, we write it as $s \sim \rho$. Let $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot \mid \cdot]$, respectively, denote the expectation and conditional expectation of a random variable. Let $\mathbb{R}$ denote the set of real numbers. For a vector x, we use $x^\top$ to denote the transpose of x and use $\langle x, y\rangle$ to denote the inner product $x^\top y$. We use the convention that $\|x\|_1 = \sum_i |x_i|$, $\|x\|_2 = \sqrt{\sum_i x_i^2}$, and $\|x\|_\infty = \max_i |x_i|$. For a set $X \subset \mathbb{R}^p$, let $\mathrm{cl}(X)$ denote the closure of X. Let $\mathcal{P}_X$ denote the projection onto X, defined as $\mathcal{P}_X(y) \coloneqq \arg\min_{x\in X} \|x - y\|_2$. For a matrix A, let $\|A\|_2$ stand for the spectral norm, i.e., $\|A\|_2 = \max_{\|x\|_2\neq 0} \{\|Ax\|_2/\|x\|_2\}$. For a function $f(x)$, let $\arg\min f(x)$ (resp. $\arg\max f(x)$) denote any global minimum (resp. global maximum) of $f(x)$.

2 Problem Formulation

Standard CMDP Consider an infinite-horizon CMDP over a finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$ with a discount factor $\gamma \in [0,1)$. Let ρ be the initial distribution. The transition dynamics is given by $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\mathbb{P}(s'|s,a)$ is the probability of transition from state s to state s' when action a is taken. A policy is a function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that represents the decision rule that the agent uses, i.e., the agent takes action a with probability $\pi(a|s)$ in state s. We denote the set of all stochastic policies as Π. The goal of the agent is to find a policy that maximizes some long-term objective. In standard CMDPs, the agent aims at maximizing the expected (discounted) cumulative reward for a given initial distribution ρ while satisfying constraints on the expected (discounted) cumulative cost, i.e.,

$$\begin{aligned} \max_{\pi\in\Pi} V^\pi(r) &\coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \,\middle|\, a_t \sim \pi(\cdot|s_t), s_0 \sim \rho\right], \\ \text{s.t. } V^\pi(c) &\coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t,a_t) \,\middle|\, a_t \sim \pi(\cdot|s_t), s_0 \sim \rho\right] \le 0, \end{aligned} \tag{1}$$

where the expectation is taken over all possible trajectories, and $r(\cdot,\cdot)$ and $c(\cdot,\cdot)$ denote the reward and cost functions, respectively. For given reward function $r(\cdot,\cdot)$, we define the action-value function (Q-function) under policy π as

$$Q^\pi(r; s, a) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \,\middle|\, a_t \sim \pi(\cdot|s_t), s_0 = s, a_0 = a\right], \tag{2}$$

which can be interpreted as the expected total reward with an initial state $s_0 = s$ and an initial action $a_0 = a$. For each policy $\pi \in \Pi$ and state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, the discounted state-action visitation distribution is defined as

$$\lambda^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a \mid \pi, s_0 \sim \rho). \tag{3}$$

We use Λ to denote the set of all possible state-action visitation distributions, which is a convex polytope (cf. (26)). By using λ as decision variables, the CMDP problem in (1) can be re-parameterized as follows:

$$\max_{\lambda\in\Lambda} \frac{1}{1-\gamma}\langle r, \lambda\rangle, \quad \text{s.t.} \quad \frac{1}{1-\gamma}\langle c, \lambda\rangle \le 0. \tag{4}$$

This is known as the linear programming formulation of the CMDP [10]. Once a solution $\lambda^\star$ is computed, the corresponding policy can be recovered using the relation $\pi(a|s) = \lambda^\pi(s,a)/\sum_{a'\in\mathcal{A}} \lambda^\pi(s,a')$.

Convex CMDP In this work, we consider a more general problem where the agent's goal is to find a policy that maximizes a concave function of the state-action visitation distribution λ subject to a single convex constraint on λ, namely

$$\max_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad g(\lambda) \le 0, \tag{5}$$

where f is concave and g is convex. As (5) is a convex program in λ, we refer to the problem as Convex CMDP. We emphasize that the method proposed in this paper directly generalizes to multiple constraints and we present the single constraint setting only for brevity.

Example 2.1 (Safety-aware apprenticeship learning (AL)) *In AL, instead of maximizing the long-term reward, the agent learns to mimic an expert's demonstrations. When there are critical safety requirements, the learner will also strive to satisfy given constraints on the expected total cost [28]. This problem can be formulated as*

$$\max_{\lambda \in \Lambda} f(\lambda) = -\operatorname{dist}(\lambda, \lambda_e) \quad \textit{s.t.} \quad g(\lambda) = \frac{1}{1-\gamma}\langle c, \lambda\rangle \le 0, \tag{6}$$

where λ_e corresponds to the expert demonstration, c denotes the cost function, and $\operatorname{dist}(\cdot,\cdot)$ can be any distance function on Λ, e.g., ℓ^2-distance or Kullback-Liebler (KL) divergence.

Example 2.2 (Feasibility constrained MDPs) *As an extension to standard CMDPs, the designer may desire to control the MDP through more general constraints described by a convex feasibility region C [31] (e.g., a single point representing a known safe policy) such that the learned policy is not too far away from C. In this case, the problem can be cast as*

$$\max_{\lambda \in \Lambda} f(\lambda) = \frac{1}{1-\gamma}\langle r, \lambda\rangle \quad \textit{s.t.} \quad g(\lambda) = \operatorname{dist}(\lambda, C) - d_0 \le 0, \tag{7}$$

where $d_0 \ge 0$ denotes the threshold of the allowable deviation.

Policy Parameterization Since recovering a policy from its corresponding state-action visitation distribution is toilless, a natural approach to solving the convex CMDP problem is to optimize (5) directly (or equivalently (4) for standard CMDPs). However, since the decision variable λ has the size $|\mathcal{S}||\mathcal{A}|$, such approaches lack scalability and converge extremely slowly for large state and action spaces. In this work, we consider the direct policy search method, which can handle the curse of dimensionality via the policy parameterization. We assume that policy $\pi = \pi_\theta$ is parameterized by a general soft-max function, meaning that

$$\pi_\theta(a|s) = \frac{\exp\{\psi(\theta; s, a)\}}{\sum_{a' \in \mathcal{A}} \exp\{\psi(\theta; s, a')\}}, \ \forall \ (s, a) \in \mathcal{S} \times \mathcal{A}, \tag{8}$$

where $\psi(\cdot\,; s, a)$ is some smooth function, $\theta \in \Theta$ is the parameter vector, and Θ is a convex feasible set. We assume that θ over-parameterizes the set of all stochastic policies in the sense that $\operatorname{cl}(\lambda(\Theta)) = \Lambda$. Further assumptions on the parameterization will be formally stated in Section 4. In practice, the function ψ can be chosen to be a deep neural network, where θ is the parameter and the state-action pair (s, a) is the input. Under parameterization (8), problem (5) can be re-written as

$$\max_{\theta \in \Theta} F(\theta) \coloneqq f(\lambda(\theta)) \quad \text{s.t.} \quad G(\theta) \coloneqq g(\lambda(\theta)) \le 0, \tag{9}$$

where we use the shorthand notations $\lambda(\theta) \coloneqq \lambda^{\pi_\theta}$ and $\lambda(\theta; s, a) \coloneqq \lambda^{\pi_\theta}(s, a)$. It is worth mentioning that (9) is a nonconvex problem due to its nonconcave objective function and nonconvex constraints with respect to θ.

Lagrangian Duality Consider the Lagrangian function associated with (9), $\mathcal{L}(\theta, \mu) \coloneqq F(\theta) - \mu G(\theta)$. For the ease of theoretical analysis, we define $L(\lambda, \mu) \coloneqq f(\lambda) - \mu g(\lambda)$, which is concave in λ when $\mu \ge 0$. It is clear that $\mathcal{L}(\theta, \mu) = L(\lambda(\theta), \mu)$. The dual function is defined as $\mathcal{D}(\mu) \coloneqq \max_{\theta \in \Theta} \mathcal{L}(\theta, \mu)$. Let $\pi_{\theta^\star}$ be the optimal policy such that $\theta^\star$ is the optimal solution to (9), and $\mu^\star$ be the optimal dual variable.

In constrained optimization, strict feasibility can induce many desirable properties. Assume that the following Slater's condition holds.

Assumption 2.1 (Slater's condition) *There exist $\widetilde{\theta} \in \Theta$ and $\xi > 0$ such that $G(\widetilde{\theta}) = g(\lambda(\widetilde{\theta})) \le -\xi$.*

The Slater's condition is a standard assumption and it holds when the feasible region has an interior point. In practice, such a point is often easy to find using prior knowledge of the problem. The following result is a direct consequence of the Slater's condition [10].

Lemma 2.2 (Strong duality and boundedness of $\mu^\star$) *Let Assumption 2.1 hold and suppose that* $\mathrm{cl}\left(\lambda(\Theta)\right) = \Lambda$. *We have: **(I)** $F\left(\theta^\star\right) = \mathcal{D}\left(\mu^\star\right) = \mathcal{L}\left(\theta^\star, \mu^\star\right)$, **(II)** $0 \le \mu^\star \le (F(\theta^\star) - F(\widetilde{\theta}))/\xi$.*

For completeness, we provide a proof for Lemma 2.2 in Appendix A. The strong duality implies that (9) is equivalent to the following saddle point problem:

$$\max_{\theta\in\Theta} \min_{\mu\ge 0} \mathcal{L}(\theta, \mu) = \min_{\mu\ge 0} \max_{\theta\in\Theta} \mathcal{L}(\theta, \mu). \tag{10}$$

Motivated by this equivalence, we seek to develop a primal-dual algorithm to solve the problem.

3 Safe Policy Search Beyond Cumulative Rewards/Utilities

To solve (10), we propose the following Primal-Dual Projected Gradient Algorithm (PDPG):

$$\theta^{t+1} = \mathcal{P}_\Theta\left(\theta^t + \eta_1 \nabla_\theta \mathcal{L}(\theta^t, \mu^t)\right), \quad \mu^{t+1} = \mathcal{P}_U\left(\mu^t - \eta_2 \nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right), \quad \text{for } t = 0, 1, 2, \ldots, \tag{11}$$

where $\eta_1 > 0$, $\eta_2 > 0$ are constant step-sizes, and the dual feasible region $U := [0, C_0]$ is an interval that contains $\mu^\star$. By Lemma 2.2, choosing $C_0 \ge (F(\theta^\star) - F(\widetilde{\theta}))/\xi$ satisfies the requirement. The method (11) adopts an alternating update scheme: the primal step performs the projected gradient ascent in the policy space, whereas dual step updates the multiplier with projected sub-gradient descent such that μ^{t+1} is obtained by adding a multiple of the constraint violation to μ^t. The values of C_0, η_1, η_2 will be specified later in the paper.

Unlike standard CMDPs (1) where the value function is defined as discounted cumulative rewards/utilities and admits an additive structure, performing and analyzing algorithm (11) are far more challenging for convex CMDPs.

3.1 Gradient Evaluation of the Lagrangian Function

Computing the primal update in (11) involves evaluating the gradient of the Lagrangian with respect to θ, i.e., $\nabla_\theta \mathcal{L}(\theta, \mu) = \nabla_\theta \left[f(\lambda(\theta)) - \mu g(\lambda(\theta))\right]$. When $f(\cdot)$ and $g(\cdot)$ are linear as in the standard CMDP, i.e., $f(\lambda) = \langle r, \lambda\rangle/(1-\gamma)$ and $g(\lambda) = \langle c, \lambda\rangle/(1-\gamma)$, the *Policy gradient* theorem (cf. Lemma G.1) can be applied, implying that

$$\nabla_\theta \mathcal{L}(\theta, \mu) = \nabla_\theta V^{\pi_\theta}(r - \mu c) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \cdot Q^{\pi_\theta}(r - \mu c; s, a)\right], \tag{12}$$

where $d^\pi(s) := \sum_{a\in\mathcal{A}} \lambda^\pi(s, a)$ is the discounted state visitation distribution. However, this favorable result is not applicable when $f(\cdot)$ and $g(\cdot)$ are general concave/convex functions. Instead, we present two alternative approaches.

Variational Policy Gradient By leveraging the Fenchel duality, [5] showed that the gradient $\nabla_\theta \mathcal{L}(\theta, \mu)$ can be computed by solving a stochastic saddle point problem, in particular

$$\nabla_\theta \mathcal{L}(\theta, \mu) = \lim_{\delta\to 0_+} \operatorname*{argmax}_x \inf_z \left\{(1-\gamma)\left[V^{\pi_\theta}(z) + \delta \langle \nabla_\theta V^{\pi_\theta}(z), x\rangle\right] - L_*(z, \mu) - \frac{\delta}{2}\|x\|^2\right\}, \tag{13}$$

where $L_*(z, \mu) := \inf_\lambda \{\langle z, \lambda\rangle - L(\lambda, \mu)\}$ is the concave conjugate of $L(\lambda, \mu)$ with respect to λ. As $L_*(z, \mu)$ is concave in z and $V^{\pi_\theta}(z) = \langle z, \lambda(\theta)\rangle$ is linear in z, the max-min problem in (13) is a concave-convex saddle point problem.

REINFORCE-based Policy Gradient By noticing the relation $\nabla_\theta V^{\pi_\theta}(r) = [\nabla_\theta \lambda(\theta)]^\top \cdot r/(1-\gamma)$, one can view $\nabla_\theta \mathcal{L}(\theta, \mu)$ as the standard policy gradient for the value function with the reward $\nabla_\lambda L(\lambda(\theta), \mu)$ (assuming that $f(\lambda)$, $g(\lambda)$, and $\lambda(\theta)$ are all differentiable), i.e.,

$$\nabla_\theta \mathcal{L}(\theta, \mu) = [\nabla_\theta \lambda(\theta)]^\top \cdot \nabla_\lambda L(\lambda(\theta), \mu) = (1-\gamma) \nabla_\theta V^{\pi_\theta}(\nabla_\lambda L(\lambda(\theta), \mu)), \tag{14}$$

where the first equality follows from the chain rule. Thus, the gradient $\nabla_\theta \mathcal{L}(\theta, \mu)$ can be estimated with the REINFORCE algorithm [32] as long as we can choose an approximation of $\nabla_\lambda L(\lambda(\theta), \mu)$ as the reward. We refer the reader to [17, Section 4] for more details.

Since $\nabla_\mu \mathcal{L}(\theta, \mu) = -G(\theta) = -g(\lambda(\theta))$, performing the dual update (11) requires evaluating the constraint function. In cases where an efficient oracle for computing $g(\lambda(\theta))$ from θ is not available, we can formulate it as another convex problem using the Fenchel duality to avoid directly estimating the current state-action visitation distribution $\lambda(\theta)$:

$$\nabla_\mu \mathcal{L}(\theta, \mu) = -g(\lambda(\theta)) = -\sup_z \{\langle z, \lambda\rangle - g^*(z)\} = -\sup_z \{(1-\gamma)V^{\pi_\theta}(z) - g^*(z)\}. \tag{15}$$

where $g^*(z) \coloneqq \sup_\lambda \{\langle z, \lambda\rangle - g(\lambda)\}$ is the convex conjugate of $g(\cdot)$ and we use the fact that the biconjugate of a convex function equals itself, i.e., $g^{**}(\lambda) \coloneqq \sup_z \{\langle z, \lambda\rangle - g^*(z)\} = g(\lambda)$.

3.2 Exploiting the Hidden Convexity

By itself, (10) is a nonconvex-linear maximin problem. The existing results for the analysis of the gradient ascent descent method (11) for such problems can only guarantee to find a ϵ-stationary point in $\mathcal{O}(\epsilon^{-6})$ iterations [33]. To obtain an improved convergence rate and achieve a global optimality, it is necessary to exploit the "hidden convexity" of (9) with respect to λ.

However, standard analyses based on the performance difference lemma (cf. Lemma G.4) do not apply to convex CMDPs [34, 13]. A key insight is that, due to the loss of linearity, the performance difference lemma together with concavity can only provide an upper bound for the single-step improvement with the gradient information at the current step (the derivation can be found in Appendix A.1):

$$\mathcal{L}(\theta^{t+1}, \mu^t) - \mathcal{L}(\theta^t, \mu^t) \le \mathbb{E}_{s\sim d^{\pi_{\theta^t}}} \left\langle \pi_{\theta^{t+1}}(\cdot|s) - \pi_{\theta^t}(\cdot|s), Q^{\pi_{\theta^{t+1}}}\left(\nabla_\lambda L(\lambda(\theta^t), \mu^t); s, \cdot\right)\right\rangle. \tag{16}$$

Thus, this prompts us to introduce a new analysis to bound the average performance in terms of the Lagrangian (cf. (18)).

Following [5], we leverage the fact that the primal update implies the formula

$$\mathcal{L}(\theta^{t+1}, \mu^t) = \max_{\theta\in\Theta} \left\{\mathcal{L}(\theta^t, \mu^t) + (\theta - \theta^t)^\top \nabla_\theta \mathcal{L}(\theta^t, \mu^t) - \frac{1}{2\eta_1}\|\theta - \theta^t\|_2^2\right\}. \tag{17}$$

With a proper step-size η_1, (17) guarantees a strict single-step improvement. The basis of our analysis lies in designing a special point θ from θ^t and $\theta^\star$ to lower-bound $\mathcal{L}(\theta^{t+1}, \mu^t)$ through (17). The hidden convexity of (9) with respect to λ plays a central role in bounding the improvement $\mathcal{L}(\theta^{t+1}, \mu^t) - \mathcal{L}(\theta^t, \mu^t)$ and relating it to the sub-optimality gap $\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)$. The details are deferred to Section 4.

4 Convergence Analysis

In this section, we establish the global convergence of the primal-dual projected gradient algorithm (11) by exploiting the hidden convexity of (9) with respect to λ. We refer the reader to the supplement in Appendix B for the proofs of this section.

First, we formally state our assumption about the parameterization (8). To avoid introducing an additional bias, it is natural to assume that the parameterization has enough expressibility to represent any policy, i.e., $\forall \pi \in \Pi$, $\exists \theta \in \Theta$ such that $\pi = \pi_\theta$. However, assuming a one-to-one correspondence between $\pi \in \Pi$ and $\pi_\theta, \theta \in \Theta$ is too restrictive. In practice, using a deep neural network to represent the policy can often arrive at an over-parameterization. Therefore, following [17], we assume that π_θ is defined such that it can represent any policy and that $\lambda(\cdot)$ is locally continuously invertible. A more detailed discussion can be found in Appendix D.

Assumption 4.1 (Parameterization) *The policy parameterization $\pi = \pi_\theta$ over-parameterizes the set of all stochastic policies and satisfies:* ***(I)*** *For every $\theta \in \Theta$, there exists a neighborhood $\mathcal{U}_\theta \ni \theta$ such that the restriction of $\lambda(\cdot)$ to $\mathcal{U}_\theta$ is a bijection between $\mathcal{U}_\theta$ and $\mathcal{V}_{\lambda(\theta)} \coloneqq \lambda(\mathcal{U}_\theta)$;* ***(II)*** *Let $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta)}} : \mathcal{V}_{\lambda(\theta)} \to \mathcal{U}_\theta$ be the local inverse of $\lambda(\cdot)$, i.e., $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta)}}(\lambda(\theta_0)) = \theta_0,\ \forall \theta_0 \in \mathcal{U}_\theta$. Then, there exists a universal constant ℓ_Θ such that $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta)}}$ is ℓ_Θ-Lipschitz continuous for all $\theta \in \Theta$;* ***(III)*** *There exists $\bar{\varepsilon} > 0$ such that $(1-\varepsilon)\lambda(\theta) + \varepsilon\lambda(\theta^\star) \in \mathcal{V}_{\lambda(\theta)},\ \forall \varepsilon \le \bar{\varepsilon},\ \forall \theta \in \Theta$.*

We also make the following assumption about the smoothness of the objective and constraint functions.

Assumption 4.2 (Smoothness) *$F(\theta)$ is ℓ_F-smooth with respect to (w.r.t.) θ and $G(\theta)$ is ℓ_G-smooth w.r.t. θ, i.e., $\|\nabla_\theta F(\theta_1) - \nabla_\theta F(\theta_2)\|_2 \le \ell_F \|\theta_1 - \theta_2\|_2$ and $\|\nabla_\theta G(\theta_1) - \nabla_\theta G(\theta_2)\|_2 \le \ell_G \|\theta_1 - \theta_2\|_2$, $\forall\ \theta_1, \theta_2 \in \Theta$.*

In optimization, smoothness is important when analyzing the convergence rate of an algorithm. In Appendix E, we provide a discussion which shows that Assumption 4.2 is mild in the sense that if $f(\lambda)$ is smooth with respect to λ, then $F(\theta) = f(\lambda(\theta))$ is smooth with respect to θ under some regularity conditions.

The following property about $\mathcal{L}(\theta, \mu)$ is the direct consequence of Assumption 4.2.

Lemma 4.3 *The functions $f(\cdot)$ and $g(\cdot)$ are bounded on Λ. Define M_F and M_G such that $|f(\lambda)| \le M_F$ and $|g(\lambda)| \le M_G$, for all $\lambda \in \Lambda$. Then, it holds that $|\mathcal{L}(\theta, \mu)| \le M_L,\ \forall\ \theta \in \Theta, \mu \in U$, where $M_L := M_F + C_0 M_G$. Furthermore, under Assumption 4.2, $\mathcal{L}(\cdot, \mu)$ is ℓ_L-smooth on Θ, for all $\mu \in U$, where $\ell_L := \ell_F + C_0 \ell_G$.*

To quantify the quality of a given solution θ to (9), the measures we consider are the optimality gap $F(\theta^\star) - F(\theta)$ and the constrained violation $[G(\theta)]_+$, where $[x]_+ := \max\{x, 0\}$. Unlike the unconstrained setting where the last-iterate convergence is of more interest, a primal-dual algorithm for constrained optimization often cannot ensure an effective improvement in every iteration due to the change of the multiplier. Therefore, we focus on the global convergence of algorithm (11) in the time-average sense.

We first bound the average performance in terms of the Lagrangian below.

Proposition 4.4 *Let Assumptions 4.1 and 4.2 hold and assume that $\varepsilon \le \bar{\varepsilon}$. Then, for every $T > 0$, the iterates $\{(\theta^t, \mu^t)\}_{t=0}^{T-1}$ produced by algorithm* (11) *with $\eta_1 = 1/\ell_L$ satisfy*

$$\frac{1}{T} \sum_{t=0}^{T-1} \left[\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)\right] \le \frac{\mathcal{L}(\theta^\star, \mu^0) - \mathcal{L}(\theta^0, \mu^0)}{\varepsilon T} + 2\varepsilon \ell_L \ell_\Theta^2 + \frac{2\eta_2 M_G^2}{\varepsilon}. \tag{18}$$

We remark that by choosing $\varepsilon = T^{-1/3}$ and $\eta_2 = T^{-2/3}$, the bound (18) given by Proposition 4.4 has the order of $\mathcal{O}(T^{-1/3})$. The core idea in proving Proposition 4.4 is that one can relate the primal update (17) to the sub-optimality gap $\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)$ by leveraging the hidden convexity of (9) with respect to λ. Then, as $|\mu^{t+1} - \mu^t| = \mathcal{O}(\eta_2)$, we are able to draw a recursion between the sub-optimality gaps for two consecutive periods (cf. (35)).

The average performance in terms of the Lagrangian can be decomposed into the summation of the average optimality gap and the weighted average "constraint violation", i.e.,

$$\frac{1}{T} \sum_{t=0}^{T-1} \left[\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)\right] = \frac{1}{T} \sum_{t=0}^{T-1} \left[F(\theta^\star) - F(\theta^t)\right] + \frac{1}{T} \sum_{t=0}^{T-1} \mu^t \left[G(\theta^t) - G(\theta^\star)\right]. \tag{19}$$

Since $\theta^\star$ must be a feasible solution, the term $\left[G(\theta^t) - G(\theta^\star)\right]$ can be interpreted as an approximate of the constraint violation. To obtain separate bounds for the optimality gap and the true constraint violation, we need to decouple the bound for the average performance.

Theorem 4.5 (General concavity) *Let Assumptions 2.1, 4.1, and 4.2 hold. For every $T \ge (\bar{\varepsilon})^{-3}$, we choose $C_0 = 1 + \left(M_F - F(\widetilde{\theta})\right)/\xi$, $\mu^0 = 0$, $\eta_1 = 1/\ell_L$, and $\eta_2 = T^{-2/3}$. Then, the sequence $\{(\theta^t, \mu^t)\}_{t=0}^{T-1}$ generated by algorithm (11) converges with the rate $\mathcal{O}(T^{-1/3})$, in particular*

$$\textit{Optimality Gap: } \frac{1}{T} \sum_{t=0}^{T-1} \left[F(\theta^\star) - F(\theta^t)\right] \le \frac{2M_F + M_G^2/2}{T^{2/3}} + \frac{2\ell_L \ell_\Theta^2 + 2M_G^2}{T^{1/3}}, \tag{20a}$$

$$\textit{Constraint Violation: } \frac{1}{T} \left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ \le \frac{2M_F + M_G^2/2}{T^{2/3}} + \frac{2\ell_L \ell_\Theta^2 + 2M_G^2 + C_0^2/2}{T^{1/3}}. \tag{20b}$$

Theorem 4.5 shows that algorithm (11) achieves a global convergence in the average sense such that the optimality gap and the constraint violation decay to zero with the rate $\mathcal{O}(T^{-1/3})$. In other words, to obtain an $\mathcal{O}(\epsilon)$-accurate solution, the iteration complexity is $\mathcal{O}(\epsilon^{-3})$. When $f(\cdot)$ and $g(\cdot)$ are linear functions as in standard CMDPs (cf. (1)), Theorem 4.5 matches the rate of the natural policy gradient primal-dual algorithm [13, Theorem 2][1].

In Theorem 4.5, the dual feasible region $U = [0, C_0]$ is set by taking $C_0 = 1 + \left(M_F - F(\widetilde{\theta})\right)/\xi$. By Lemma 2.2, $\mu^\star \le (F(\theta^\star) - F(\widetilde{\theta}))/\xi \le C_0 - 1$, which implies that $\widehat{\mu} := \mu^\star + 1 \in U$. This "slackness" plays an important role when bounding the constraint violation, as we can write $\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = \left[(\mu^\star - \widehat{\mu}) \sum_{t=0}^{T-1} \nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right]_+$, where the latter term can be related to the first-order expansion of $\mathcal{L}(\theta, \cdot)$ and bounded through the use of telescoping sums.

When the objective function $f(\lambda(\theta))$ is strongly concave with respect to λ, we can further improve the convergence rate of algorithm 11 by a similar line of analysis. Firstly, we establish the average performance bound in terms of the Lagrangian.

Proposition 4.6 *Let Assumptions 4.1 and 4.2 hold. Suppose that $f(\cdot)$ is σ-strongly concave w.r.t. λ on Λ. Then, for every $T > 0$, the iterates $\left\{(\theta^t, \mu^t)\right\}_{t=0}^{T-1}$ produced by algorithm (11) with $\eta_1 = 1/\ell_L$ satisfy*

$$\frac{1}{T} \sum_{t=0}^{T-1} [\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)] \le \frac{\mathcal{L}(\theta^T, \mu^{T-1}) - \mathcal{L}(\theta^0, \mu^0)}{\tilde{\varepsilon} T} + \frac{\eta_2 M_G^2}{\tilde{\varepsilon}}, \tag{21}$$

where $\tilde{\varepsilon} := \min\{\bar{\varepsilon}, \sigma/\left(\sigma + 2\ell_\Theta^2 \ell_L\right)\}$.

Different from the general concave case (18), the bound (21) does not contain the constant error term $\mathcal{O}(\varepsilon)$. Thus, by choosing $\eta_2 = T^{-1/2}$, the average performance has the order $\mathcal{O}\left(T^{-1/2}\right)$. In a similar manner as in Theorem 4.5, we can decouple the average performance to bound the optimality gap and constraint violation.

Theorem 4.7 (Strong concavity) *Let Assumptions 2.1, 4.1, and 4.2 hold. Suppose that $f(\cdot)$ is σ-strongly concave w.r.t. λ on Λ. For every $T > 0$, we choose $C_0 = 1 + \left(M_F - F(\widetilde{\theta})\right)/\xi$, $\mu^0 = 0$, $\eta_1 = 1/\ell_L$, and $\eta_2 = T^{-1/2}$. Then, the sequence $\left\{(\theta^t, \mu^t)\right\}_{t=0}^{T-1}$ generated by algorithm (11) converges with the rate $\mathcal{O}\left(T^{-1/2}\right)$, in particular*

$$\textit{Optimality Gap: } \frac{1}{T} \sum_{t=0}^{T-1} \left[F(\theta^\star) - F(\theta^t)\right] \le \frac{M_L + M_F}{\tilde{\varepsilon} T} + \left(\frac{M_G^2}{\tilde{\varepsilon}} + \frac{M_G^2}{2}\right) \frac{1}{\sqrt{T}}, \tag{22a}$$

$$\textit{Constraint Violation: } \frac{1}{T} \left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ \le \frac{M_L + M_F}{\tilde{\varepsilon} T} + \left(\frac{M_G^2}{\tilde{\varepsilon}} + \frac{M_G^2 + C_0^2}{2}\right) \frac{1}{\sqrt{T}}, \tag{22b}$$

where $\tilde{\varepsilon} := \min\{\bar{\varepsilon}, \sigma/\left(\sigma + 2\ell_\Theta^2 \ell_L\right)\}$.

Theorem 4.7 shows that when $f(\cdot)$ is strongly concave, algorithm (11) admits an improved convergence rate of $\mathcal{O}\left(T^{-1/2}\right)$ by taking the dual step-size $\eta_2 = \mathcal{O}\left(T^{-1/2}\right)$. Equivalently, the iteration complexity is $\mathcal{O}(\epsilon^{-2})$ to compute an $\mathcal{O}(\epsilon)$-accurate solution.

Remark 4.8 (Direct parameterization) *As a special case, the direct parameterization satisfies Assumption 4.1 once there is a universal positive lower bound for the state visitation distribution d^π. Under the direct parameterization, it can be shown that the primal update (17) also enjoys the so-called variational gradient dominance property for standard MDPs (see, e.g., [34, Lemma 4.1]). This evidence gives a clearer intuition of how the hidden convexity enables us to prove the global convergence of algorithm (11). We refer the reader to Appendix F for a detailed discussion.*

5 Zero Constraint Violation

In safety-critical systems where violating the constraint may induce an unexpected cost, having a zero constraint violation is of great importance. Following the recent works [35, 26], we will show

[1]Although the convergence rate presented in [13, Theorem 2] is $\left(\mathcal{O}\left(T^{-1/2}\right), \mathcal{O}\left(T^{-1/4}\right)\right)$, we note that it can be converted to $\left(\mathcal{O}\left(T^{-1/3}\right), \mathcal{O}\left(T^{-1/3}\right)\right)$ by choosing $\eta_1 = \eta_2 = T^{-1/3}$.

that a zero constraint violation can be achieved while maintaining the same order of convergence rate for the optimality gap. Consider the pessimistic counterpart of (9):

$$\max_{\theta\in\Theta} F(\theta) = f(\lambda(\theta)) \quad \text{s.t.} \quad G(\theta) = g(\lambda(\theta)) \leq -\delta, \tag{23}$$

where $\delta > 0$ is the pessimistic term to be determined. In the following theorem, we show that by applying algorithm (11) to the pessimistic problem (23) with a carefully chosen δ, the constraint violation will be zero for the original problem (9) when T is reasonably large. We refer to this variate as the Primal-dual Policy Gradient-Zero Algorithm (PDPG-0). The informal version of the theorem is stated below and we direct the reader to Appendix C for a detailed statement as well as the proof.

Theorem 5.1 *Let Assumptions 2.1, 4.1, and 4.2 hold.* ***(I)*** *For every reasonably large $T > 0$, the sequence $\left\{\left(\theta^t, \mu^t\right)\right\}_{t=0}^{T-1}$ generated by the PDPG-0 algorithm with $\delta = \mathcal{O}(T^{-1/3})$ satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] = \mathcal{O}(T^{-1/3}), \quad \frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = 0. \tag{24}$$

(II) *When $f(\cdot)$ is σ-strongly concave w.r.t. λ on Λ, the sequence $\left\{\left(\theta^t, \mu^t\right)\right\}_{t=0}^{T-1}$ generated by the PDPG-0 algorithm with $\delta = \mathcal{O}(T^{-1/2})$ satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] = \mathcal{O}(T^{-1/2}), \quad \frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = 0. \tag{25}$$

We briefly introduce the ideas behind Theorem 5.1 here. Adding the pessimistic term δ would shift the optimal solution from $\theta^\star$ to another point $\theta^\star_\delta$. By leveraging the Slater's condition, we can upper-bound the sub-optimality gap $|F(\theta^\star) - F(\theta^\star_\delta)|$ by $\mathcal{O}(\delta)$. Since the orders of convergence rates are the same for optimality gap and constraint violation (cf. Theorem 4.5 and 4.7), we can choose δ to have the same order and then offset the constraint violation for the pessimistic problem (23). As a result, the constraint violation becomes zero for the original problem (9) and the optimality gap preserves its previous order.

6 Conclusion

In this work, we proposed a primal-dual projected gradient algorithm to solve convex CMDP problems. Under the general soft-max parameterization with an over-parameterization assumption, it is proved that the proposed method enjoys an $\mathcal{O}\left(T^{-1/3}\right)$ global convergence rate in terms of the optimality gap and constraint violation. When the objective is strongly concave in the state-action visitation distribution, we showed an improved convergence rate of $\mathcal{O}\left(T^{-1/2}\right)$. By considering a pessimistic counterpart of the original problem, we also proved that a zero constraint violation can be achieved while maintaining the same convergence rate for the optimality gap.

One important direction of future work lies in establishing a lower bound for convex CMDP problems under a general soft-max parameterization to verify the optimality of our upper bounds. Also, an extension to this work is studying the sample complexity of the PDPG method. Furthermore, it is interesting to study whether geometric structures, such as entropy regularization [25–27] or policy mirror descent [36], can be exploited to accelerate the convergence.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[3] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[4] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

[5] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

[6] Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

[8] Xiangyuan Zhang, Kaiqing Zhang, Erik Miehling, and Tamer Basar. Non-cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84, 2010.

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

[11] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[12] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[13] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

[14] Stefan Schaal. Learning from demonstration. *Advances in neural information processing systems*, 9, 1996.

[15] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

[16] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486. PMLR, 2019.

[17] Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

[19] Vivek S Borkar. An actor-critic algorithm for constrained markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

[20] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[21] Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. *arXiv preprint arXiv:2201.11965*, 2022.

[22] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[23] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[24] Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

[25] Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained markov decision processes with entropy regularization. *arXiv preprint arXiv:2110.08923*, 2021.

[26] Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Fast global convergence of policy optimization for constrained mdps. *arXiv preprint arXiv:2111.00552*, 2021.

[27] Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained markov decision process. *arXiv preprint arXiv:2110.10351*, 2021.

[28] Weichao Zhou and Wenchao Li. Safety-aware apprenticeship learning. In *International Conference on Computer Aided Verification*, pages 662–680. Springer, 2018.

[29] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for concave utility constrained reinforcement learning via primal-dual approach, 2021.

[30] Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

[31] Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement learning with convex constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[33] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[34] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[35] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

[37] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[38] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[39] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

Appendix A Supplementary Materials for Sections 2 and 3

Lemma A.1 (Restatement of Lemma 2.2) *Let Assumption 2.1 hold and suppose that* $\mathrm{cl}\left(\lambda(\Theta)\right) = \Lambda$. *We have:* ***(I)*** $F\left(\theta^{\star}\right) = \mathcal{D}\left(\mu^{\star}\right) = \mathcal{L}\left(\theta^{\star}, \mu^{\star}\right)$*,* ***(II)*** $0 \le \mu^{\star} \le (F(\theta^{\star}) - F(\widetilde{\theta}))/\xi$.

Proof. We note that Λ, the set of all possible state-action visitation distributions, is a convex polytope having the expression

$$\Lambda = \left\{\lambda \in \mathbb{R}^{|S||A|} \middle| \lambda \ge \mathbf{0}, \sum_{a} \lambda(s,a) = (1-\gamma)\cdot\rho(s) + \gamma \sum_{s',a'} \mathbb{P}\left(s|s',a'\right)\cdot\lambda\left(s',a'\right), \forall s \in S\right\}. \quad (26)$$

Then, since $\mathrm{cl}\left(\lambda(\Theta)\right) = \Lambda$, the nonconvex problem (9) is equivalent to the convex problem (5):

$$\max_{\lambda\in\Lambda} f(\lambda) \quad \text{s.t.} \quad g(\lambda) \le 0.$$

Therefore, the strong duality ***(I)*** naturally holds under Assumption 2.1 [37].

To prove ***(II)***, let $C \in \mathbb{R}$. For every $\mu \ge 0$ such that $\mathcal{D}(\mu) \le C$, it holds that

$$C \ge \mathcal{D}(\mu) \overset{(i)}{\ge} F(\widetilde{\theta}) - \mu G(\widetilde{\theta}) \overset{(ii)}{\ge} F(\widetilde{\theta}) + \mu\xi, \quad (27)$$

where (i) follows from the definition of $\mathcal{D}(\mu)$ and (ii) is due to Assumption 2.1.

Since $\xi > 0$, (27) gives rise to the bound $\mu \le \left(C - F(\widetilde{\theta})\right)/\xi$. Now, by letting $C = F(\theta^{\star})$, it results from the strong duality that $\{\mu \ge 0 \mid \mathcal{D}(\mu) \le C\}$ becomes the set of optimal dual variables. This completes the proof. ◻

A.1 Supplementary Materials for Section 3.2

We elaborate on the reason why the standard analysis based on the performance difference lemma does not apply. When $f(\lambda) = \langle r, \lambda\rangle/(1-\gamma)$ and $g(\lambda) = \langle c, \lambda\rangle/(1-\gamma)$, the Lagrangian $L(\lambda(\theta_{\rangle}, \mu)$ is linear in λ. Thus,

$$\begin{aligned}\mathcal{L}(\theta^{t+1},\mu^t) - \mathcal{L}(\theta^t,\mu^t) &= (1-\gamma)\left[V^{\pi_{\theta^{t+1}}}(r-\mu^t c) - V^{\pi_{\theta^t}}(r-\mu^t c)\right] \\ &= \sum_{s\in\mathcal{S}} d^{\pi_{\theta^t}}(s)\sum_{a\in\mathcal{A}}\left(\pi_{\theta^{t+1}}(a|s) - \pi_{\theta^t}(a|s)\right)\cdot Q^{\pi_{\theta^{t+1}}}(r-\mu^t c; s, a),\end{aligned}$$

where the second step follows from the performance difference lemma (cf. Lemma G.4). This provides a way to measure the improvement of the primal update. In particular, suppose that the primal update adopts the natural policy gradient [38], meaning that

$$\theta^{t+1} = \theta^t + \eta_1\left(\mathcal{F}^{\theta^t}\right)^{\dagger}\nabla_\theta\mathcal{L}(\theta^t,\mu^t),$$

where $\left(\mathcal{F}^{\theta^t}\right)^{\dagger}$ denotes the Moore–Penrose inverse of the Fisher-information matrix with respect to π_{θ^t}. The corresponding policy update follows that

$$\pi_{\theta^{t+1}}(a|s) \propto \pi_{\theta^t}(a|s)\exp\left(\frac{\eta_1 Q^{\pi_{\theta^t}}(r-\mu^t c; s,a)}{1-\gamma}\right)\Big/ Z^t(s),$$

where $Z^t(\cdot)$ denotes the normalization term. Then, the single step improvement has the following lower bound [13, Lemma 6]:

$$\mathcal{L}(\theta^{t+1},\mu^t) - \mathcal{L}(\theta^t,\mu^t) \ge \frac{(1-\gamma)^2}{\eta_1}\mathbb{E}_{s\sim\rho}\log Z^t(s) \ge 0.$$

However, when $L(\lambda(\theta),\mu)$ loses the linearity structure as in convex CMDPs, such argument no longer holds true. The reason is that with concavity we can only obtain an upper bound for the single-step improvement as follows (cf. (16)):

$$\begin{aligned}&\mathcal{L}(\theta^{t+1},\mu^t) - \mathcal{L}(\theta^t,\mu^t) \\ &\le \left\langle \nabla_\lambda L(\lambda(\theta^t),\mu^t), \lambda(\theta^{t+1}) - \lambda(\theta^t)\right\rangle \\ &= (1-\gamma)\left[V^{\pi_{\theta^{t+1}}}(\nabla_\lambda L(\lambda(\theta^t),\mu^t)) - V^{\pi_{\theta^t}}(\nabla_\lambda L(\lambda(\theta^t),\mu^t))\right] \\ &= \sum_{s\in\mathcal{S}} d^{\pi_{\theta^t}}(s)\sum_{a\in\mathcal{A}}\left(\pi_{\theta^{t+1}}(a|s) - \pi_{\theta^t}(a|s)\right)\cdot Q^{\pi_{\theta^{t+1}}}(\nabla_\lambda L(\lambda(\theta^t),\mu^t); s,a) \\ &= \mathbb{E}_{s\sim d^{\pi_{\theta^t}}}\left\langle \pi_{\theta^{t+1}}(\cdot|s) - \pi_{\theta^t}(\cdot|s), Q^{\pi_{\theta^{t+1}}}\left(\nabla_\lambda L(\lambda(\theta^t),\mu^t); s,\cdot\right)\right\rangle.\end{aligned}$$

Appendix B Supplementary Materials for Section 4

Lemma B.1 (Restatement of Lemma 4.3) *The functions $f(\cdot)$ and $g(\cdot)$ are bounded on Λ. Define M_F and M_G such that $|f(\lambda)| \leq M_F$ and $|g(\lambda)| \leq M_G$, for all $\lambda \in \Lambda$. Then, it holds that $|\mathcal{L}(\theta,\mu)| \leq M_L,\ \forall\ \theta \in \Theta, \mu \in U$, where $M_L := M_F + C_0 M_G$. Furthermore, under Assumption 4.2, $\mathcal{L}(\cdot,\mu)$ is ℓ_L-smooth on Θ, for all $\mu \in U$, where $\ell_L := \ell_F + C_0\ell_G$.*

Proof. Being a polytope means that Λ is closed and compact (cf. (26)). Since f is concave and g is convex on $\Lambda \subset \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, they are also continuous. Thus, we have that f and g are bounded on Λ. As $U = [0, C_0]$, it follows that

$$|\mathcal{L}(\theta,\mu)| = |F(\theta) - \mu G(\theta)| = |f(\lambda(\theta)) - \mu g(\lambda(\theta))| \leq |f(\lambda(\theta))| + \mu|g(\lambda(\theta))| \leq M_F + C_0 M_G$$

for all $\mu \in U$. Similarly, as $F(\theta)$ is ℓ_F-smooth and $G(\theta)$ is ℓ_G-smooth, we have that $\mathcal{L}(\theta,\mu) = F(\theta) - \mu G(\theta)$ is $(\ell_F + C_0\ell_G)$-smooth. □

B.1 Proof of Theorem 4.5

Proposition B.2 (Restatement of Proposition 4.4) *Let Assumptions 4.1 and 4.2 hold and assume that $\varepsilon \leq \bar{\varepsilon}$. Then, for every $T > 0$, the iterates $\{(\theta^t, \mu^t)\}_{t=0}^{T-1}$ produced by algorithm (11) with $\eta_1 = 1/\ell_L$ satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^t) - \mathcal{L}(\theta^t,\mu^t)\right] \leq \frac{\mathcal{L}(\theta^\star,\mu^0) - \mathcal{L}(\theta^0,\mu^0)}{\varepsilon T} + 2\varepsilon\ell_L\ell_\Theta^2 + \frac{2\eta_2 M_G^2}{\varepsilon}.$$

Proof. We note that computing the primal update in algorithm (11) is equivalent to solving the following sub-problem (cf. (17)):

$$\begin{aligned} \theta^{t+1} &= \mathcal{P}_\Theta\left(\theta^t + \eta_1\nabla_\theta\mathcal{L}(\theta^t,\mu^t)\right) \\ &= \operatorname*{argmax}_{\theta\in\Theta}\left\{\mathcal{L}(\theta^t,\mu^t) + (\theta-\theta^t)^\top\nabla_\theta\mathcal{L}(\theta^t,\mu^t) - \frac{1}{2\eta_1}\|\theta-\theta^t\|_2^2\right\} \\ &= \operatorname*{argmax}_{\theta\in\Theta}\left\{\mathcal{L}(\theta^t,\mu^t) + (\theta-\theta^t)^\top\nabla_\theta\mathcal{L}(\theta^t,\mu^t) - \frac{\ell_L}{2}\|\theta-\theta^t\|_2^2\right\}. \end{aligned} \tag{28}$$

Since $\mathcal{L}(\theta,\mu)$ is ℓ_L-smooth by Lemma 4.3, we obtain for every $\theta \in \Theta$ that

$$\left|\mathcal{L}(\theta,\mu^t) - \mathcal{L}(\theta^t,\mu^t) - (\theta-\theta^t)^\top\nabla_\theta\mathcal{L}(\theta^t,\mu^t)\right| \leq \frac{\ell_L}{2}\|\theta-\theta^t\|_2^2.$$

Thus, the following ascent property holds:

$$\mathcal{L}(\theta,\mu^t) \geq \mathcal{L}(\theta^t,\mu^t) + (\theta-\theta^t)^\top\nabla_\theta\mathcal{L}(\theta^t,\mu^t) - \frac{\ell_L}{2}\|\theta-\theta^t\|_2^2 \geq \mathcal{L}(\theta,\mu^t) - \ell_L\|\theta-\theta^t\|_2^2. \tag{29}$$

On the basis of (28) and (29), it holds that

$$\begin{aligned} \mathcal{L}(\theta^{t+1},\mu^t) &\geq \mathcal{L}(\theta^t,\mu^t) + (\theta^{t+1}-\theta^t)^\top\nabla_\theta\mathcal{L}(\theta^t,\mu^t) - \frac{\ell_L}{2}\|\theta^{t+1}-\theta^t\|_2^2 \\ &= \max_{\theta\in\Theta}\left\{\mathcal{L}(\theta^t,\mu^t) + (\theta-\theta^t)^\top\nabla_\theta\mathcal{L}(\theta^t,\mu^t) - \frac{\ell_L}{2}\|\theta-\theta^t\|_2^2\right\} \\ &\geq \max_{\theta\in\Theta}\left\{\mathcal{L}(\theta,\mu^t) - \ell_L\|\theta-\theta^t\|_2^2\right\}, \end{aligned} \tag{30}$$

Now, we leverage the local invertibility of $\lambda(\cdot)$ to lower-bound the right-hand side of (30). We define

$$\theta_\varepsilon := \lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left((1-\varepsilon)\lambda(\theta^t) + \varepsilon\lambda(\theta^\star)\right). \tag{31}$$

According to Assumption 4.1, since $\varepsilon \leq \bar{\varepsilon}$, we have $(1-\varepsilon)\lambda(\theta^t) + \varepsilon\lambda(\theta^\star) \in \mathcal{V}_{\lambda(\theta^t)}$. Thus, θ_ε is well-defined and $\theta_\varepsilon \in \mathcal{U}_{\theta^t}$. By definition, the composition of $\lambda : \Theta \to \Lambda$ and $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}} : \mathcal{V}_{\lambda(\theta^t)} \to \mathcal{U}_{\theta^t}$ is the identity map on $\mathcal{V}_{\lambda(\theta^t)}$. Together with the facts that $\mathcal{L}(\theta,\mu) = L(\lambda(\theta),\mu)$ and $L(\cdot,\mu)$ is concave, we have that

$$\begin{aligned}
\mathcal{L}(\theta_\varepsilon,\mu^t) &= L(\lambda(\theta_\varepsilon),\mu^t) \\
&= L\left(\lambda\circ\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left((1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star)\right),\mu^t\right) \\
&= L\left((1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star),\mu^t\right) \\
&\geq (1-\varepsilon)L(\lambda(\theta^t),\mu^t)+\varepsilon L(\lambda(\theta^\star),\mu^t) \\
&= (1-\varepsilon)\mathcal{L}(\theta^t,\mu^t)+\varepsilon\mathcal{L}(\theta^\star,\mu^t).
\end{aligned} \tag{32}$$

Additionally, the Lipschitz continuity of $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}$ implies that

$$\begin{aligned}
\|\theta_\varepsilon-\theta^t\|_2^2 &= \left\|\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left((1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star)\right)-\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left(\lambda(\theta^t)\right)\right\|_2^2 \\
&\leq \ell_\Theta^2\left\|(1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2 \\
&\leq \varepsilon^2\ell_\Theta^2\left\|\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2 \\
&\leq 2\varepsilon^2\ell_\Theta^2,
\end{aligned} \tag{33}$$

where the last inequality uses the diameter of the probability simplex Λ, i.e., $\max_{\lambda_1,\lambda_2\in\Lambda}\|\lambda_1-\lambda_2\|_2 \leq \sqrt{2}$. By substituting θ_ε into (30) and using inequalities (32) and (33), it holds that

$$\begin{aligned}
\mathcal{L}(\theta^{t+1},\mu^t) &\geq \max_{\theta\in\Theta}\left\{\mathcal{L}(\theta,\mu^t)-\ell_L\|\theta-\theta^t\|_2^2\right\} \\
&\geq \mathcal{L}(\theta_\varepsilon,\mu^t)-\ell_L\|\theta_\varepsilon-\theta^t\|_2^2 \\
&\geq (1-\varepsilon)\mathcal{L}(\theta^t,\mu^t)+\varepsilon\mathcal{L}(\theta^\star,\mu^t)-2\varepsilon^2\ell_L\ell_\Theta^2,
\end{aligned}$$

which implies that

$$\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^{t+1},\mu^t) \leq (1-\varepsilon)\left(\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right)+2\varepsilon^2\ell_L\ell_\Theta^2. \tag{34}$$

Consequently, one can obtain the recursion

$$\begin{aligned}
&\mathcal{L}(\theta^\star,\mu^{t+1})-\mathcal{L}(\theta^{t+1},\mu^{t+1}) \\
&= \left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^{t+1},\mu^t)\right]+\left[\mathcal{L}(\theta^\star,\mu^{t+1})-\mathcal{L}(\theta^\star,\mu^t)\right]+\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^{t+1},\mu^{t+1})\right] \\
&\overset{(i)}{\leq} (1-\varepsilon)\left(\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right)+2\varepsilon^2\ell_L\ell_\Theta^2 \\
&\qquad +\left[\mathcal{L}(\theta^\star,\mu^{t+1})-\mathcal{L}(\theta^\star,\mu^t)\right]+\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^{t+1},\mu^{t+1})\right] \\
&\overset{(ii)}{\leq} (1-\varepsilon)\left(\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right)+2\varepsilon^2\ell_L\ell_\Theta^2+2\eta_2M_G^2,
\end{aligned} \tag{35}$$

where we use (34) in (i) . Step (ii) is due to the bound

$$\begin{aligned}
|\mathcal{L}(\theta,\mu^t)-\mathcal{L}(\theta,\mu^{t+1})| &= |(\mu^t-\mu^{t+1})G(\theta)| \\
&= \left|\left[\mu^t-\mathcal{P}_U\left(\mu^t-\eta_2\nabla_\mu\mathcal{L}(\theta^t,\mu^t)\right)\right]G(\theta)\right| \\
&\leq \left|\eta_2\nabla_\mu\mathcal{L}(\theta^t,\mu^t)G(\theta)\right| \\
&= |\eta_2G(\theta^t)G(\theta)| \\
&\leq \eta_2M_G^2,\ \forall\ \theta\in\Theta,
\end{aligned} \tag{36}$$

where the two inequalities above result from the non-expansive property of the projection operator and the boundedness of $G(\theta)$, i.e., $|G(\theta)| \leq M_G$, respectively. Utilizing the recursion (35), we derive that

$$\begin{aligned}
&\mathcal{L}(\theta^\star,\mu^{t+1})-\mathcal{L}(\theta^{t+1},\mu^{t+1}) \\
&\leq (1-\varepsilon)\left(\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right)+2\varepsilon^2\ell_L\ell_\Theta^2+2\eta_2M_G^2 \\
&\leq (1-\varepsilon)^2\left(\mathcal{L}(\theta^\star,\mu^{t-1})-\mathcal{L}(\theta^{t-1},\mu^{t-1})\right)+(1+1-\varepsilon)\left(2\varepsilon^2\ell_L\ell_\Theta^2+2\eta_2M_G^2\right) \\
&\leq (1-\varepsilon)^{t+1}\left(\mathcal{L}(\theta^\star,\mu^0)-\mathcal{L}(\theta^0,\mu^0)\right)+\sum_{i=0}^{t}(1-\varepsilon)^i\left(2\varepsilon^2\ell_L\ell_\Theta^2+2\eta_2M_G^2\right) \\
&= (1-\varepsilon)^{t+1}\left(\mathcal{L}(\theta^\star,\mu^0)-\mathcal{L}(\theta^0,\mu^0)\right)+\frac{1-(1-\varepsilon)^{t+1}}{\varepsilon}\left(2\varepsilon^2\ell_L\ell_\Theta^2+2\eta_2M_G^2\right),
\end{aligned}$$

which is equivalent to

$$
\begin{aligned}
&\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\\
\le\ &(1-\varepsilon)^t\left(\mathcal{L}(\theta^\star,\mu^0)-\mathcal{L}(\theta^0,\mu^0)\right)+\left(1-(1-\varepsilon)^t\right)\left(2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2M_G^2}{\varepsilon}\right),\ \forall\ t\ge 0.
\end{aligned}
$$

Summing the above inequality over $t=0,1,\dots,T-1$ yields that

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]\\
\le\ &\sum_{t=0}^{T-1}(1-\varepsilon)^t\left(\mathcal{L}(\theta^\star,\mu^0)-\mathcal{L}(\theta^0,\mu^0)\right)+\left(1-(1-\varepsilon)^t\right)\left(2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2M_G^2}{\varepsilon}\right)\\
=\ &\frac{1-(1-\varepsilon)^T}{\varepsilon}\left(\mathcal{L}(\theta^\star,\mu^0)-\mathcal{L}(\theta^0,\mu^0)\right)+\left(T-\frac{1-(1-\varepsilon)^T}{\varepsilon}\right)\left(2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2M_G^2}{\varepsilon}\right)\\
\le\ &\frac{1}{\varepsilon}\left(\mathcal{L}(\theta^\star,\mu^0)-\mathcal{L}(\theta^0,\mu^0)\right)+T\left(2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2M_G^2}{\varepsilon}\right).
\end{aligned}
$$

The proof is completed by dividing T on both sides of the inequality. □

Theorem B.3 (Restatement of Theorem 4.5) *Let Assumptions 2.1, 4.1, and 4.2 hold. For every $T\ge(\bar\varepsilon)^{-3}$, we choose $C_0=1+\left(M_F-F(\widetilde\theta)\right)/\xi$, $\mu^0=0$, $\eta_1=1/\ell_L$, and $\eta_2=T^{-2/3}$. Then, the sequence $\left\{\left(\theta^t,\mu^t\right)\right\}_{t=0}^{T-1}$ generated by algorithm (11) converges with the rate $\mathcal{O}\left(T^{-1/3}\right)$, in particular*

$$
\begin{aligned}
&\textit{Optimality Gap: }\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-F(\theta^t)\right]\le\frac{2M_F+M_G^2/2}{T^{2/3}}+\frac{2\ell_L\ell_\Theta^2+2M_G^2}{T^{1/3}},\\
&\textit{Constraint Violation: }\frac{1}{T}\left[\sum_{t=0}^{T-1}G(\theta^t)\right]_+\le\frac{2M_F+M_G^2/2}{T^{2/3}}+\frac{2\ell_L\ell_\Theta^2+2M_G^2+C_0^2/2}{T^{1/3}}.
\end{aligned}
$$

Proof of the optimality gap (20a). By the definition of the Lagrangian function $\mathcal{L}(\theta,\mu)$, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-F(\theta^t)\right]&=\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-\mathcal{L}(\theta^t,\mu^t)-\mu^tG(\theta^t)\right]\\
&=\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-\mathcal{L}(\theta^t,\mu^t)\right]-\frac{1}{T}\sum_{t=0}^{T-1}\mu^tG(\theta^t).
\end{aligned}
\tag{37}
$$

The first term in the right-hand side of (37) can be upper-bounded as

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-\mathcal{L}(\theta^t,\mu^t)\right]&=\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^\star)-\mathcal{L}(\theta^t,\mu^t)\right]\\
&\overset{(i)}{\le}\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]\\
&\overset{(ii)}{\le}\frac{\mathcal{L}(\theta^\star,\mu^0)-\mathcal{L}(\theta^0,\mu^0)}{\varepsilon T}+2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2M_G^2}{\varepsilon}\\
&=\frac{F(\theta^\star)-F(\theta^0)}{\varepsilon T}+2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2M_G^2}{\varepsilon}\\
&\le\frac{2M_F}{\varepsilon T}+2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2M_G^2}{\varepsilon},
\end{aligned}
\tag{38}
$$

where the first equality holds due to strong duality (cf. Lemma 2.2), and step (i) is due to the fact that $\mu^\star=\operatorname{argmin}_{\mu\ge0}\mathcal{L}(\theta^\star,\mu)$. By Proposition 4.4 and Assumption 4.1, step (ii) holds true for all $\varepsilon\le\bar\varepsilon$. Finally, we use the fact that $\mu^0=0$ in the second equality.

Next, we upper-bound the second term in the right-hand side of (37). By the update rule of μ^t in algorithm (11) and the non-expansive property of the projection operator, we obtain that

$$\begin{aligned}
(\mu^{t+1} - \mu)^2 &= \left[\mathcal{P}_U\left(\mu^t - \eta_2 \nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right) - \mu\right]^2 \\
&\le \left[\mu^t - \eta_2 \nabla_\mu \mathcal{L}(\theta^t, \mu^t) - \mu\right]^2 \\
&= (\mu^t - \mu)^2 - 2\eta_2 (\mu^t - \mu) \cdot \nabla_\mu \mathcal{L}(\theta^t, \mu^t) + \left[\eta_2 \nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right]^2 \\
&\le (\mu^t - \mu)^2 + 2\eta_2 (\mu^t - \mu) \cdot G(\theta^t) + (\eta_2 M_G)^2, \quad \forall \mu \in U,
\end{aligned} \tag{39}$$

where the last inequality results from $\nabla_\mu \mathcal{L}(\theta, \mu) = -G(\theta)$ and the boundedness of $G(\theta)$. By setting $\mu = 0$ and rearranging terms, we conclude that

$$-\mu^t G(\theta^t) \le \frac{1}{2\eta_2}\left[(\mu^t)^2 - (\mu^{t+1})^2 + (\eta_2 M_G)^2\right]. \tag{40}$$

We sum both sides of (40) from $t = 0$ to $T - 1$ and divide both sides by T to obtain that

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} -\mu^t G(\theta^t) &\le \frac{1}{2\eta_2 T}\sum_{t=0}^{T-1}\left[(\mu^t)^2 - (\mu^{t+1})^2 + (\eta_2 M_G)^2\right] \\
&\le \frac{1}{2\eta_2 T}\left[(\mu^0)^2 - (\mu^T)^2 + T\cdot(\eta_2 M_G)^2\right] \\
&\le \frac{\eta_2 M_G^2}{2},
\end{aligned} \tag{41}$$

where the last inequality is resulted from dropping the non-positive term $-(\mu^T)^2$ and plugging in $\mu^0 = 0$. By substituting (38) and (41) back into (37), it follows that

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] \le \frac{2M_F}{\varepsilon T} + 2\varepsilon \ell_L \ell_\Theta^2 + \frac{2\eta_2 M_G^2}{\varepsilon} + \frac{\eta_2 M_G^2}{2}.$$

The proof is completed by taking $\eta_2 = T^{-2/3}$ and $\varepsilon = T^{-1/3}$. We note that $T \ge (\bar{\varepsilon})^{-3}$ ensures $\varepsilon \le \bar{\varepsilon}$. □

Proof of the constraint violation (20b). If $\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = 0$, the bound is trivially satisfied. Therefore, from now on, we assume $\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ > 0$, which implies $\sum_{t=0}^{T-1} G(\theta^t) = \left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+$. Define $\widehat{\mu} := \mu^\star + 1 \ge 1$ as $\mu^\star \ge 0$. By the boundedness of $\mu^\star$ (cf. Lemma 2.2), we have that

$$\widehat{\mu} = \mu^\star + 1 \le \frac{F(\theta^\star) - F(\widetilde{\theta})}{\xi} + 1 \le \frac{M_F - F(\widetilde{\theta})}{\xi} + 1 = C_0,$$

which implies $\widehat{\mu} \in U$. Thus, it follows that

$$\begin{aligned}
\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ &= (\widehat{\mu} - \mu^\star)\cdot\frac{1}{T}\sum_{t=0}^{T-1} G(\theta^t) \\
&= (\mu^\star - \widehat{\mu})\cdot\frac{1}{T}\sum_{t=0}^{T-1}\nabla_\mu \mathcal{L}(\theta^t, \mu^t) \\
&\le \max_{\mu\in U}\left\{(\mu^\star - \mu)\cdot\frac{1}{T}\sum_{t=0}^{T-1}\nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right\},
\end{aligned} \tag{42}$$

where we used the fact that $\widehat{\mu} - \mu^\star = 1$ in the first step. To upper-bound the last line in (42), we note that

$$\begin{aligned}
(\mu^\star - \mu)\cdot\nabla_\mu \mathcal{L}(\theta^t, \mu^t) &= (\mu^\star - \mu^t + \mu^t - \mu)\cdot\nabla_\mu \mathcal{L}(\theta^t, \mu^t) \\
&= \left[(\mu^\star - \mu^t)\cdot\nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right] + \left[(\mu^t - \mu)\cdot\nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right] \\
&\overset{(i)}{=} \left[\mathcal{L}(\theta^t, \mu^\star) - \mathcal{L}(\theta^t, \mu^t)\right] - \left[(\mu^t - \mu)\cdot G(\theta^t)\right] \\
&\le \left[\mathcal{L}(\theta^\star, \mu^\star) - \mathcal{L}(\theta^t, \mu^t)\right] - \left[(\mu^t - \mu)\cdot G(\theta^t)\right] \\
&\overset{(ii)}{\le} \left[\mathcal{L}(\theta^\star, \mu^\star) - \mathcal{L}(\theta^t, \mu^t)\right] + \frac{(\mu^t - \mu)^2 - (\mu^{t+1} - \mu)^2}{2\eta_2} + \frac{\eta_2 M_G^2}{2},
\end{aligned} \tag{43}$$

where we use the linearity of $\mathcal{L}(\theta,\mu)$ with respect to μ in (i). The first inequality follows from the fact that $\theta^\star$ maximizes $\mathcal{L}(\cdot,\mu^\star)$ and inequality (ii) follows from rearranging the terms in (39).

Summing both sides of (43) over $t=0,\ldots,T-1$, dividing them by T, and plugging into (42) yield that

$$
\begin{aligned}
&\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ \\
&\le \max_{\mu\in U}\left\{(\mu^\star-\mu)\cdot\frac{1}{T}\sum_{t=0}^{T-1}\nabla_\mu\mathcal{L}(\theta^t,\mu^t)\right\} \\
&\le \max_{\mu\in U}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^\star)-\mathcal{L}(\theta^t,\mu^t)\right]+\frac{1}{T}\sum_{t=0}^{T-1}\frac{(\mu^t-\mu)^2-(\mu^{t+1}-\mu)^2}{2\eta_2}+\frac{\eta_2 M_G^2}{2}\right\} \\
&\le \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^\star)-\mathcal{L}(\theta^t,\mu^t)\right]+\frac{\eta_2 M_G^2}{2}+\max_{\mu\in U}\left\{\frac{(\mu^0-\mu)^2-(\mu^T-\mu)^2}{2\eta_2 T}\right\} \\
&\overset{(i)}{\le} \frac{2M_F}{\varepsilon T}+2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2 M_G^2}{\varepsilon}+\frac{\eta_2 M_G^2}{2}+\max_{\mu\in U}\left\{\frac{(\mu^0-\mu)^2}{2\eta_2 T}\right\} \\
&\overset{(ii)}{\le} \frac{2M_F}{\varepsilon T}+2\varepsilon\ell_L\ell_\Theta^2+\frac{2\eta_2 M_G^2}{\varepsilon}+\frac{\eta_2 M_G^2}{2}+\frac{C_0^2}{2\eta_2 T},
\end{aligned}
\tag{44}
$$

where we upper-bound $(1/T)\cdot\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^\star)-\mathcal{L}(\theta^t,\mu^t)\right]$ with (38) in (i) and drop the non-positive term $-(\mu^T-\mu)^2$. Step (ii) holds due to $\mu^0=0$ and $\mu\le C_0,\ \forall\mu\in U$. The proof is completed by substituting $\eta_2=T^{-2/3}$ and $\varepsilon=T^{-1/3}$ into (44). □

B.2 Proof of Theorem 4.7

Proposition B.4 (Restatement of Proposition 4.6) *Let Assumptions 4.1 and 4.2 hold. Suppose that $f(\cdot)$ is σ-strongly concave w.r.t. λ on Λ. Then, for every $T>0$, the iterates $\left\{(\theta^t,\mu^t)\right\}_{t=0}^{T-1}$ produced by algorithm (11) with $\eta_1=1/\ell_L$ satisfy*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]\le\frac{\mathcal{L}(\theta^T,\mu^{T-1})-\mathcal{L}(\theta^0,\mu^0)}{\tilde\varepsilon T}+\frac{\eta_2 M_G^2}{\tilde\varepsilon},
$$

where $\tilde\varepsilon:=\min\{\bar\varepsilon,\sigma/\left(\sigma+2\ell_\Theta^2\ell_L\right)\}$.

Proof. We begin with (30):

$$
\mathcal{L}(\theta^{t+1},\mu^t)\ge\max_{\theta\in\Theta}\left\{\mathcal{L}(\theta,\mu^t)-\ell_L\|\theta-\theta^t\|_2^2\right\}. \tag{45}
$$

For $\varepsilon\le\bar\varepsilon$, we define $\theta_\varepsilon:=\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left((1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star)\right)$ similarly to (31). Combining the definition of $\mathcal{L}(\theta,\mu)=L(\lambda(\theta),\mu)$ with the fact that $L(\cdot,\mu)$ is σ-strongly concave in λ, which is due to the σ-strongly concavity of $f(\cdot)$ and the convexity of $g(\cdot)$, we have that

$$
\begin{aligned}
\mathcal{L}(\theta_\varepsilon,\mu^t)&=L(\lambda(\theta_\varepsilon),\mu^t)\\
&=L\left(\lambda\circ\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left((1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star)\right),\mu^t\right)\\
&=L\left((1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star),\mu^t\right)\\
&\ge(1-\varepsilon)L(\lambda(\theta^t),\mu^t)+\varepsilon L(\lambda(\theta^\star),\mu^t)+\frac{\sigma}{2}\varepsilon(1-\varepsilon)\left\|\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2\\
&=(1-\varepsilon)\mathcal{L}(\theta^t,\mu^t)+\varepsilon\mathcal{L}(\theta^\star,\mu^t)+\frac{\sigma}{2}\varepsilon(1-\varepsilon)\left\|\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2.
\end{aligned}
\tag{46}
$$

By Assumption 4.1, the Lipschitz continuity of $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}$ implies that

$$
\begin{aligned}
\|\theta_\varepsilon-\theta^t\|_2^2&=\left\|\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left((1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star)\right)-\lambda^{-1}_{\mathcal{V}_{\lambda(\theta^t)}}\left(\lambda(\theta^t)\right)\right\|_2^2\\
&\le\ell_\Theta^2\left\|(1-\varepsilon)\lambda(\theta^t)+\varepsilon\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2\\
&\le\varepsilon^2\ell_\Theta^2\left\|\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2.
\end{aligned}
\tag{47}
$$

Substitute θ_ε into the right-hand side of (45), we have that

$$\begin{aligned}\mathcal{L}(\theta^{t+1},\mu^t) &\geq \max_{\theta\in\Theta}\left\{\mathcal{L}(\theta,\mu^t)-\ell_L\|\theta-\theta^t\|_2^2\right\}\\ &\geq \max_{0\leq\varepsilon\leq\bar{\varepsilon}}\left\{\mathcal{L}(\theta_\varepsilon,\mu^t)-\ell_L\|\theta_\varepsilon-\theta^t\|_2^2\right\}\\ &\geq \max_{0\leq\varepsilon\leq\bar{\varepsilon}}\left\{(1-\varepsilon)\mathcal{L}(\theta^t,\mu^t)+\varepsilon\mathcal{L}(\theta^\star,\mu^t)+\left(\frac{\sigma}{2}\varepsilon(1-\varepsilon)-\varepsilon^2\ell_L\ell_\Theta^2\right)\left\|\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2\right\},\end{aligned}$$

where we use (46) and (47) in the last inequality. Consequently,

$$\begin{aligned}\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^{t+1},\mu^t)\leq \min_{0\leq\varepsilon\leq\bar{\varepsilon}}\Big\{&(1-\varepsilon)\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]\\ &-\left(\frac{\sigma\varepsilon(1-\varepsilon)}{2}-\varepsilon^2\ell_L\ell_\Theta^2\right)\left\|\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2\Big\}.\end{aligned}\tag{48}$$

We note that

$$\frac{\sigma\varepsilon(1-\varepsilon)}{2}-\varepsilon^2\ell_L\ell_\Theta^2\geq 0,\ \text{ if }\ 0\leq\varepsilon\leq\frac{\sigma}{\left(\sigma+2\ell_\Theta^2\ell_L\right)}.\tag{49}$$

By letting $\tilde{\varepsilon}\coloneqq\min\{\bar{\varepsilon},\sigma/\left(\sigma+2\ell_\Theta^2\ell_L\right)\}\leq\bar{\varepsilon}$, it follows from (48) that

$$\begin{aligned}&\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^{t+1},\mu^t)\\ &\leq(1-\tilde{\varepsilon})\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]-\left(\frac{\sigma\tilde{\varepsilon}(1-\tilde{\varepsilon})}{2}-\tilde{\varepsilon}^2\ell_L\ell_\Theta^2\right)\left\|\lambda(\theta^\star)-\lambda(\theta^t)\right\|_2^2\\ &\leq(1-\tilde{\varepsilon})\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right],\end{aligned}\tag{50}$$

where the second inequality results from (49). Now, we rearrange terms in (50) to obtain that

$$\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^{t+1},\mu^t)\leq\frac{1-\tilde{\varepsilon}}{\tilde{\varepsilon}}\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right],$$

which implies that

$$\begin{aligned}\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)&=\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^{t+1},\mu^t)\right]+\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]\\ &\leq\left(\frac{1-\tilde{\varepsilon}}{\tilde{\varepsilon}}+1\right)\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]\\ &=\frac{1}{\tilde{\varepsilon}}\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right].\end{aligned}$$

Summing it over $t=0,\ldots,T-1$, we have that

$$\begin{aligned}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]&\leq\frac{1}{\tilde{\varepsilon}}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^t,\mu^t)\right]\\ &=\frac{1}{\tilde{\varepsilon}}\left(\mathcal{L}(\theta^T,\mu^{T-1})-\mathcal{L}(\theta^0,\mu^0)+\sum_{t=0}^{T-2}\left[\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^{t+1},\mu^{t+1})\right]\right)\\ &\overset{(i)}{\leq}\frac{1}{\tilde{\varepsilon}}\left[\mathcal{L}(\theta^T,\mu^{T-1})-\mathcal{L}(\theta^0,\mu^0)+(T-1)\eta_2M_G^2\right],\end{aligned}$$

where we use (36) to bound the difference $\mathcal{L}(\theta^{t+1},\mu^t)-\mathcal{L}(\theta^{t+1},\mu^{t+1})$ in (i). The proof is completed by dividing T on both sides of the inequality. □

Theorem B.5 (Restatement of Theorem 4.7) *Let Assumptions 2.1, 4.1, and 4.2 hold. Suppose that $f(\cdot)$ is σ-strongly concave w.r.t. λ on Λ. For every $T>0$, we choose $C_0=1+\left(M_F-F(\widetilde{\theta})\right)/\xi$, $\mu^0=0$, $\eta_1=1/\ell_L$, and $\eta_2=T^{-1/2}$. Then, the sequence $\left\{\left(\theta^t,\mu^t\right)\right\}_{t=0}^{T-1}$ generated by algorithm (11) converges with the rate $\mathcal{O}\left(T^{-1/2}\right)$, in particular*

$$\textit{Optimality Gap: }\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-F(\theta^t)\right]\leq\frac{M_L+M_F}{\tilde{\varepsilon}T}+\left(\frac{M_G^2}{\tilde{\varepsilon}}+\frac{M_G^2}{2}\right)\frac{1}{\sqrt{T}},$$

$$\textit{Constraint Violation: }\frac{1}{T}\left[\sum_{t=0}^{T-1}G(\theta^t)\right]_+\leq\frac{M_L+M_F}{\tilde{\varepsilon}T}+\left(\frac{M_G^2}{\tilde{\varepsilon}}+\frac{M_G^2+C_0^2}{2}\right)\frac{1}{\sqrt{T}},$$

where $\tilde{\varepsilon}\coloneqq\min\{\bar{\varepsilon},\sigma/\left(\sigma+2\ell_\Theta^2\ell_L\right)\}$.

Proof of the optimality gap (22a). We follow the same proof as the concave case (cf. (20a) in Theorem 4.5), except for inequality (38). In step (ii) of (38), we use Proposition 4.6 instead of Proposition 4.4. This gives rise to

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-\mathcal{L}(\theta^t,\mu^t)\right] &\le \frac{\mathcal{L}(\theta^T,\mu^{T-1})-\mathcal{L}(\theta^0,\mu^0)}{\tilde{\varepsilon}T}+\frac{\eta_2 M_G^2}{\tilde{\varepsilon}} \\ &= \frac{\mathcal{L}(\theta^T,\mu^{T-1})-F(\theta^0)}{\tilde{\varepsilon}T}+\frac{\eta_2 M_G^2}{\tilde{\varepsilon}} \\ &\le \frac{M_L+M_F}{\tilde{\varepsilon}T}+\frac{\eta_2 M_G^2}{\tilde{\varepsilon}},\end{aligned} \tag{51}$$

where we use $\mu^0=0$ in the second step. Following (37) and (41), we conclude that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-F(\theta^t)\right] &= \frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star)-\mathcal{L}(\theta^t,\mu^t)\right]-\frac{1}{T}\sum_{t=0}^{T-1}\mu^t G(\theta^t) \\ &\le \frac{M_L+M_F}{\tilde{\varepsilon}T}+\frac{\eta_2 M_G^2}{\tilde{\varepsilon}}+\frac{\eta_2 M_G^2}{2}.\end{aligned}$$

The proof is completed by taking $\eta_2=T^{-1/2}$. □

Proof of the constraint violation (22b). We follow the same lines as in the proof of (20b) for the concave case. By substituting (51) into (44), it holds that

$$\begin{aligned}\frac{1}{T}\left[\sum_{t=0}^{T-1}G(\theta^t)\right]_+ &\le \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star,\mu^\star)-\mathcal{L}(\theta^t,\mu^t)\right]+\frac{\eta_2 M_G^2}{2}+\max_{\mu\in U}\left\{\frac{(\mu^0-\mu)^2-(\mu^T-\mu)^2}{2\eta_2 T}\right\} \\ &\le \frac{M_L+M_F}{\tilde{\varepsilon}T}+\frac{\eta_2 M_G^2}{\tilde{\varepsilon}}+\frac{\eta_2 M_G^2}{2}+\max_{\mu\in U}\left\{\frac{(\mu^0-\mu)^2}{2\eta_2 T}\right\} \\ &\le \frac{M_L+M_F}{\tilde{\varepsilon}T}+\frac{\eta_2 M_G^2}{\tilde{\varepsilon}}+\frac{\eta_2 M_G^2}{2}+\frac{C_0^2}{2\eta_2 T},\end{aligned}$$

which, together with $\eta_2=T^{-1/2}$, completes the proof. □

Appendix C Supplementary Materials for Section 5

In this section, we formally state the Primal-dual Policy Gradient-Zero Algorithm (PDPG-0) and Theorem 5.1. First, we note that the pessimistic problem (23) is equivalent to

$$\max_{\theta\in\Theta} F(\theta) \quad \text{s.t.} \quad G_\delta(\theta)\le 0, \tag{52}$$

where $G_\delta(\theta):=G(\theta)+\delta$. Suppose that $\delta<\xi$, i.e., the pessimistic term is smaller than the strict feasibility of the Slater point. This implies that

$$|G_\delta(\theta)|\le M_G+\xi, \text{ and } G_\delta(\widetilde{\theta})\le-(\xi-\delta)<0. \tag{53}$$

This gives the constraint upper bound and slackness for the pessimistic problem (52).

The PDPG-0 method is simply a variate of algorithm (11) applied to the new problem (52), i.e.,

$$\theta^{t+1}=\mathcal{P}_\Theta\left(\theta^t+\eta_1\nabla_\theta\mathcal{L}_\delta(\theta^t,\mu^t)\right),\ \mu^{t+1}=\mathcal{P}_U\left(\mu^t-\eta_2\nabla_\mu\mathcal{L}_\delta(\theta^t,\mu^t)\right),\ \text{for } t=0,1,2,\ldots, \tag{54}$$

where $\mathcal{L}_\delta(\theta,\mu):=F(\theta)-\mu G_\delta(\theta)$ is the Lagrangian function for (52).

Theorem C.1 (Restatement of Theorem 5.1) *Let Assumptions 2.1, 4.1, and 4.2 hold.*

(I) *For fixed $T>0$, let $\delta=\mathcal{O}(T^{-1/3})$ be the solution to the equation*

$$\frac{2M_F+(M_G+\xi)^2/2}{T^{2/3}}+\frac{2\ell_L\ell_\Theta^2+2(M_G+\xi)^2+C_0^2/2}{T^{1/3}}-\delta=0,$$

where $C_0 = 1 + \left(M_F - F(\widetilde{\theta})\right)/(\xi - \delta)$. *For* $T > 0$ *such that* $\delta < \xi$, *choose* $\mu^0 = 0$, $\eta_1 = 1/\ell_L$, *and* $\eta_2 = T^{-2/3}$. *Then, the sequence* $\left\{(\theta^t, \mu^t)\right\}_{t=0}^{T-1}$ *generated by algorithm* (54) *satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] \le \frac{2\delta M_F}{\xi} + \frac{2M_F + (M_G+\xi)^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2(M_G+\xi)^2}{T^{1/3}},$$

$$\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = 0.$$

(II) *Assume that* $f(\cdot)$ *is* σ*-strongly concave w.r.t.* λ *on* Λ. *For fixed* $T > 0$, *let* $\delta = \mathcal{O}(T^{-1/2})$ *be the solution to the equation*

$$\frac{M_L + M_F + C_0\xi}{\tilde{\varepsilon}T} + \left(\frac{(M_G+\xi)^2}{\tilde{\varepsilon}} + \frac{(M_G+\xi)^2 + C_0^2}{2}\right)\frac{1}{\sqrt{T}} - \delta = 0,$$

where $C_0 = 1 + \left(M_F - F(\widetilde{\theta})\right)/(\xi - \delta)$. *For* $T > 0$ *such that* $\delta < \xi$, *choose* $\mu^0 = 0$, $\eta_1 = 1/\ell_L$, *and* $\eta_2 = T^{-1/2}$. *Then, the sequence* $\left\{(\theta^t, \mu^t)\right\}_{t=0}^{T-1}$ *generated by algorithm* (54) *satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] \le \frac{2\delta M_F}{\xi} + \frac{M_L + M_F + C_0\xi}{\tilde{\varepsilon}T} + \left(\frac{(M_G+\xi)^2}{\tilde{\varepsilon}} + \frac{(M_G+\xi)^2}{2}\right)\frac{1}{\sqrt{T}},$$

$$\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = 0.$$

Proof. We begin with general arguments that apply to both cases ***(I)*** and ***(II)***. Let $\theta_\delta^\star$ be an optimal solution to the pessimistic problem (52). Then,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] &= \frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta_\delta^\star)\right] + \frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta_\delta^\star) - F(\theta^t)\right] \\ &= \left[F(\theta^\star) - F(\theta_\delta^\star)\right] + \frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta_\delta^\star) - F(\theta^t)\right].\end{aligned} \tag{55}$$

To upper-bound the first term in (55), we define a feasible point θ_δ to (52) through the state-action visitation distribution such that

$$\lambda(\theta_\delta) = \frac{\xi - \delta}{\xi}\lambda(\theta^\star) + \frac{\delta}{\xi}\lambda(\widetilde{\theta}), \tag{56}$$

where we assume $0 < \delta < \xi$. We remark that the policy corresponds to $\lambda(\theta_\delta)$ is unique and given by

$$\pi_{\theta_\delta}(a|s) = \frac{\lambda(\theta_\delta; s, a)}{\sum_{a'\in\mathcal{A}}\lambda(\theta_\delta; s, a')}.$$

In contrast, due to the assumption of over-parameterization (cf. Assumption 4.1), θ_δ may not be unique. It suffices to choose one such θ_δ that satisfies (56). The feasibility of θ_δ can be verified as follows:

$$\begin{aligned}G_\delta(\theta_\delta) &= g\left(\lambda(\theta_\delta)\right) + \delta \\ &= g\left(\frac{\xi-\delta}{\xi}\lambda(\theta^\star) + \frac{\delta}{\xi}\lambda(\widetilde{\theta})\right) + \delta \\ &\overset{(i)}{\le} \frac{\xi-\delta}{\xi}g\left(\lambda(\theta^\star)\right) + \frac{\delta}{\xi}g\left(\lambda(\widetilde{\theta})\right) + \delta \\ &\overset{(ii)}{\le} 0 + \frac{\delta}{\xi}\cdot(-\xi) + \delta \\ &= 0,\end{aligned}$$

where (i) follows from the concavity of $g(\cdot)$ and (ii) uses the feasibility of $\theta^\star$ to (9) as well as Assumption 2.1. This proves the feasiblity of θ_δ to (52), and thus implies $F(\theta_\delta^\star) \ge F(\theta_\delta)$.

Consequently,

$$
\begin{aligned}
F(\theta^\star) - F(\theta^\star_\delta) &\le F(\theta^\star) - F(\theta_\delta) \\
&= f(\lambda(\theta^\star)) - f(\lambda(\theta_\delta)) \\
&= f(\lambda(\theta^\star)) - f\left(\frac{\xi-\delta}{\xi}\lambda(\theta^\star) + \frac{\delta}{\xi}\lambda(\widetilde{\theta})\right) \\
&\le f(\lambda(\theta^\star)) - \left(\frac{\xi-\delta}{\xi} f(\lambda(\theta^\star)) + \frac{\delta}{\xi} f\left(\lambda(\widetilde{\theta})\right)\right) \\
&= \frac{\delta}{\xi}\left[f(\lambda(\theta^\star)) - f\left(\lambda(\widetilde{\theta})\right)\right] \\
&\le \frac{2\delta M_F}{\xi}.
\end{aligned}
\tag{57}
$$

By (53), for every $\delta < \xi$, choosing $C_0 = 1 + \left(M_F - F(\widetilde{\theta})\right)/(\xi - \delta)$ ensures that the optimal dual variable of problem (52) belongs to the dual feasible region $U = [0, C_0]$ (cf. Lemma 2.2).

(I) When $f(\cdot)$ is a general convex function, we apply Theorem 4.5 to obtain that

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star_\delta) - F(\theta^t)\right] &\le \frac{2M_F + (M_G+\xi)^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2(M_G+\xi)^2}{T^{1/3}}, \\
\frac{1}{T}\left[\sum_{t=0}^{T-1} G_\delta(\theta^t)\right]_+ &\le \frac{2M_F + (M_G+\xi)^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2(M_G+\xi)^2 + C_0^2/2}{T^{1/3}}.
\end{aligned}
$$

where the term $M_G + \xi$ results from the upper bound for $|G_\delta(\cdot)|$ in (53). Therefore, together with (55) and (57), we have the following optimality gap for (9):

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] &\le \frac{2\delta M_F}{\xi} + \frac{2M_F + (M_G+\xi)^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2(M_G+\xi)^2}{T^{1/3}} \\
&= \mathcal{O}(\delta) + \mathcal{O}\left(T^{-1/3}\right).
\end{aligned}
\tag{58}
$$

For the constraint violation, we have that

$$
\begin{aligned}
\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ &= \frac{1}{T}\left[\sum_{t=0}^{T-1}\left(G(\theta^t)+\delta\right) - \delta\right]_+ \\
&= \left[\frac{1}{T}\left[\sum_{t=0}^{T-1}\left(G(\theta^t)+\delta\right)\right]_+ - \delta\right]_+ \\
&= \left[\frac{1}{T}\left[\sum_{t=0}^{T-1} G_\delta(\theta^t)\right]_+ - \delta\right]_+ \\
&\le \left[\frac{2M_F + (M_G+\xi)^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2(M_G+\xi)^2 + C_0^2/2}{T^{1/3}} - \delta\right]_+.
\end{aligned}
\tag{59}
$$

By choosing δ such that

$$
\frac{2M_F + (M_G+\xi)^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2(M_G+\xi)^2 + C_0^2/2}{T^{1/3}} - \delta = 0,
\tag{60}
$$

the constraint violation (59) becomes 0. As (60) implies $\delta = \mathcal{O}(T^{-1/3})$, the convergence rate of the optimality gap (58) is $\mathcal{O}(T^{-1/3})$. Finally, we remark that the requirement $\delta < \xi$ is naturally satisfied when T is reasonably large.

(II) When $f(\cdot)$ is σ-strongly concave, we apply Theorem 4.7 to obtain that

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] &\le \frac{M_L + M_F + C_0\xi}{\tilde{\varepsilon}T} + \left(\frac{(M_G+\xi)^2}{\tilde{\varepsilon}} + \frac{(M_G+\xi)^2}{2}\right)\frac{1}{\sqrt{T}}, \\
\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ &\le \frac{M_L + M_F + C_0\xi}{\tilde{\varepsilon}T} + \left(\frac{(M_G+\xi)^2}{\tilde{\varepsilon}} + \frac{(M_G+\xi)^2 + C_0^2}{2}\right)\frac{1}{\sqrt{T}},
\end{aligned}
$$

where the terms $C_0\xi$ and $M_G + \xi$ result from the upper bound for $G_\delta(\cdot)$ in (53). Together with (55) and (57), we derive the optimality gap such that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] &\leq \frac{2\delta M_F}{\xi} + \frac{M_L + M_F + C_0\xi}{\tilde{\varepsilon}T} + \left(\frac{(M_G+\xi)^2}{\tilde{\varepsilon}} + \frac{(M_G+\xi)^2}{2}\right)\frac{1}{\sqrt{T}} \\ &= \mathcal{O}(\delta) + \mathcal{O}\left(T^{-1/2}\right).\end{aligned} \tag{61}$$

For the constraint violation, similarly to (59), we have that

$$\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ \leq \left[\frac{M_L + M_F + C_0\xi}{\tilde{\varepsilon}T} + \left(\frac{(M_G+\xi)^2}{\tilde{\varepsilon}} + \frac{(M_G+\xi)^2 + C_0^2}{2}\right)\frac{1}{\sqrt{T}} - \delta\right]_+. \tag{62}$$

We choose δ such that

$$\frac{M_L + M_F + C_0\xi}{\tilde{\varepsilon}T} + \left(\frac{(M_G+\xi)^2}{\tilde{\varepsilon}} + \frac{(M_G+\xi)^2 + C_0^2}{2}\right)\frac{1}{\sqrt{T}} - \delta = 0, \tag{63}$$

which guarantees the zero constraint violation (62). As (63) implies $\delta = \mathcal{O}(T^{-1/2})$, the convergence rate of the optimality gap (61) is $\mathcal{O}(T^{-1/2})$. □

Appendix D Discussions About Assumption 4.1

To leverage the hidden convexity of problem (9) with respect to λ, it is natural to assume that there exists some desirable correspondence between $\lambda(\theta)$ and θ. However, as briefly discussed in Section 4, requiring such correspondence to be one-to-one or invertible is too restrictive. Although we can show that a one-to-one correspondence indeed exists under the direct parameterization and that the inverse map is Lipschitz continuous as long as there is a universal positive lower bound for the state visitation distribution $d^\pi(\cdot)$ (cf. Lemma D.1), this is not the case for many other parameterizations. The soft-max policy, defined as

$$\pi_\theta(a|s) = \frac{\exp(\theta_{sa})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{sa'})},\ \forall\ (s,a)\in\mathcal{S}\times\mathcal{A}, \tag{64}$$

serves as a counterexample. For a fixed vector $\theta_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, consider the set of parameters

$$\{\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mid \theta_{sa} = (\theta_0)_{sa} + k,\ \forall\ (s,a)\in\mathcal{S}\times\mathcal{A},\ \forall\ k\in\mathbb{R}\}.$$

Then, it is clear that all parameters in the set correspond to the same policy π_{θ_0}. Thus, a one-to-one correspondence does not exist. This motivates Assumption 4.1, which only requires the local existence of a continuous inverse λ^{-1}. Assumption 4.1 is able to accommodate the soft-max policy defined in (64).

Lemma D.1 (Lipschitz continuity of λ^{-1} under direct parameterization)

Suppose that $d_0 := \min_{s\in\mathcal{S},\pi\in\Pi} d^\pi(s) > 0$[2]*. For every two discounted state-action visitation distributions $\lambda_1, \lambda_2 \in \Lambda$, it holds that*

$$\left\|\lambda^{-1}(\lambda_1) - \lambda^{-1}(\lambda_2)\right\|_2 \leq \frac{\sqrt{2(1+|\mathcal{A}|)}}{d_0}\|\lambda_1 - \lambda_2\|_2,$$

where $\lambda^{-1}(\cdot)$ maps a discounted state-action visitation distribution to the corresponding policy, defined as $\pi(a|s) = \left[\lambda^{-1}(\lambda^\pi)\right]_{s,a} := \lambda^\pi(s,a)/\sum_{a'\in\mathcal{A}}\lambda^\pi(s,a')$.

Proof. Let $d_1(s) = \sum_a \lambda_1(s,a)$ and $d_2(s) = \sum_a \lambda_2(s,a)$ be the corresponding state visitation distributions. Then,

$$\begin{aligned}\left[\lambda^{-1}(\lambda_1)\right]_{s,a} - \left[\lambda^{-1}(\lambda_2)\right]_{s,a} &= \frac{\lambda_1(s,a)}{d_1(s)} - \frac{\lambda_2(s,a)}{d_2(s)} \\ &= \frac{1}{d_1(s)}\left(\lambda_1(s,a) - \lambda_2(s,a) + \frac{d_2(s) - d_1(s)}{d_2(s)}\cdot\lambda_2(s,a)\right).\end{aligned}$$

[2]Since $d^\pi(s) \geq (1-\gamma)\rho(s)$, this assumption is satisfied when there is an exploratory initial distribution, i.e., $\rho_0 := \min_{s\in\mathcal{S}}\rho(s) > 0$.

Therefore, one can compute

$$
\begin{aligned}
&\left\|\lambda^{-1}(\lambda_1)-\lambda^{-1}(\lambda_2)\right\|_2^2\\
=&\sum_{s,a}\left(\left[\lambda^{-1}(\lambda_1)\right]_{s,a}-\left[\lambda^{-1}(\lambda_2)\right]_{s,a}\right)^2\\
=&\sum_{s}\frac{1}{\left[d_1(s)\right]^2}\sum_{a}\left(\lambda_1(s,a)-\lambda_2(s,a)+\left[d_2(s)-d_1(s)\right]\cdot\frac{\lambda_2(s,a)}{d_2(s)}\right)^2\\
\le&\sum_{s}\frac{2}{\left[d_1(s)\right]^2}\left(\sum_{a}\left[\lambda_1(s,a)-\lambda_2(s,a)\right]^2+\left(\left[d_2(s)-d_1(s)\right]\cdot\frac{\lambda_2(s,a)}{d_2(s)}\right)^2\right),
\end{aligned}
\tag{65}
$$

where the last line follows from the inequality $(x+y)^2\le 2x^2+2y^2$. For the second term inside the summation, we have that

$$
\begin{aligned}
\sum_{a}\left(\left[d_2(s)-d_1(s)\right]\cdot\frac{\lambda_2(s,a)}{d_2(s)}\right)^2&=\left[d_2(s)-d_1(s)\right]^2\cdot\left\|\frac{\lambda_2(s,\cdot)}{d_2(s)}\right\|_2^2\\
&\overset{i}{\le}\left[d_2(s)-d_1(s)\right]^2\cdot\left\|\frac{\lambda_2(s,\cdot)}{d_2(s)}\right\|_1^2\\
&=\left[d_2(s)-d_1(s)\right]^2\\
&=\left[\sum_{a}\lambda_2(s,a)-\sum_{a}\lambda_1(s,a)\right]^2\\
&\le|\mathcal{A}|\cdot\sum_{a}\left[\lambda_2(s,a)-\lambda_1(s,a)\right]^2,
\end{aligned}
\tag{66}
$$

where (i) is due to $\|\cdot\|_2\le\|\cdot\|_1$ and the last step follows from the Cauchy-Schwarz inequality.

By substituting (66) into (65) and noting that $d^\pi(s)\ge d_0,\ \forall s\in\mathcal{S},\pi\in\Pi$, it holds that

$$
\begin{aligned}
\left\|\lambda^{-1}(\lambda_1)-\lambda^{-1}(\lambda_2)\right\|_2^2&\le\sum_{s}\frac{2(1+|\mathcal{A}|)}{\left[d_1(s)\right]^2}\left(\sum_{a}\left[\lambda_1(s,a)-\lambda_2(s,a)\right]^2\right)\\
&\le\frac{2(1+|\mathcal{A}|)}{d_0^2}\sum_{s,a}\left[\lambda_1(s,a)-\lambda_2(s,a)\right]^2\\
&\le\frac{2(1+|\mathcal{A}|)}{d_0^2}\left\|\lambda_1-\lambda_2\right\|_2^2.
\end{aligned}
$$

The proof is completed by taking square root of both sides of the inequality. □

Appendix E Discussions About Assumption 4.2

In this section, we validate Assumption 4.2 for both the general soft-max parameterization (8) and the direct parameterization.

E.1 General Soft-max Parameterization

The following result is cited from [17]. In short, it states that, under mild conditions on the smoothness of $\psi(\cdot;s,a)$, $f(\cdot)$, and $g(\cdot)$, Assumption 4.2 is satisfied. We remark that neither concavity nor convexity of the objective function is required for Proposition E.1.

Proposition E.1 (Smoothness of $f(\lambda(\theta))$ w.r.t. θ, [17]) *Under the general soft-max parameterization (8), suppose that $\psi(\cdot;s,a)$ is twice differentiable for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and there exist $\ell_{\psi,1},\ell_{\psi,2}>0$ such that*

$$
\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sup_{\theta}\left\|\nabla_\theta\psi(\theta;s,a)\right\|\le\ell_{\psi,1}\quad\textit{and}\quad\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sup_{\theta}\left\|\nabla_\theta^2\psi(\theta;s,a)\right\|\le\ell_{\psi,2}.
$$

Assume that $f(\lambda)$ has a bounded and Lipschitz gradient in Λ, namely, there exist $\ell_{f,1}, \ell_{f,2} > 0$ such that

$$\|\nabla_\lambda f(\lambda)\|_\infty \le \ell_{f,1}, \quad \|\nabla_\lambda f(\lambda) - \nabla_\lambda f(\lambda')\|_\infty \le \ell_{f,2}\|\lambda - \lambda'\|_2, \quad \forall\ \lambda, \lambda' \in \Lambda.$$

The following statements hold:

(I) *For every $\theta \in \Theta$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, it holds that*

$$\begin{cases} \|\nabla_\theta \log \pi_\theta(a \mid s)\|_2 \le 2\ell_{\psi,1}, \\ \left\|\nabla_\theta^2 \log \pi_\theta(a \mid s)\right\| \le 2\left(\ell_{\psi,2} + \ell_{\psi,1}^2\right), \end{cases} \quad \textit{and} \quad \|\nabla_\theta f(\lambda(\theta))\| \le \frac{2\ell_{\psi,1} \cdot \ell_{f,1}}{(1-\gamma)^2}.$$

(II) *For every $\theta_1, \theta_2 \in \Theta$, it holds that*

$$\|\lambda(\theta_1) - \lambda(\theta_2)\|_1 \le \frac{2\ell_{\psi,1}}{(1-\gamma)^2} \cdot \|\theta_1 - \theta_2\|.$$

(III) *The function $f(\lambda(\theta))$ is ℓ_F-smooth with respect to θ, where*

$$\ell_F = \frac{4\ell_{f,2} \cdot \ell_{\psi,1}^2}{(1-\gamma)^4} + \frac{8\ell_{\psi,1}^2 \cdot \ell_{f,1}}{(1-\gamma)^3} + \frac{2\ell_{f,1} \cdot \left(\ell_{\psi,2} + \ell_{\psi,1}^2\right)}{(1-\gamma)^2}.$$

E.2 Direct Parameterization

To give a clearer characterization, we further validate Assumption 4.2 for the direct parameterization. Recall that the discounted state-action visitation distribution for a given policy π is denoted as λ^π. We begin by showing that the one-to-one correspondence $\pi \to \lambda^\pi$ is Lipschitz continuous in Lemma E.2. Then, by leveraging the Lipschitz continuity, we show that $f(\lambda^\pi)$ is smooth w.r.t. π once $f(\lambda)$ is smooth w.r.t. λ in Proposition E.3. Again, we do not need to assume the concavity/convexity of $f(\cdot)$.

Lemma E.2 (Lipschitz continuity of λ^π w.r.t. π) *For every two policies π and π', it holds that*

$$\|\lambda^\pi - \lambda^{\pi'}\|_1 \le \frac{|\mathcal{A}|}{1-\gamma} \cdot \|\pi - \pi'\|_2.$$

Proof. Fix π' and define $h(\pi) = \|\lambda^\pi - \lambda^{\pi'}\|_1$. Then, we have

$$\nabla_\pi h(\pi) = \sum_{s,a} \mathrm{sign}(\lambda^\pi(s,a) - \lambda^{\pi'}(s,a)) \cdot \nabla_\pi \lambda^\pi(s,a),$$

where $\mathrm{sign}(x) = 1$ if $x \ge 0$, otherwise $\mathrm{sign}(x) = -1$. Let $\mathbf{1}_{s',a'} : \mathcal{S} \times \mathcal{A} \to \{0,1\}$ denote the indicator vector of the state-action pair (s',a') such that $\mathbf{1}_{s',a'}(s,a) = 1$ if and only if $(s,a) = (s',a')$. Then, we can view $\lambda^\pi(s,a)$ as a scaled value function with the reward function $\mathbf{1}_{s,a}$, i.e.,

$$\lambda^\pi(s,a) = \mathbf{1}_{s,a}^\top \lambda^\pi = (1-\gamma)V^\pi(\mathbf{1}_{s,a}). \tag{67}$$

We use $\mathbf{1}_{|\mathcal{S}||\mathcal{A}|}$ to denote the vector of ones with dimension $|\mathcal{S}||\mathcal{A}|$. Then, it holds that

$$\begin{aligned}
\|\nabla_\pi h(\pi)\|_2 &= \|\sum_{s,a} \mathrm{sign}(\lambda^\pi(s,a) - \lambda^{\pi'}(s,a)) \cdot \nabla_\pi \lambda^\pi(s,a)\|_2 \\
&\le \sum_{s,a} \|\nabla_\pi \lambda^\pi(s,a)\|_2 \\
&\le \sum_{s,a} \|\nabla_\pi \lambda^\pi(s,a)\|_1 \\
&\overset{(i)}{=} \sum_{s,a} \sum_{s',a'} d^\pi(s') \cdot Q^\pi(\mathbf{1}_{s,a}; s', a') \\
&= \sum_{s',a'} d^\pi(s') \cdot Q^\pi(\mathbf{1}_{|\mathcal{S}||\mathcal{A}|}; s', a') \\
&= \frac{|\mathcal{A}|}{1-\gamma},
\end{aligned}$$

where we use (67) and the policy gradient under direct parameterization (cf. Lemma G.2) in (i). The last line follows from the fact that $Q^\pi(\mathbf{1}_{|\mathcal{S}||\mathcal{A}|}; s', a') = 1/(1-\gamma)$ for all $(s', a') \in \mathcal{S} \times \mathcal{A}$ and $\sum_s d^\pi(s) = 1$. Therefore, we conclude that

$$
\begin{aligned}
\|\lambda^\pi - \lambda^{\pi'}\|_1 &= h(\pi) \\
&\leq h(\pi') + \max_{\pi_0}\{\|\nabla_\pi h(\pi_0)\|_2\} \cdot \|\pi - \pi'\|_2 \\
&\leq \|\lambda^{\pi'} - \lambda^{\pi'}\|_1 + \frac{|\mathcal{A}|}{1-\gamma}\|\pi - \pi'\|_2 \\
&= \frac{|\mathcal{A}|}{1-\gamma}\|\pi - \pi'\|_2,
\end{aligned}
$$

which completes the proof. □

Proposition E.3 (Smoothness of $f(\lambda^\pi)$ w.r.t. π) *Suppose that $f(\lambda)$ has a bounded and Lipschitz gradient in Λ, namely, there exist $\ell_{f,1}, \ell_{f,2} > 0$ such that*

$$
\|\nabla_\lambda f(\lambda)\|_\infty \leq \ell_{f,1}, \quad \|\nabla_\lambda f(\lambda) - \nabla_\lambda f(\lambda')\|_\infty \leq \ell_{f,2}\|\lambda - \lambda'\|_2, \quad \forall\ \lambda, \lambda' \in \Lambda.
$$

Then, $f(\lambda^\pi)$ is ℓ_F-smooth w.r.t. π, i.e.,

$$
\|\nabla_\pi f(\lambda^{\pi_1}) - \nabla_\pi f(\lambda^{\pi_2})\|_2 \leq \ell_F\|\pi_1 - \pi_2\|_2, \quad \forall\ \pi_1, \pi_2 \in \Pi,
$$

where

$$
\ell_F = \frac{4\ell_{f,1}\gamma|\mathcal{A}| + \ell_{f,2}|\mathcal{A}|^{3/2}}{(1-\gamma)^2}.
$$

Proof. By using the chain rule, we can write $\nabla_\pi f(\lambda^\pi) = (\nabla_\pi \lambda^\pi)^\top \nabla_\lambda f(\lambda^\pi)$. Thus,

$$
\begin{aligned}
&\|\nabla_\pi f(\lambda^{\pi_1}) - \nabla_\pi f(\lambda^{\pi_2})\|_2 \\
=\ &\|(\nabla_\pi \lambda^{\pi_1})^\top \nabla_\lambda f(\lambda^{\pi_1}) - (\nabla_\pi \lambda^{\pi_2})^\top \nabla_\lambda f(\lambda^{\pi_2})\|_2 \\
=\ &\|(\nabla_\pi \lambda^{\pi_1})^\top \nabla_\lambda f(\lambda^{\pi_1}) - (\nabla_\pi \lambda^{\pi_1})^\top \nabla_\lambda f(\lambda^{\pi_2}) \\
&\quad + (\nabla_\pi \lambda^{\pi_1})^\top \nabla_\lambda f(\lambda^{\pi_2}) - (\nabla_\pi \lambda^{\pi_2})^\top \nabla_\lambda f(\lambda^{\pi_2})\|_2 \\
=\ &\|(\nabla_\pi \lambda^{\pi_1})^\top \nabla_\lambda f(\lambda^{\pi_1}) - (\nabla_\pi \lambda^{\pi_1})^\top \nabla_\lambda f(\lambda^{\pi_2})\|_2 \\
&\quad + \|(\nabla_\pi \lambda^{\pi_1})^\top \nabla_\lambda f(\lambda^{\pi_2}) - (\nabla_\pi \lambda^{\pi_2})^\top \nabla_\lambda f(\lambda^{\pi_2})\|_2 \\
=\ &\underbrace{\|(\nabla_\pi \lambda^{\pi_1})^\top [\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})]\|_2}_{T_1} + \underbrace{\|[(\nabla_\pi \lambda^{\pi_1}) - (\nabla_\pi \lambda^{\pi_2})]^\top \nabla_\lambda f(\lambda^{\pi_2})\|_2}_{T_2}.
\end{aligned}
$$

To bound T_1, we notice that, by the definition $V^\pi(r) = r^\top \lambda^\pi/(1-\gamma)$,

$$
(\nabla_\pi \lambda^{\pi_1})^\top [\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})] = (1-\gamma)\nabla_\pi V^\pi([\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})])\Big|_{\pi=\pi_1} \qquad (68)
$$

Therefore, by the policy gradient under direct parameterization (cf. Lemma G.2), it holds that

$$
\begin{aligned}
T_1 &= \left\| (1-\gamma)\nabla_\pi V^\pi([\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})])\big|_{\pi=\pi_1} \right\|_2 \\
&= \left\{ \sum_{s,a} [d^{\pi_1}(s) \cdot Q^{\pi_1}([\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})]; s, a)]^2 \right\}^{\frac{1}{2}} \\
&= \left\{ \sum_{s} (d^{\pi_1}(s))^2 \cdot \sum_{a} [Q^{\pi_1}([\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})]; s, a)]^2 \right\}^{\frac{1}{2}} \\
&= \left\{ \sum_{s} (d^{\pi_1}(s))^2 \right\}^{\frac{1}{2}} \cdot \max_s \left\{ \sum_{a} [Q^{\pi_1}([\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})]; s, a)]^2 \right\}^{\frac{1}{2}} \\
&\overset{(i)}{\leq} \|d^{\pi_1}\|_1 \cdot \sqrt{|\mathcal{A}|} \cdot \max_{s,a} |Q^{\pi_1}([\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})]; s, a)| \\
&\overset{(ii)}{\leq} \sqrt{|\mathcal{A}|} \cdot \frac{\|\nabla_\lambda f(\lambda^{\pi_1}) - \nabla_\lambda f(\lambda^{\pi_2})\|_\infty}{1-\gamma} \\
&\overset{(iii)}{\leq} \frac{\ell_{f,2}\sqrt{|\mathcal{A}|}}{1-\gamma} \|\lambda^{\pi_1} - \lambda^{\pi_2}\|_2 \\
&\leq \frac{\ell_{f,2}\sqrt{|\mathcal{A}|}}{1-\gamma} \|\lambda^{\pi_1} - \lambda^{\pi_2}\|_1 \\
&\leq \frac{\ell_{f,2}|\mathcal{A}|^{3/2}}{(1-\gamma)^2} \|\pi_1 - \pi_2\|_2,
\end{aligned} \tag{69}
$$

where (i) uses the inequality $\|\cdot\|_2 \leq \|\cdot\|_1$ and (ii) is due to $\|d^\pi\|_1 = 1$ and $|Q^\pi(r; s, a)| \leq \|r\|_\infty/(1-\gamma)$. (iii) results from the smoothness assumption of $f(\lambda)$. The last step follows from the Lipschitz continuity of λ^π w.r.t. π (cf. Lemma E.2).

Now, to bound T_2, we use the relation as described in (68) again, namely

$$
\begin{aligned}
&(\nabla_\pi \lambda^{\pi_1} - \nabla_\pi \lambda^{\pi_2})^\top [\nabla_\lambda f(\lambda^{\pi_2})] \\
=&(1-\gamma) \left[\nabla_\pi V^\pi(\nabla_\lambda f(\lambda^{\pi_2}))\big|_{\pi=\pi_1} - \nabla_\pi V^\pi(\nabla_\lambda f(\lambda^{\pi_2}))\big|_{\pi=\pi_2} \right]
\end{aligned}
$$

Thus, by Lemma G.3, it holds that

$$
\begin{aligned}
T_2 &= \| [(\nabla_\pi \lambda^{\pi_1}) - (\nabla_\pi \lambda^{\pi_2})]^\top \nabla_\lambda f(\lambda^{\pi_2}) \|_2 \\
&= (1-\gamma) \left\| \nabla_\pi V^\pi(\nabla_\lambda f(\lambda^{\pi_2}))\big|_{\pi=\pi_1} - \nabla_\pi V^\pi(\nabla_\lambda f(\lambda^{\pi_2}))\big|_{\pi=\pi_2} \right\|_2 \\
&\leq (1-\gamma) \cdot \frac{4\gamma|\mathcal{A}|}{(1-\gamma)^3} \cdot \|\nabla_\lambda f(\lambda^{\pi_2})\|_\infty \cdot \|\pi_1 - \pi_2\|_2 \\
&\leq \frac{4\ell_{f,1}\gamma|\mathcal{A}|}{(1-\gamma)^2} \|\pi_1 - \pi_2\|_2,
\end{aligned} \tag{70}
$$

where we use the assumption $\|\nabla_\lambda f(\lambda)\|_\infty \leq \ell_{f,1}$ in the last inequality. The proof is completed by combining (69) and (70). □

Appendix F Further Discussions About Direct Parameterization

In this section, as we focus on the direct parameterization, we adopt the notation $F(\pi) = f(\lambda^\pi)$, $G(\pi) = g(\lambda^\pi)$, and $\mathcal{L}(\pi, \mu) = L(\lambda^\pi, \mu) = f(\lambda^\pi) - \mu g(\lambda^\pi)$. We denote the optimal policy by $\pi^\star$. The algorithm (11) then becomes

$$
\pi^{t+1} = \mathcal{P}_\Pi\left(\pi^t + \eta_1 \nabla_\pi \mathcal{L}(\pi^t, \mu^t)\right), \quad \mu^{t+1} = \mathcal{P}_U\left(\mu^t - \eta_2 \nabla_\mu \mathcal{L}(\pi^t, \mu^t)\right), \quad \text{for } t = 0, 1, 2, \ldots \tag{71}
$$

As a special case of (8), the direct parameterization satisfies a stronger version of Assumption 4.1. Since there is a bijection between the policy π and the state-action visitation distribution λ^π, the inverse map $\lambda^{-1}(\cdot)$ is well-defined globally on Λ. Furthermore, when the state visitation distribution

d^π is universally bounded away from 0, the inverse $\lambda^{-1}(\cdot)$ is Lipschitz continuous (see Lemma D.1 in Appendix D). We still assume that $F(\pi)$ is ℓ_F-smooth and $G(\pi)$ is ℓ_G-smooth as in Assumption 4.2. It is shown in Proposition E.3 that this assumption is satisfied when $f(\cdot)$ and $g(\cdot)$ are smooth with respect to λ.

Below, in Lemma F.1, we show that the update (71) also enjoys the variational gradient dominance property for standard MDPs (see, e.g., [34, Lemma 4.1]), i.e.,

$$\mathcal{L}(\pi^\star,\mu^t) - \mathcal{L}(\pi^{t+1},\mu^t) \le 2\sqrt{2|S|}\cdot\left\|\frac{d^{\pi^\star}}{d^{\pi^t}}\right\|_\infty \cdot \ell_L \left\|\pi^t - \pi^{t+1}\right\|_2, \quad \text{for } t = 0,1,2,\ldots. \tag{72}$$

Since the projected gradient update (71) has the following descent property [39, Theorem 1]:

$$\mathcal{L}(\pi^{t+1},\mu^t) - \mathcal{L}(\pi^t,\mu^t) \ge \frac{\ell_L}{2}\|\pi^t - \pi^{t+1}\|_2^2,$$

together with (72), we have that

$$\left[\max\{0,\mathcal{L}(\pi^\star,\mu^t) - \mathcal{L}(\pi^{t+1},\mu^t)\}\right]^2 \le 16|\mathcal{S}|\cdot\left\|\frac{d^{\pi^\star}}{d^{\pi^t}}\right\|_\infty^2 \cdot \ell_L\left[\mathcal{L}(\pi^{t+1},\mu^t) - \mathcal{L}(\pi^t,\mu^t)\right]. \tag{73}$$

This is the counterpart to (34). Following this line of argument, we derive a similar bound for the average performance in terms of the Lagrangian:

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\pi^\star,\mu^t) - \mathcal{L}(\pi^t,\mu^t)\right]\\ &\le \frac{2\sqrt{2|S|}\cdot\|d^{\pi^\star}\|_\infty}{d_0}\cdot\sqrt{2\ell_L\left(\frac{2M_L}{T}+\eta_2 M_G^2\right)} + \left(\frac{2M_L}{T}+\eta_2 M_G^2\right).\end{aligned} \tag{74}$$

By taking $\eta^2 = T^{-2/3}$ in (74), this yields a similar result as Proposition 4.4. We summarize this result in Proposition F.2 below.

Lemma F.1 (Variational gradient dominance) *Let Assumption 4.2 hold and choose $\eta_1 = 1/\ell_L$. The sequence $\left\{(\pi^t,\mu^t)\right\}_t$ generated by algorithm (11) satisfies*

$$\mathcal{L}(\pi^\star,\mu^t) - \mathcal{L}(\pi^{t+1},\mu^t) \le 2\sqrt{2|S|}\cdot\left\|\frac{d^{\pi^\star}}{d^{\pi^t}}\right\|_\infty \cdot \ell_L \left\|\pi^t - \pi^{t+1}\right\|_2, \quad \textit{for } t = 0,1,2,\ldots.$$

Proof. By the concavity of $L(\cdot,\mu^t)$, when $t \ge 0$, it yields that

$$\begin{aligned}\mathcal{L}(\pi^\star,\mu^t) - \mathcal{L}(\pi^t,\mu^t) &= L(\lambda^{\pi^\star},\mu^t) - L(\lambda^{\pi^t},\mu^t)\\ &\le (\lambda^{\pi^\star} - \lambda^{\pi^t})^\top \nabla_\lambda L(\lambda^{\pi^t},\mu^t)\\ &= (1-\gamma)\left(V^{\pi^\star}(\nabla_\lambda L(\lambda^{\pi^t},\mu^t)) - V^{\pi^t}(\nabla_\lambda L(\lambda^{\pi^t},\mu^t))\right)\\ &\overset{(i)}{=} \sum_{s,a} d^{\pi^\star}(s)\cdot\pi^\star(a|s)\cdot A^{\pi^t}\left(\nabla_\lambda L(\lambda^{\pi^t},\mu^t);s,a\right)\\ &\le \sum_s d^{\pi^\star}(s)\cdot\max_a\left\{A^{\pi^t}\left(\nabla_\lambda L(\lambda^{\pi^t},\mu^t);s,a\right)\right\}\\ &= \sum_s \frac{d^{\pi^\star}(s)}{d^{\pi^t}(s)}\cdot d^{\pi^t}(s)\cdot\max_a\left\{A^{\pi^t}\left(\nabla_\lambda L(\lambda^{\pi^t},\mu^t);s,a\right)\right\}\\ &\overset{(ii)}{\le} \left\|\frac{d^{\pi^\star}}{d^{\pi^t}}\right\|_\infty\cdot\sum_s d^{\pi^t}(s)\cdot\max_a\left\{A^{\pi^t}\left(\nabla_\lambda L(\lambda^{\pi^t},\mu^t);s,a\right)\right\},\end{aligned} \tag{75}$$

where we use the performance difference lemma (cf. Lemma G.4) in (i). Recall that $A^\pi(r;s,a)$ denotes the advantage function with reward $r(\cdot,\cdot)$ under policy π (cf. (87)). The inequality (ii) holds since $\max_a\left\{A^{\pi^t}\left(\nabla_\lambda L(\lambda^{\pi^t},\mu^t);s,a\right)\right\} \ge 0$. The summation term in the last line can be analyzed as

$$
\begin{aligned}
&\sum_s d^{\pi^t}(s) \cdot \max_a \left\{ A^{\pi^t} \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t); s, a \right) \right\} \\
&\overset{(i)}{=} \max_{\pi' \in \Pi} \left\{ \sum_{s,a} d^{\pi^t}(s) \cdot \pi'(a|s) \cdot A^{\pi^t} \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t); s, a \right) \right\} \\
&\overset{(ii)}{=} \max_{\pi' \in \Pi} \left\{ \sum_{s,a} d^{\pi^t}(s) \cdot [\pi'(a|s) - \pi^t(a|s)] \cdot A^{\pi^t} \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t); s, a \right) \right\} \\
&\overset{(iii)}{=} \max_{\pi' \in \Pi} \left\{ \sum_{s,a} d^{\pi^t}(s) \cdot [\pi'(a|s) - \pi^t(a|s)] \cdot Q^{\pi^t} \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t); s, a \right) \right\} \\
&\overset{(iv)}{=} (1-\gamma) \max_{\pi' \in \Pi} \left\{ (\pi' - \pi^t)^\top \nabla_\pi \left[V^\pi \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t) \right) \right] \Big|_{\pi = \pi^t} \right\},
\end{aligned} \tag{76}
$$

where (i) holds as the maximum policy is attained at the action that maximizes $A^{\pi^t} \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t); s, \cdot \right)$. Step (ii) follows since $\sum_a \pi^t(a|s) \cdot A^{\pi^t} \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t); s, a \right) = 0$. Then $\sum_a [\pi'(a|s) - \pi^t(a|s)] \cdot V^{\pi^t} = 0$ makes (iii) holds. The last step (iv) uses the policy gradient under direct parameterization (cf. Lemma G.2), which can be further written as

$$
\begin{aligned}
\nabla_\pi \left[V^\pi \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t) \right) \right] \Big|_{\pi = \pi^t} &\overset{(i)}{=} \frac{1}{1-\gamma} \nabla_\pi \left[\left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t) \right)^\top \lambda^\pi \right] \Big|_{\pi = \pi^t} \\
&= \frac{1}{1-\gamma} \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t) \right)^\top \nabla_\pi \lambda^{\pi^t} \\
&\overset{(ii)}{=} \frac{1}{1-\gamma} \nabla_\pi L(\lambda^{\pi^t}, \mu^t) \\
&= \frac{1}{1-\gamma} \nabla_\pi \mathcal{L}(\pi^t, \mu^t),
\end{aligned} \tag{77}
$$

where (i) follows from the definition of $V^\pi \left(\nabla_\lambda L(\lambda^{\pi^t}, \mu^t) \right)$, and step (ii) is obtained by the chain rule. Thus, it follows from (75)-(77) that

$$
\mathcal{L}(\pi^\star, \mu^t) - \mathcal{L}(\pi^t, \mu^t) \le \left\| \frac{d^{\pi^\star}}{d^{\pi^t}} \right\|_\infty \cdot \max_{\pi' \in \Pi} \left\{ \left(\pi' - \pi^t \right)^\top \nabla_\pi \mathcal{L}(\pi^t, \mu^t) \right\}, \quad \text{for } t = 0, 1, 2, \ldots. \tag{78}
$$

Let $\left\{ \left(\pi^t, \mu^t \right) \right\}_t$ be the sequence generated by the algorithm (71) with the primal step-size $\eta_1 = 1/\ell_L$. Following [39, Theorem 1], the update (71) satisfies that

$$
\left(\pi' - \pi^{t+1} \right)^\top \nabla_\pi \mathcal{L}(\pi^{t+1}, \mu^t) \le 2\ell_L \|\pi^t - \pi^{t+1}\|_2 \cdot \|\pi' - \pi^{t+1}\|_2, \quad \text{for } t = 0, 1, 2, \ldots. \tag{79}
$$

Maximizing both sides of (79) in terms of π' yields that

$$
\begin{aligned}
\max_{\pi' \in \Pi} \left\{ \left(\pi' - \pi^{t+1} \right)^\top \nabla_\pi \mathcal{L}(\pi^{t+1}, \mu^t) \right\} &\le \max_{\pi' \in \Pi} \left\{ \|\pi' - \pi^{t+1}\|_2 \right\} \cdot 2\ell_L \|\pi^t - \pi^{t+1}\|_2 \\
&\le \sqrt{2|S|} \cdot 2\ell_L \|\pi^t - \pi^{t+1}\|_2,
\end{aligned} \tag{80}
$$

where we use $\max_{\pi_1, \pi_2 \in \Pi} \{\|\pi_1 - \pi_2\|_2\} \le \sqrt{2|S|}$ in the last inequality. Combining (78) and (80) leads to the desired result. $\square$

Proposition F.2 *Let Assumption 4.2 hold and suppose that $d_0 := \min_{s \in \mathcal{S}, \pi \in \Pi} d^\pi(s) > 0$. Then, for every $T > 0$, the iterates $\left\{ \left(\pi^t, \mu^t \right) \right\}_{t=0}^{T-1}$ produced by algorithm (71) with $\eta_1 = 1/\ell_L$ satisfy*

$$
\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \left[\mathcal{L}(\pi^\star, \mu^t) - \mathcal{L}(\pi^t, \mu^t) \right] \\
&\le \frac{2\sqrt{2|S|} \cdot \|d^{\pi^\star}\|_\infty}{d_0} \cdot \sqrt{2\ell_L \left(\frac{2M_L}{T} + \eta_2 M_G^2 \right)} + \left(\frac{2M_L}{T} + \eta_2 M_G^2 \right).
\end{aligned}
$$

Proof. By applying the descent property [39, Theorem 1] of the projected gradient algorithm to the primal update (71), it holds that

$$\mathcal{L}(\pi^{t+1}, \mu^t) - \mathcal{L}(\pi^t, \mu^t) \geq \frac{\ell_L}{2} \|\pi^t - \pi^{t+1}\|_2^2, \tag{81}$$

By Lemma F.1, we have that

$$\begin{aligned}\mathcal{L}(\pi^\star, \mu^t) - \mathcal{L}(\pi^{t+1}, \mu^t) &\leq 2\sqrt{2|S|} \cdot \left\| \frac{d^{\pi^\star}}{d^{\pi^t}} \right\|_\infty \cdot \ell_L \left\| \pi^t - \pi^{t+1} \right\|_2 \\ &\leq \frac{2\sqrt{2|S|} \cdot \|d^{\pi^\star}\|_\infty \cdot \ell_L}{d_0} \left\| \pi^t - \pi^{t+1} \right\|_2,\end{aligned}$$

which implies that

$$\|\pi^t - \pi^{t+1}\|_2^2 \geq \left(\frac{d_0}{2\sqrt{2|S|} \cdot \|d^{\pi^\star}\|_\infty \cdot \ell_L} \right)^2 \cdot \left[\max\{0, \mathcal{L}(\pi^\star, \mu^t) - \mathcal{L}(\pi^{t+1}, \mu^t)\} \right]^2. \tag{82}$$

Combining inequalities (81) and (82) yields that

$$\mathcal{L}(\pi^{t+1}, \mu^t) - \mathcal{L}(\pi^t, \mu^t) \geq \frac{1}{2C_1^2 \ell_L} \cdot \left[\max\{0, \mathcal{L}(\pi^\star, \mu^t) - \mathcal{L}(\pi^{t+1}, \mu^t)\} \right]^2,$$

where we denote $C_1 := 2\sqrt{2|S|} \cdot \|d^{\pi^\star}\|_\infty / d_0$. Summing over $t = 0, \dots, T-1$, we obtain that

$$\begin{aligned}\sum_{t=0}^{T-1} \left[\mathcal{L}(\pi^{t+1}, \mu^t) - \mathcal{L}(\pi^t, \mu^t) \right] &\geq \frac{1}{2C_1^2 \ell_L} \cdot \sum_{t=0}^{T-1} \left[\max\{0, \mathcal{L}(\pi^\star, \mu^t) - \mathcal{L}(\pi^{t+1}, \mu^t)\} \right]^2 \\ &\geq \frac{1}{2C_1^2 \ell_L} \cdot \frac{1}{T} \left[\sum_{t=0}^{T-1} \max\{0, \mathcal{L}(\pi^\star, \mu^t) - \mathcal{L}(\pi^{t+1}, \mu^t)\} \right]^2,\end{aligned} \tag{83}$$

where the last line in (83) results from the Cauchy-Schwarz inequality.

We then provide an upper bound on the left-hand side of (83). By the definition of Lagrangian $\mathcal{L}(\pi, \mu)$,

$$\begin{aligned}&\sum_{t=0}^{T-1} \left[\mathcal{L}(\pi^{t+1}, \mu^t) - \mathcal{L}(\pi^t, \mu^t) \right] \\ &= \sum_{t=0}^{T-1} \left[F(\pi^{t+1}) - \mu^t G(\pi^{t+1}) - F(\pi^t) + \mu^t G(\pi^t) \right] \\ &\overset{(i)}{=} F(\pi^T) - F(\pi^0) + \mu^0 G(\pi^0) - \mu^{T-1} G(\pi^T) + \sum_{t=1}^{T-1} (\mu^t - \mu^{t-1}) G(\pi^t) \\ &\leq 2M_F + 2C_0 M_G + \sum_{t=1}^{T-1} (\mu^t - \mu^{t-1}) G(\pi^t) \\ &= 2M_L + \sum_{t=1}^{T-1} (\mu^t - \mu^{t-1}) G(\pi^t),\end{aligned} \tag{84}$$

where we take telescoping sums and change the index of summation in (i). The summation term in the last line of (84) has the order $\mathcal{O}(T)$, in particular

$$\begin{aligned}\sum_{t=1}^{T-1} (\mu^t - \mu^{t-1}) G(\pi^t) &\leq \sum_{t=1}^{T-1} |\mu^t - \mu^{t-1}| \cdot |G(\pi^t)| \\ &\leq M_G \sum_{t=1}^{T-1} \left| \mathcal{P}_U \left(\mu^{t-1} - \eta_2 \nabla_\mu \mathcal{L}(\pi^{t-1}, \mu^{t-1}) \right) - \mu^{t-1} \right| \\ &\leq M_G \sum_{t=1}^{T-1} \left| \mu^{t-1} - \eta_2 \nabla_\mu \mathcal{L}(\pi^{t-1}, \mu^{t-1}) - \mu^{t-1} \right| \\ &= \eta_2 M_G \sum_{t=1}^{T-1} |G(\pi^{t-1})| \\ &\leq \eta_2 M_G^2 T.\end{aligned}$$

Thus, we obtain an upper bound such that

$$\sum_{t=0}^{T-1}\left[\mathcal{L}(\pi^{t+1},\mu^t)-\mathcal{L}(\pi^t,\mu^t)\right]\leq 2M_L+\eta_2 M_G^2 T, \tag{85}$$

which further implies, by (83), that

$$\frac{1}{2C_1^2\ell_L}\cdot\frac{1}{T}\left[\sum_{t=0}^{T-1}\max\{0,\mathcal{L}(\pi^\star,\mu^t)-\mathcal{L}(\pi^{t+1},\mu^t)\}\right]^2\leq 2M_L+\eta_2 M_G^2 T. \tag{86}$$

Therefore, we can bound the average performance as

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\pi^\star,\mu^t)-\mathcal{L}(\pi^t,\mu^t)\right]\\
&=\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\pi^\star,\mu^t)-\mathcal{L}(\pi^{t+1},\mu^t)\right]+\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\pi^{t+1},\mu^t)-\mathcal{L}(\pi^t,\mu^t)\right]\\
&\leq\frac{1}{T}\sum_{t=0}^{T-1}\max\{0,\mathcal{L}(\pi^\star,\mu^t)-\mathcal{L}(\pi^{t+1},\mu^t)\}+\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\pi^{t+1},\mu^t)-\mathcal{L}(\pi^t,\mu^t)\right]\\
&\overset{(i)}{\leq} C_1\sqrt{2\ell_L\left(\frac{2M_L}{T}+\eta_2M_G^2\right)}+\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\pi^{t+1},\mu^t)-\mathcal{L}(\pi^t,\mu^t)\right]\\
&\overset{(ii)}{\leq} C_1\sqrt{2\ell_L\left(\frac{2M_L}{T}+\eta_2M_G^2\right)}+\left(\frac{2M_L}{T}+\eta_2M_G^2\right)\\
&=\frac{2\sqrt{2|S|}\cdot\|d^{\pi^\star}\|_\infty}{d_0}\cdot\sqrt{2\ell_L\left(\frac{2M_L}{T}+\eta_2M_G^2\right)}+\left(\frac{2M_L}{T}+\eta_2M_G^2\right),
\end{aligned}$$

where we use (86) in (i) and (85) in (ii). This completes the proof. □

Appendix G Auxiliary Lemmas

In this section, we present a few auxiliary lemmas that we needed for the proofs of main results in this paper. These lemmas are standard results on Markov decision processes. We refer the reader to Section 2 for necessary definitions and [34] for the proofs of these results.

Lemma G.1 (Policy gradient under general parameterization) *Let $V^{\pi_\theta}(r)$ be the value function under policy π_θ with an arbitrary reward function $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$. The gradient of $V^{\pi_\theta}(r)$ with respect to θ is given by*

$$\nabla_\theta V^{\pi_\theta}(r)=\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta\log\pi_\theta(a|s)\cdot Q^{\pi_\theta}(r;s,a)\right].$$

Lemma G.2 (Policy gradient under direct parameterization) *Let $V^\pi(r)$ be the value function under policy π with an arbitrary reward function $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$. The gradient of $V^\pi(r)$ with respect to π is given by*

$$\frac{\partial V^\pi(r)}{\partial\pi(a|s)}=\frac{1}{1-\gamma}d^\pi(s)\cdot Q^\pi(r;s,a),\ \forall\ (s,a)\in\mathcal{S}\times\mathcal{A}.$$

Lemma G.3 (Smoothness of $V^\pi(r)$ w.r.t. π) *Let $V^\pi(r)$ be the value function under policy π with an arbitrary reward function $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$. For every two policies π and π', it holds that*

$$\left\|\nabla_\pi V^\pi(r)-\nabla_\pi V^{\pi'}(r)\right\|_2\leq\frac{4\gamma|\mathcal{A}|}{(1-\gamma)^3}\cdot\|r\|_\infty\cdot\|\pi-\pi'\|_2.$$

Lemma G.4 (Performance difference) *Let $V^{\pi}(r)$ be the value function under policy π with an arbitrary reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. For every two policies π and π', it holds that*

$$\begin{aligned} V^{\pi'}(r) - V^{\pi}(r) &= \frac{1}{1-\gamma} \left\langle r, \lambda^{\pi'} - \lambda^{\pi} \right\rangle \\ &= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \left(\pi'(a|s) - \pi(a|s)\right) \cdot Q^{\pi'}(r; s, a) \\ &= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d^{\pi'}(s) \sum_{a \in \mathcal{A}} \pi'(a|s) \cdot A^{\pi}(r; s, a) \end{aligned}$$

where $A^{\pi}(r; s, a)$ denotes the advantage function with reward $r(\cdot, \cdot)$ under policy π, defined as

$$A^{\pi}(r; s, a) := Q^{\pi}(r; s, a) - \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, a_t \sim \pi(\cdot|s_t), s_0 = s \right], \ \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{87}$$