A Dual Approach to Constrained Markov Decision Processes with Entropy Regularization

Donghao Ying
University of California, Berkeley
donghaoy@berkely.edu

Yuhao Ding
University of California, Berkeley
yuhao_ding@berkely.edu

Javad Lavaei
University of California, Berkeley
lavaei@berkeley.edu

Abstract

We study entropy-regularized constrained Markov decision processes (CMDPs) under the soft-max parameterization, in which an agent aims to maximize the entropy-regularized value function while satisfying constraints on the expected total utility. By leveraging the entropy regularization, our theoretical analysis shows that its Lagrangian dual function is smooth and the Lagrangian duality gap can be decomposed into the primal optimality gap and the constraint violation. Furthermore, we propose an accelerated dual-descent method for entropy-regularized CMDPs. We prove that our method achieves the global convergence rate $\widetilde{\mathcal{O}}(1/T)$ for both the optimality gap and the constraint violation for entropy-regularized CMDPs. A discussion about a linear convergence rate for CMDPs with a single constraint is also provided.

1 INTRODUCTION

In many sequential decision-making problems for safety-critical systems, e.g. autonomous driving (Fisac et al., 2018) and cyber-physical systems (Zhang et al., 2019), the optimality of an objective function by itself is not sufficient and a variety of constraints must be satisfied. This has naturally led to a generalization of the model of Markov Decision Processes (MDPs) to Constrained MDPs (CMDPs) (Altman, 1999), in which an agent aims to maximize the value function while satisfying given constraints on the expected total utility.

Direct policy search methods, including the policy gradient and the natural policy gradient (NPG) methods, have had substantial empirical successes in solving CMDPs (Achiam et al., 2017; Chow et al., 2017; Bhatnagar and Lakshmanan, 2012; Borkar, 2005; Uchibe and Doya, 2007; Achiam et al., 2017). Recently, a major progress in understanding the theoretical non-asymptotic global convergence behavior of policy-based methods for CMDPs has also been achieved (Ding et al., 2020, 2021; Xu et al., 2021; Efroni et al., 2020; Chen et al., 2021).

For policy-based methods, entropy regularization is a popular technique for encouraging the exploration of an unknown environment and preventing a premature convergence (Williams and Peng, 1991; Mnih et al., 2016; Haarnoja et al., 2018; Zang et al., 2020). From a theoretical optimization perspective, it is shown in Mei et al. (2020) and Cen et al. (2021) that the entropy regularization can make the policy optimization landscape benign and achieve faster convergence rates even in the exact value evaluation setting. Nevertheless, most existing theoretical guarantees for the entropy-regularized policy optimization are restricted to unconstrained MDPs. The scope of the power of entropy regularization for CMDPs remains unknown even for the tabular setting with the exact value evaluation.

Inspired by the recent theoretical advances towards understanding entropy-regularized policy gradient methods (Mei et al., 2020; Cen et al., 2021) together with the global convergence of Lagrangian-based methods for CMDPs (Ding et al., 2020, 2021; Paternain et al., 2019; Xu et al., 2021), we investigate the optimization properties induced by the entropy regularization for CMDPs under the soft-max policy parameterization. We focus on the study of tabular CMDPs with the exact gradient evaluation. This is the setting commonly investigated in the literature since its understanding assists in demystifying the effectiveness of entropy-regularization in CMDPs with more complex settings.

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1.1 Contributions

This work is the first one that certifies the effectiveness of entropy regularization in CMDPs from an optimization perspective. We summarize our contributions below:

- We first show that although the underlying problem is nonconcave, the Lagrangian dual function of CMDPs with the entropy regularization is smooth under the Slater condition and the exploratory initial distribution assumption. Under the same conditions, an $\mathcal{O}(\varepsilon)$ error bound for the dual optimality gap leads to an $\mathcal{O}\left(\sqrt{\varepsilon}\right)$ error bound for the primal optimality gap and the constraint violation.
- To leverage the smoothness of the Lagrangian dual function, we propose a new accelerated dual-descent method for entropy-regularized CMDPs, which updates the dual variable via projected accelerated gradient descent and uses the natural policy gradient method in the inner loop.
- We prove that the proposed method achieves a global convergence with the rate $\widetilde{\mathcal{O}}(1/T)$ for both the optimality gap and the constraint violation for entropy-regularized CMDPs.
- In the special case where CMDPs only have a single constraint, we show that a bisection-based dual approach can achieve a linear convergence rate.

1.2 Related Work

CMDPs Our work is related to policy-based CMDP algorithms (Altman, 1999; Borkar, 2005; Bhatnagar and Lakshmanan, 2012; Chow et al., 2017; Ding et al., 2020, 2021; Xu et al., 2021; Chen et al., 2021; Efroni et al., 2020). The papers Ding et al. (2020) and Xu et al. (2021) are closely related to our work. In Ding et al. (2020), the authors propose a natural policy gradient primal-dual method for CMDPs and prove that it achieves global convergence with the rate $\mathcal{O}(1/\sqrt{T})$ for both the optimality gap and the constraint violation under the soft-max policy parameterization. The work Xu et al. (2021) achieves a similar global convergence rate as Ding et al. (2020) using a primal-based approach. However, the entropy regularization, which is an effective technique for unconstrained MDPs, is not used in these algorithms.

Entropy-regularized RL Maximum entropy reinforcement learning optimizes policies to jointly maximize the expected return and the expected entropy of the policy. This framework has been used in many contexts. It has been shown that the maximum entropy formulation provides a substantial improvement in exploration and robustness (Ziebart, 2010). It is robust in the face of model and estimation errors (Haarnoja et al., 2017) in both on-policy and off-policy settings (Haarnoja et al., 2018). More recently, the theoretical results in Mei et al. (2020) and Cen et al. (2021) have shown that the entropy regularization can help policy-based methods improve the convergence rate and the sample complexity compared with standard MDPs without the entropy regularization. However, despite the tremendous successes of the entropy regularization in unconstrained MDPs, the impact of the entropy regularization for CMDPs remains unknown.

1.3 Notations

Let $\Delta(\mathcal{S})$ denote the probability simplex over the set $\mathcal{S}$, and let $|\mathcal{S}|$ denote its cardinality. For a set $T \subset \mathbb{R}^p$, let $\mathrm{cl}(T)$ denote the closure of T. 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 number $a \in \mathbb{R}$, let $\mathrm{sign}(a)$ denote the sign of a, i.e. $\mathrm{sign}(a) = +1$ if $a \geq 0$ and $\mathrm{sign}(a) = -1$ if $a < 0$. Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For a vector x, we use x^T to denote the transpose of x, and use x_i or $(x)_i$ to denote the i-th entry of x. When applying a scalar function to x, e.g. $\log x$, the operation is understood as entry-wise. For vectors x and y, we use $x \geq y$ to denote an entry-wise inequality. 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 matrix A, we use A_{ij} to denote its (i,j)-th entry, and let $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$. Let I_n denote the $n \times n$ identity matrix. For square matrices A and B, we use $A \succeq B$ to denote that $A - B$ is positive semi-definite. For a function $f(x)$, let $\nabla_x f(x)$ (resp. $\nabla_{xx}^2 f(x)$) denote its gradient (resp. Hessian) with respect to x, and we may omit x in the subscript when it is clear from the context. Let $\arg\min f(x)$ (resp. $\arg\max f(x)$) denote any arbitrary global minimum (resp. global maximum) of $f(x)$. We use boldface symbols for constraint-related vectors, e.g. $\boldsymbol{\lambda}$.

2 PROBLEM FORMULATION

Markov Decision Processes An infinite-horizon Markov Decision Process $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with a finite state-action space is specified by: a finite state space $\mathcal{S}$; a finite action space $\mathcal{A}$; a transition dynamics $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $P(s'|s,a)$ is the probability of transition from state s to state s' when action a is taken; a reward function $r : \mathcal{S} \times \mathcal{A} \to [0,1]$, where $r(s,a)$ is the instantaneous reward when taking action a in state s; a discount factor $\gamma \in [0,1)$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$

represents that the decision rule the agent uses, i.e. the agent takes action a with probability $\pi(a|s)$ in state s. We can also interpret a policy π as a vector in $\Delta(\mathcal{A})^{|\mathcal{S}|} \subset \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Given a policy π, the value function $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined to characterize the discounted sum of the rewards earned under π, i.e.

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s\right], \ \forall s \in \mathcal{S} \tag{1}$$

where the expectation is taken over all possible trajectories, in which $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. When the initial state is sampled from some distribution ρ, we slightly abuse the notation and define the value function as

$$V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right] \tag{2}$$

One classical property of the value function is that it is sufficiently smooth with respect to the policy if we view $V^\pi(\rho)$ as a function from the policy space $\Delta(\mathcal{A})^{|\mathcal{S}|}$ to $\mathbb{R}$. Especially, $V^\pi(\rho)$ has the following Lipschitz property.

Lemma 2.1 *For arbitrary policies π_1 and π_2, it holds*

$$|V^{\pi_1}(\rho) - V^{\pi_2}(\rho)| \le \ell_c \left\|\pi_1 - \pi_2\right\|_2, \tag{3}$$

where $\ell_c = \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}$.

Lemma 2.1 follows from the bounded gradient of $V^\pi(\rho)$ with respect to π. We refer the reader to the supplement in Appendix A for more details.

The action-value function (or Q-function) $Q^\pi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ under policy π is defined as

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a\right] \tag{4}$$

which can be interpreted as the expected total reward with an initial state $s_0 = s$ and an initial action $a_0 = a$. Since $r(s,a) \in [0,1]$ by assumption, we have that both $Q^\pi(s,a)$ and $V^\pi(\rho)$ are bounded between $[0, 1/(1-\gamma)]$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and initial distribution ρ.

For theoretical analysis, it is useful to define the so-called discounted state visitation distribution $d^\pi_{s_0}$ of a policy π:

$$d^\pi_{s_0}(s) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi, s_0), \ \forall s \in \mathcal{S} \tag{5}$$

and we write $d^\pi_\rho(s) := \mathbb{E}_{s_0\sim\rho}\left[d^\pi_{s_0}(s)\right]$ as the visitation distribution when the initial state follows ρ.

Soft-max Parameterization Parameterization is commonly deployed to model unknown policies to help with the optimization process. One natural choice is the soft-max parameterization:

$$\pi_\theta(a|s) := \frac{\exp(\theta(s,a))}{\sum_{a'\in\mathcal{A}} \exp\left(\theta(s,a')\right)}, \ \forall (s,a) \in \mathcal{S}\times\mathcal{A} \tag{6}$$

where $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is an unconstrained vector. We denote the class of all soft-max parameterized policies by Π. This policy class is complete in the sense that its closure $\mathrm{cl}(\Pi)$ contains all stationary policies. In what follows, we will discard the subscript θ and just write $\pi \in \Pi$, whenever it is clear from the context.

Entropy Regularization To encourage exploration and accelerate convergence to the optimal policy, entropy regularization is widely used in solving MDPs. In the regularized setting, the agent seeks to optimize the entropy-regularized value function

$$V^\pi_\tau(\rho) := V^\pi(\rho) + \tau \cdot \mathcal{H}(\rho, \pi), \tag{7}$$

where $\tau \ge 0$ specifies the weight of regularization and $\mathcal{H}(\rho,\pi)$ is the discounted entropy defined by

$$\mathcal{H}(\rho,\pi) := \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log\pi(a_t|s_t) \,\middle|\, \pi, s_0 \sim \rho\right]. \tag{8}$$

We can also define the Q-function under regularization, which is referred to as the soft Q-function

$$Q^\pi_\tau(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V^\pi_\tau(s')\right]. \tag{9}$$

Constrained MDP With Entropy Regularization In a Constrained Markov Decision Process $\mathrm{CMDP}(\mathcal{S},\mathcal{A},P,r,\mathbf{g},\mathbf{b},\gamma)$, besides the reward function r, we have a utility function $\mathbf{g} = (g_1,\dots,g_n) : \mathcal{S}\times\mathcal{A} \to [0,1]^n$ and a threshold $\mathbf{b} \in [0, 1/(1-\gamma)]^n$. Under entropy regularization, the agent seeks to maximize the regularized value function $V^\pi_\tau(\rho)$ without violating the utility constraint $U^\pi_{\mathbf{g}}(\rho) \ge \mathbf{b}$, where the discounted utility $U^\pi_{\mathbf{g}}(\rho) := \left(U^\pi_{g_1}(\rho),\dots,U^\pi_{g_n}(\rho)\right) \in \mathbb{R}^n$ is defined by

$$U^\pi_{g_i}(\rho) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g_i(s_t,a_t) \,\middle|\, \pi, s_0 \sim \rho\right]. \tag{10}$$

Equivalently, the agent solves the optimization problem

$$\max_{\pi\in\Pi} \ V^\pi_\tau(\rho) \quad \text{s.t. } U^\pi_{\mathbf{g}}(\rho) \ge \mathbf{b}, \tag{11}$$

where $V^\pi_\tau(\rho) := V^\pi(\rho) + \tau\mathcal{H}(\rho,\pi)$. Consider the associated Lagrangian function $L(\pi,\boldsymbol{\lambda})$ and the dual function $D(\boldsymbol{\lambda})$ defined as:

$$L(\pi,\boldsymbol{\lambda}) := V^\pi_\tau(\rho) + \boldsymbol{\lambda}^T\left(U^\pi_{\mathbf{g}}(\rho) - \mathbf{b}\right), \tag{12a}$$

$$D(\boldsymbol{\lambda}) := \max_{\pi\in\Pi} L(\pi,\boldsymbol{\lambda}), \tag{12b}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the dual variable. For brevity, we omit the dependency of L and D on ρ and τ in the notations.

It can be seen from (12a) that $L(\pi, \boldsymbol{\lambda})$ can be viewed as an entropy-regularized value function with the reward $r_{\boldsymbol{\lambda}}(s,a) \coloneqq r(s,a) + \boldsymbol{\lambda}^T \mathbf{g}(s,a)$ subtracted by the scalar $\boldsymbol{\lambda}^T \mathbf{b}$. Let $V_{\boldsymbol{\lambda}}^{\pi}(\rho)$ (resp. $Q_{\boldsymbol{\lambda}}^{\pi}(s,a)$) denote the value function (resp. Q-function) with the reward function $r_{\boldsymbol{\lambda}}$, i.e. $V_{\boldsymbol{\lambda}}^{\pi}(\rho) \coloneqq V_{\tau}^{\pi}(\rho) + \boldsymbol{\lambda}^T U_{\mathbf{g}}^{\pi}(\rho)$.

It is worth mentioning that (11) is non-convex due to its non-concave objective function and non-convex constraints, thus making the problem challenging to solve. Henceforth, we slightly abuse the notation and still denote $\pi_{\tau}^{\star}$ as an arbitrary optimal policy to the constrained problem (11). Let $\boldsymbol{\lambda}^{\star}$ denote an optimal multiplier, i.e.

$$\boldsymbol{\lambda}^{\star} \coloneqq \arg\min_{\boldsymbol{\lambda} \geq 0} D(\boldsymbol{\lambda}), \tag{13}$$

and $\pi_{\boldsymbol{\lambda}}$ be the Lagrangian maximizer associated with the multiplier $\boldsymbol{\lambda}$, i.e.

$$\pi_{\boldsymbol{\lambda}} \coloneqq \arg\max_{\pi \in \Pi} L(\pi, \boldsymbol{\lambda}). \tag{14}$$

We use the shorthand notations $V_{\tau}^{\star} \coloneqq V_{\tau}^{\pi_{\tau}^{\star}}(\rho)$ and $D_{\tau}^{\star} \coloneqq \min_{\boldsymbol{\lambda} \geq 0} D(\boldsymbol{\lambda})$. As before, we hide the dependency of $\boldsymbol{\lambda}^{\star}$, $\pi_{\boldsymbol{\lambda}}$ on ρ and τ, as well as the dependency of $\pi_{\tau}^{\star}$, $V_{\tau}^{\star}$, $D_{\tau}^{\star}$ on ρ.

3 PROPERTIES OF CMDP WITH ENTROPY REGULARIZATION

Despite its non-convex nature, entropy-regularized CMDPs enjoy desirable properties, which we will discuss below. We refer the reader to the supplement in Appendix A for all the proofs in this section.

Assume that the Slater condition holds, i.e. there exists a strictly feasible policy.

Assumption 3.1 (Slater Condition) *There exist a policy $\overline{\pi} \in \Pi$ and $\boldsymbol{\xi} > 0$ such that $U_{\mathbf{g}}^{\overline{\pi}}(\rho) - \mathbf{b} \geq \boldsymbol{\xi}$.*

The Slater condition is standard in constrained optimization. It holds when the feasible region contains an interior point. In practical, such a point is often easy to find given prior knowledge of the problem. One direct consequence of the Slater condition is the strong duality (Altman, 1999; Paternain et al., 2019).

Lemma 3.2 (Strong Duality) *Under Assumption 3.1, there exist a primal-dual pair $(\pi_{\tau}^{\star}, \boldsymbol{\lambda}^{\star})$ such that $V_{\tau}^{\star} = D_{\tau}^{\star} = L(\pi_{\tau}^{\star}, \boldsymbol{\lambda}^{\star})$.*

In the remainder of the paper, we always assume that $(\pi_{\tau}^{\star}, \boldsymbol{\lambda}^{\star})$ is a primal-dual pair. From the strong duality, we can derive an upper bound on $\boldsymbol{\lambda}^{\star}$.

Lemma 3.3 *Under Assumption 3.1, it holds that*

$$0 \leq \lambda_i^{\star} \leq \frac{V_{\tau}^{\star} - V_{\tau}^{\overline{\pi}}(\rho)}{\xi_i}, \quad \forall i \in [n]. \tag{15}$$

Define

$$\Lambda \coloneqq \left\{ \boldsymbol{\lambda} \mid 0 \leq \lambda_i \leq \frac{V_{\tau}^{\star} - V_{\tau}^{\overline{\pi}}(\rho)}{\xi_i}, \text{ for all } i \in [n] \right\}. \tag{16}$$

Since the dual function $D(\boldsymbol{\lambda})$ is always convex, Lemmas 3.2 and 3.3 together imply that, instead of directly solving the non-convex primal problem (11), one can seek to solve the convex dual problem

$$\min_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda}) \quad s.t.\ \boldsymbol{\lambda} \in \Lambda. \tag{17}$$

However, there are two open problems that need to be addressed. The first one is that although algorithms in convex optimization can be used to solve the dual problem, it is not clear how fast they will converge without discovering key properties of the dual function. The second problem is that optimizing the dual function gives a dual optimality bound, while our goal is find a primal solution and analyze the primal optimality gap together with the constraint violation. In the following sections, we will show that the entropy-regularized CMDP has special structures that can be leveraged to address the above issues.

3.1 Dual Smoothness

In optimization, smoothness plays an important role in establishing the convergence rate of an algorithm. Recall that a function $f : X \to \mathbb{R}$ is said to be ℓ-smooth if

$$\|\nabla f(x_1) - \nabla f(x_2)\|_2 \leq \ell \|x_1 - x_2\|_2, \tag{18}$$

for all $x_1, x_2 \in X$. In constrained optimization, however, smoothness is not always guaranteed, even when the primal problem is convex (Necoara et al., 2019). In addition, while the subgradient of the dual function exists in general, the dual function is not always differentiable due to the non-uniqueness of Lagrangian multipliers.

By leveraging the entropy regularization, we will show that the dual function $D(\boldsymbol{\lambda})$ in CMDPs is both differentiable and ℓ-smooth for some constant $\ell > 0$, under the following assumption on the discounted state visitation distribution.

Assumption 3.4 *The discounted state visitation distribution d_{ρ}^{π} is uniformly bounded away from 0 for all $\pi \in \Pi$, i.e. there exists $d > 0$, such that $d_{\rho}^{\pi}(s) \geq d$, $\forall s \in \mathcal{S}$, $\pi \in \Pi$.*

Assumption 3.4 ensures that the MDP *sufficiently explores* the state space. Since $d_{\rho}^{\pi}(s) \geq (1-\gamma)\rho(s)$, it is satisfied when the initial distribution ρ lies in the interior of the probability simplex $\Delta(\mathcal{S})$. Similar assumptions are used in the prior literature (Agarwal et al., 2021; Mei et al., 2020, 2021).

The following proposition is crucial for the development of our main result.

Proposition 3.5 *For all policy π and $\boldsymbol{\lambda} \geq 0$, it holds that*

$$L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - L(\pi, \boldsymbol{\lambda}) \geq \frac{\tau d}{2(1-\gamma)\ln 2} \|\pi - \pi_{\boldsymbol{\lambda}}\|_2^2. \quad (19)$$

Under Assumption 3.4, Proposition 3.5 implies that the Lagrangian function $L(\pi, \boldsymbol{\lambda})$ has a negative curvature at $\pi_{\boldsymbol{\lambda}}$ in all directions. The proof relies on the soft sub-optimality lemma (cf. Lemma D.4) and a lower bound on the KL divergence (Cover, 1999).

With the quadratic lower bound given by Proposition 3.5, we derive the following result.

Proposition 3.6 *Under Assumptions 3.1 and 3.4, the dual function $D(\boldsymbol{\lambda})$ satisfies the following properties:*

1. *$D(\boldsymbol{\lambda})$ is differentiable and*

$$\begin{aligned}\nabla D(\boldsymbol{\lambda}) =& U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - \mathbf{b} \\ =& \left(U_{g_1}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - b_1, \ldots, U_{g_n}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - b_n\right). \quad (20)\end{aligned}$$

2. *$D(\boldsymbol{\lambda})$ is ℓ-smooth on Λ, where*

$$\ell = \frac{2 \times \ln 2 \times \left(n|\mathcal{A}| + (1-\gamma)^2\sqrt{n|\mathcal{A}|}\right)}{\tau(1-\gamma)^3 d}. \quad (21)$$

Proposition 3.6 asserts that the dual function $D(\boldsymbol{\lambda})$ is not only differentiable but also smooth on Λ. This is desirable since, along with the convexity, it establishes an improved convergence rate compared with the slow convergence rate of sub-gradient methods. We provide a short proof sketch for Proposition 3.6 below:

1. As subgradients of the dual function always exist for continuous problems, the differentiability follows from the uniqueness of the Lagrangian maximizer $\pi_{\boldsymbol{\lambda}}$ for every $\boldsymbol{\lambda} \in \Lambda$ (Floudas, 1995)[1].

2. The smoothness of $D(\boldsymbol{\lambda})$ is the joint result of the Lipschitz continuity of $U_{\mathbf{g}}^{\pi}(\rho)$ with respect to π (cf. Lemma 2.1) and the Lipschitz continuity of $\pi_{\boldsymbol{\lambda}}$ with respect to $\boldsymbol{\lambda}$, i.e. $\|\pi_{\boldsymbol{\lambda}_1} - \pi_{\boldsymbol{\lambda}_2}\|_2 \leq \ell_\Lambda \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2$ for some $\ell_\Lambda > 0$. To prove the latter, the main idea is to use the quadratic lower bound given by Proposition 3.5 to conclude that $\pi_{\boldsymbol{\lambda}}$ is a second-order strict local maximum. After that, we apply a standard result from perturbation analysis which states that $\pi_{\boldsymbol{\lambda}}$ is Lipschitz stable at $\boldsymbol{\lambda}$ (Bonnans and Shapiro, 2013).

[1] Although more than enough, Proposition 3.5 under Assumption 3.4 provides an intuitive way to think about the uniqueness of $\pi_{\boldsymbol{\lambda}}$.

3.2 Optimality Gap And Constraint Violation

Given a candidate solution π to the CMDP problem in (11), our primary measures of the quality of the solution π are the primal optimality gap $|V_\tau^\pi(\rho) - V_\tau^\star|$, and the constraint violation $\max_{i\in[n]}\left[b_i - U_{g_i}^\pi(\rho)\right]_+$, where $[x]_+ := \max\{x, 0\}$. However, dual-descent based methods could only guarantee a convergence bound in terms of the dual optimality gap $D(\boldsymbol{\lambda}) - D_\tau^\star$. In general, there is no guarantee that an ε-optimal dual solution $\boldsymbol{\lambda}$, namely $D(\boldsymbol{\lambda}) - D_\tau^\star \leq \varepsilon$, would imply an $\mathcal{O}\left(\varepsilon^k\right)$ bound either on the primal optimality gap or on the constraint violation for the associated primal solution $\pi_{\boldsymbol{\lambda}}$ defined in (14), for some $k \in (0, 1]$.

However, in light of the entropy regularization, it is possible to show that an ε error bound for dual functions would yield an $O\left(\sqrt{\varepsilon}\right)$ error bound for the primal optimality gap and the constraint violation. We summarize the results in the following proposition.

Proposition 3.7 *Suppose that Assumptions 3.1 and 3.4 hold. If $\boldsymbol{\lambda} \geq 0$ is an ε-optimal multiplier, i.e. $D(\boldsymbol{\lambda}) - D_\tau^\star \leq \varepsilon$, then there exist constants C_1 and C_2 such that the associated Lagrangian maximizer $\pi_{\boldsymbol{\lambda}}$ satisfies*

$$\|\pi_{\boldsymbol{\lambda}} - \pi_\tau^\star\|_2 \leq C_1\sqrt{\varepsilon}, \quad (22a)$$

$$|V_\tau^{\pi_{\boldsymbol{\lambda}}}(\rho) - V_\tau^\star| \leq 2\varepsilon + \ell_c C_1 C_2 \sqrt{\varepsilon}, \quad (22b)$$

$$\max_{i\in[n]}\left[b_i - U_{g_i}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right]_+ \leq \ell_c C_1 \sqrt{\varepsilon}, \quad (22c)$$

where ℓ_c is the Lipschitz constant defined in (3).

The values of the problem-dependent constants C_1 and C_2 can be found in Appendix A.

In a nutshell, Proposition 3.7 enables the conversion of the dual optimality bound to primal metrics of interests, at the cost of enlarging the sub-optimality by a square root. This is a non-trivial result and it does not hold in a general setting without the entropy regularization. The proof of (22a) relies on the quadratic lower bound given in Proposition 3.5. Then, using the Lipschitz continuity of $U_{\mathbf{g}}^\pi(\rho)$ with respect to π (cf. Lemma 2.1), we can derive the bound (22c) on the constraint violation. Finally, (22b) can be obtained with some primal-dual properties.

4 FIRST-ORDER DUAL-DESCENT ALGORITHM

As shown in Section 3, the dual function $D(\boldsymbol{\lambda})$ for entropy-regularized CMDPs enjoys desirable properties, including the differentiability, the smoothness and the decomposition of the dual optimality gap. These favorable properties of entropy-regularized CMDPs motivate

us to use a dual-descent approach to solve the dual problem (17). In particular, we choose first-order methods, e.g. gradient projection method or Frank-Wolfe algorithm, while using the Natural Policy Gradient (NPG) algorithm as a subroutine for evaluating $D(\boldsymbol{\lambda})$ as well as $\nabla D(\boldsymbol{\lambda})$. To streamline the presentation, we mainly focus on the gradient projection method with the Nesterov acceleration as an example in the remainder of the paper. We begin with a brief introduction about the NPG algorithm.

4.1 Preliminary Tools

NPG Algorithm With Entropy Regularization To optimize an unconstrained value function with respect to the policy, one commonly used first-order method is the Natural Policy Gradient algorithm (Kakade, 2001), which deploys a pre-conditioned gradient update and regularizes the descent direction by the Fisher-information matrix $\mathcal{F}_\rho^\theta$ (cf. Appendix B.1):

$$\theta \leftarrow \theta + \eta \left(\mathcal{F}_\rho^\theta\right)^\dagger \nabla_\theta V^{\pi_\theta}(\rho), \tag{23}$$

where η is the step-size and $(A)^\dagger$ denotes the Moore–Penrose inverse of a matrix A.

In the entropy regularized setting, the update scheme is obtained by replacing $\nabla_\theta V^{\pi_\theta}(\rho)$ in (23) with $\nabla_\theta V_\tau^{\pi_\theta}(\rho)$. Under the soft-max parameterization, the associated policy update has a fairly direct form, which is surprisingly independent from the initial distribution ρ:

$$\pi^{(t+1)}(a|s) \propto \left(\pi^{(t)}(a|s)\right)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^{\pi^{(t)}}(s,a)}{1-\gamma}\right), \tag{24}$$

where we use the shorthand $\pi^{(t)}$ for the soft-max parameterized policy with respect to $\theta^{(t)}$, and Q_τ^π is the soft Q-function defined in (9). The right-hand side of (24) can be normalized by multiplying a factor $Z^{(t)}(s)$, defined as

$$Z^{(t)}(s) \coloneqq \sum_{a\in\mathcal{A}} \left(\pi^{(t)}(a|s)\right)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^{\pi^{(t)}}(s,a)}{1-\gamma}\right), \tag{25}$$

to make $\pi^{(t+1)}$ be a valid distribution.

Cen et al. (2021) proved the global linear convergence of the entropy-regularized NPG method with a constant step-size. In particular, the error bound $\left\|\log \pi_\tau^\star - \log \pi^{(t)}\right\|_\infty \le \varepsilon$ can be achieved in

$$\frac{1}{1-\gamma}\log\left(\frac{2\left\|Q_\tau^\star - Q_\tau^{\pi^{(0)}}\right\|_\infty}{\varepsilon\tau}\right), \tag{26}$$

iterations with the step-size $\eta = (1-\gamma)/\tau$, where $\pi_\tau^\star$ is the optimal policy and $Q_\tau^\star(s,a) \coloneqq Q_\tau^{\pi_\tau^\star}(s,a)$ is the associated optimal Q-function. Furthermore, they proved that the convergence rate becomes quadratic around the optimum. We refer the reader to Appendix B.1 for more details.

Accelerated Gradient Projection Method with Inexact Gradient The Nesterov acceleration is a momentum-based approach that can be used to modify a gradient descent-type method to improve its convergence (Nesterov, 1983, 2013). Consider the optimization problem

$$\min_x f(x) \quad s.t. \quad x \in X \tag{27}$$

where $f(x)$ is convex and differentiable, and X is a convex set. The accelerated gradient projection method takes the update rule

$$\begin{cases} x^{(k+1)} = \mathcal{P}_X\left(y^{(k)} - \alpha^k \nabla f(y^{(k)})\right) \\ y^{(k)} = x^{(k)} + \beta_k\left(x^{(k)} - x^{(k-1)}\right) \end{cases}, \ k = 0, 1\ldots \tag{28}$$

where $\mathcal{P}_X$ denotes the projection onto the set X, defined as $\mathcal{P}_X(y) \coloneqq \arg\min_{x\in X}\|x-y\|_2$, and $\{\beta_k\}$ is chosen in a particular way to accelerate the convergence. The iteration (28) first computes an extrapolation point $y^{(k)}$ and then performs the gradient projection update on $y^{(k)}$ to find the next point $x^{(k+1)}$. It coincides with the standard gradient projection method when $\beta_k = 0$. For a convex and smooth function f, the accelerated gradient projection method (28) achieves an error bound of $\mathcal{O}\left(1/T^2\right)$ in T iterations (Nesterov, 2013).

When the gradient evaluation is inexact with a bounded error δ, i.e. we have access to some function $h : X \to \mathbb{R}^n$ such that $\|\nabla f(x) - h(x)\|_2 \le \delta$ for all $x \in X$, Schmidt et al. (2011) proved that the accelerated gradient projection method still works with the slightly different error bound $\mathcal{O}\left(1/T^2 + T^2\delta^2 + \delta\right)$. Despite the $\mathcal{O}\left(1/T^2\right)$ shrinking term, there is an accumulated error incurred by the inexact gradient. We refer the reader to Proposition B.3 in Appendix B for a formal statement.

4.2 Accelerated Gradient Projection Method With NPG Subroutine

Before presenting our method, we first note that the feasible region Λ, as defined in (16), makes the dual problem (17) amenable to many constrained optimization methods. Especially, the projection operator $\mathcal{P}_\Lambda(\cdot)$ maps a point $\boldsymbol{\lambda}$ coordinate-wisely onto Λ such that

$$\left(\mathcal{P}_\Lambda(\boldsymbol{\lambda})\right)_i = \text{Median}\left\{0, \frac{V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)}{\xi_i}, \lambda_i\right\} \tag{29}$$

where $\text{Median}\{\cdot,\cdot,\cdot\}$ returns the median of the input numbers.

Algorithm 1 Accelerated Gradient Projection Method with NPG Subroutine

1: **Input:** Initialization $\boldsymbol{\lambda}^{(-1)}, \boldsymbol{\lambda}^{(0)}$, $\widetilde{\pi}_{\boldsymbol{\mu}^{(-1)}}$; step-size $\{\alpha_k\}_{k\geq 0}$, η; extrapolation weight $\{\beta_k\}_{k\geq 0}$; maximum number of iterations N_1, N_2, N_3.
2: **for** $t = 0, 1, 2, \ldots, N_1 - 1$ **do**
3: Compute the extrapolation point: $\boldsymbol{\mu}^{(t)} = \boldsymbol{\lambda}^{(t)} + \beta_k \left(\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^{(t-1)}\right)$.
4: Estimate the optimal policy $\pi_{\boldsymbol{\mu}^{(t)}}$ for problem (14) through the natural policy gradient subroutine: $\widetilde{\pi}_{\boldsymbol{\mu}^{(t)}} \leftarrow \mathrm{NPG}_{Sub}\left(\boldsymbol{\mu}^{(t)}, \widetilde{\pi}_{\boldsymbol{\mu}^{(t-1)}}, \eta, N_2\right)$.
5: Compute the approximate gradient at $\boldsymbol{\mu}^{(t)}$: $\widetilde{\nabla} D\left(\boldsymbol{\mu}^{(t)}\right) := U_{\mathbf{g}}^{\widetilde{\pi}_{\boldsymbol{\mu}^{(t)}}}(\rho) - \mathbf{b}$.
6: Take a gradient projection step at $\mu^{(t)}$: $\boldsymbol{\lambda}^{(t+1)} \leftarrow \mathcal{P}_{\Lambda}\left(\boldsymbol{\mu}^{(t)} - \alpha^k \widetilde{\nabla} D\left(\boldsymbol{\mu}^{(t)}\right)\right)$, as defined by (29).
7: **end for**
8: Recover the policy from the dual variable: $\widetilde{\pi}_{\boldsymbol{\lambda}^{(N_1)}} \leftarrow \mathrm{NPG}_{Sub}\left(\boldsymbol{\lambda}^{(N_1)}, \widetilde{\pi}_{\boldsymbol{\mu}^{(N_1-1)}}, \eta, N_3\right)$.

The proposed method works in two loops. In the outer loop, we perform the accelerated gradient projection method on the dual function $D(\boldsymbol{\lambda})$, whereas we use the natural policy gradient method in the inner loop to evaluate $D(\boldsymbol{\lambda})$ by maximizing the Lagrangian $L(\pi, \boldsymbol{\lambda})$ with respect to π. We summarize the details of our method in Algorithm 1.

Specifically, in line 3, we compute the extrapolation point $\boldsymbol{\mu}^{(t)}$. Then, in line 4, we estimate the corresponding Lagrangian maximizer $\pi_{\boldsymbol{\mu}^{(t)}}$, defined in 14, using the natural policy gradient subroutine, which is displayed in Algorithm 2. With the estimated policy $\widetilde{\pi}_{\boldsymbol{\mu}^{(t)}}$, we evaluate the gradient $\nabla D\left(\boldsymbol{\mu}^{(t)}\right)$ by substituting the policy into the utility function in line 5. In line 6, we perform the gradient projection update at $\boldsymbol{\mu}^{(t)}$ using the estimated gradient $\widetilde{\nabla} D\left(\boldsymbol{\mu}^{(t)}\right)$. We remark that, as $V_\tau^\star$ is generally unknown, the projection $\mathcal{P}_\Lambda$ may not be precisely done in practical. Alternatively, one can perform the projection onto $\widetilde{\Lambda} := \{\boldsymbol{\lambda} \mid 0 \leq \lambda_i \leq (2 + 2\tau \log \mathcal{A})/((1-\gamma)\xi_i), \text{ for all } i \in [n]\}$. Since the difference $V_\tau^\star - V_\tau^{\bar{\pi}}(\rho)$ is upper bounded by $(2 + 2\tau \log \mathcal{A})/(1-\gamma)$, it holds that $\boldsymbol{\lambda}^\star \in \Lambda \subseteq \widetilde{\Lambda}$. This reduction would not influence the order of convergence. Finally, upon the termination of the outer loop, we recover the primal variable (policy) from the dual variable by running the NPG subroutine for N_3 iterations in line 8.

5 CONVERGENCE ANALYSIS

In this section, we analyze the convergence of Algorithm 1. The complete proofs of results in this section are postponed to Appendix B.3.

Algorithm 2 Natural Policy Gradient Subroutine (NPG_{Sub})

1: **Input:** Multiplier $\boldsymbol{\lambda}$; initial policy $\pi^{(0)}$; step-size η; maximum number of iterations N.
2: **for** $t = 0, 1, 2, \ldots, N-1$ **do**
3: Compute the soft Q-function associated with the Lagrangian: $Q_{\boldsymbol{\lambda}}^{\pi^{(t)}}$ of policy $\pi^{(t)}$.
4: Update the policy with $Q_{\boldsymbol{\lambda}}^{\pi^{(t)}}$ through (24).
5: **end for**

A simple insight is that the NPG subroutine in Algorithm 1 computes the optimal policy at a linear rate in the inner loop and the accelerated gradient projection algorithm converges in $\mathcal{O}\left(1/\sqrt{\varepsilon}\right)$ rate in the outer loop, leading to the overall convergence rate of $\widetilde{\mathcal{O}}\left(1/\sqrt{\varepsilon}\right)$. Then, we obtain the desired $\widetilde{\mathcal{O}}(1/\varepsilon)$ rate in terms of the primal optimality gap and constraint violation by applying Proposition 3.7.

The above high-level technique requires subtle technicalities to be addressed here. We begin with the following proposition, which evaluates the accuracy of the gradient estimator defined in line 5 of Algorithm 1.

Proposition 5.1 *Suppose that π is an approximate solution to (14) such that $\|\log \pi - \log \pi_{\boldsymbol{\lambda}}\|_\infty \leq \varepsilon$. The gradient estimator defined as $\widetilde{\nabla} D(\boldsymbol{\lambda}) := U_{\mathbf{g}}^{\pi}(\rho) - \mathbf{b} = \left(U_{g_1}^{\pi}(\rho) - b_1, \ldots, U_{g_n}^{\pi}(\rho) - b_n\right)$ satisfies the inequality $\left\|\widetilde{\nabla} D(\lambda) - \nabla D(\lambda)\right\|_2 \leq \sqrt{n}|\mathcal{A}|\varepsilon/(1-\gamma)^2$.*

This result can be deduced from the inequality $\|\pi - \pi_{\boldsymbol{\lambda}}\|_\infty \leq \|\log \pi - \log \pi_{\boldsymbol{\lambda}}\|_\infty \leq \varepsilon$ and the performance difference lemma of an unregularized value function (Lemma D.3). Recall that $L(\pi, \boldsymbol{\lambda}) = V_{\boldsymbol{\lambda}}^{\pi}(\rho) - \boldsymbol{\lambda}^T \mathbf{b}$, where $V_{\boldsymbol{\lambda}}^{\pi}(\rho) = V_\tau^\pi(\rho) + \boldsymbol{\lambda}^T U_{\mathbf{g}}^{\pi}(\rho)$ is the value function with the reward $r_{\boldsymbol{\lambda}}(s,a) := r(s,a) + \boldsymbol{\lambda}^T \mathbf{g}(s,a)$ (cf. Section 2)). Therefore, Proposition 5.1 together with (26), implies that running Algorithm 2 with the step-size $\eta = (1-\gamma)/\tau$ for

$$\frac{1}{1-\gamma} \log\left(\frac{2\sqrt{n}\,|\mathcal{A}|\left\|Q_{\boldsymbol{\lambda}}^{\pi_{\boldsymbol{\lambda}}} - Q_{\boldsymbol{\lambda}}^{\pi^{(0)}}\right\|_\infty}{\delta(1-\gamma)^2\tau}\right) \tag{30}$$

iterations can guarantee a δ-accurate gradient estimation $\widetilde{\nabla} D(\boldsymbol{\lambda})$, i.e. $\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2 \leq \delta$.

Below, we present our main convergence result of Algorithm 1.

Theorem 5.2 *Suppose that Assumptions 3.1 and 3.4 hold. For every $\varepsilon_1 > 0$, there exist some constants C_1 and $C_2 > 0$ such that Algorithm 1 with a random initialization and the parameters $\eta = (1-\gamma)/\tau$, $\alpha_k = 1/\ell$, $\beta_k = (k-1)/(k+2)$, $N_1 = T$, $N_2 = \mathcal{O}(\log T)$ and*

$N_3 = \mathcal{O}(\log 1/\varepsilon_1)$ *returns a solution pair* $(\pi, \boldsymbol{\lambda})$ *such that*

$$D(\boldsymbol{\lambda}) - D_\tau^\star \le \varepsilon_0, \tag{31a}$$

$$\|\pi - \pi_\tau^\star\|_2 \le C_1\sqrt{\varepsilon_0} + \varepsilon_1, \tag{31b}$$

$$|V_\tau^\pi(\rho) - V_\tau^\star| \le 2\varepsilon_0 + \ell_c C_1 C_2\sqrt{\varepsilon_0} + \left(\ell_c C_2 + \frac{3\gamma}{2\tau\sqrt{n}}\right)\varepsilon_1, \tag{31c}$$

$$\max_{i\in[n]}\left[b_i - U_{g_i}^\pi(\rho)\right]_+ \le \ell_c\left(C_1\sqrt{\varepsilon_0} + \varepsilon_1\right), \tag{31d}$$

where

$$\varepsilon_0 = \frac{2\ell}{(T+1)^2}\left(\left\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}^\star\right\|_2 + 1\right)^2, \tag{32}$$

and where ℓ *is the smoothness factor defined in (21) and* ℓ_c *is the Lipschitz constant defined in (3). The total iteration complexity is* $N_1 \times N_2 + N_3 = \widetilde{\mathcal{O}}(T)$ *with primal error bounds* $\mathcal{O}(1/T)$ *given by (31b)-(31d), and a dual error bound* $\mathcal{O}\left(1/T^2\right)$ *given by (31a).*

The values of the parameters N_2, N_3, and problem-dependent constants C_1, C_2 can be found in Theorem B.5, Appendix B.3, where we restate the theorem.

Theorem 5.2 shows that Algorithm 1 achieves a global convergence with the rate $\widetilde{\mathcal{O}}(1/\varepsilon)$. Specifically, it shows that with $\widetilde{\mathcal{O}}(T)$ number of iterations in total, Algorithm 1 generates a solution with $\mathcal{O}(1/T)$ error bounds in terms of the policy (primal variable), primal optimality gap, and constraint violation, as well as $\mathcal{O}(1/T^2)$ error in terms of the dual optimality gap.

We briefly describe the intuition behind the proof of Theorem 5.2 below.

1. Firstly, the linear convergence of the natural policy gradient method (cf. Proposition B.1) and Proposition 5.1 imply that running the NPG subroutine for $N_2 = \mathcal{O}(\log T)$ iterations in the inner loop guarantees a sufficiently accurate estimation of $\nabla D(\boldsymbol{\lambda})$.

2. Then, we apply the convergence result by Schmidt et al. (2011) (refer to Section 4.1 and Appendix B.2), which implies the dual optimality gap $\mathcal{O}\left(1/T^2 + T^2\delta^2 + \delta\right)$, where δ is the gradient estimation error. Since the NPG subroutine converges linearly, we can suppress the constant in $N_2 = \mathcal{O}(\log T)$, and make δ sufficiently small such that the dual optimality gap equals $\mathcal{O}\left(1/T^2\right)$.

3. Let $\boldsymbol{\lambda}$ denote the dual variable returned by the for loop. Running the NPG subroutine in line 8 for additional $N_3 = \mathcal{O}(\log(1/\varepsilon_1))$ iterations guarantees an ε_1-approximate solution π for the Lagrangian maximizer $\pi_{\boldsymbol{\lambda}}$. Again, we can make ε_1 sufficiently small by suppressing the constant in N_3.

4. Finally, by applying Proposition 3.7 and the triangular inequality, we prove (31b), stating that π is $\mathcal{O}(1/T)$-optimal. We bound the constraint violation (31d) by using the Lipschitz continuity of $U_{\mathbf{g}}^\pi(\rho)$ with respect to π, and bound the primal optimality gap (31c) by using some primal-dual properties[2].

Remark 5.3 (Quadratic Convergence of NPG) *Our analysis of Algorithm 1 in this section is inspired by the global linear convergence of the entropy-regularized NPG method. However, as Cen et al. (2021) proved, the NPG method achieves a quadratic convergence around the optimum (cf. Proposition B.2). Therefore, it may be possible to improve the hidden constants in* $N_2 = \mathcal{O}(\log T)$ *and* $N_3 = \mathcal{O}(\log(1/\varepsilon_1))$ *under extra assumptions.*

So far, we have only studied the entropy-regularized CMDP. However, adding entropy induces bias to the optimal solution of the standard unregularized CMDP. A standard way to deal with this mismatch issue is to choose the regularization parameter τ to be sufficiently small. The following corollary shows that we can compute a near-optimal policy with the rate $\widetilde{\mathcal{O}}\left(1/\sqrt{T}\right)$ for both the optimality gap and the constraint violation for the standard CMDP.

Corollary 5.4 *Suppose that Assumptions 3.1 and 3.4 hold. Then, Algorithm 1 with the choice* $\tau = \mathcal{O}(\varepsilon)$ *computes a solution* π *for the standard CMDP such that*

$$\left|V^{\pi^\star}(\rho) - V^\pi(\rho)\right| = \mathcal{O}(\varepsilon), \tag{33a}$$

$$\max_{i\in[n]}\left[b_i - U_{g_i}^\pi(\rho)\right]_+ = \mathcal{O}(\varepsilon), \tag{33b}$$

in $\widetilde{\mathcal{O}}\left(1/\varepsilon^2\right)$ *iterations, where* $\pi^\star$ *is an optimal policy to the standard CMDP.*

The proof of Corollary 5.4 relies on the following sandwich bound

$$V^{\pi_\tau^\star}(\rho) \le V^{\pi_\star}(\rho) \le V^{\pi_\tau^\star}(\rho) + \frac{\tau}{1-\gamma}\log|\mathcal{A}|. \tag{34}$$

6 CMDPS WITH A SINGLE CONSTRAINT

Since we can convert the dual optimality bound to primal metrics of interests with little extra effort (cf. Proposition 3.7), the overall complexity relies on how fast we can solve the dual problem (17). For the special case where $n = 1$, the dual problem amounts to optimizing a convex function on an closed interval, which

[2]The techniques are similar to those in the proof of Proposition 3.7.

can be efficiently solved by the bisection method. Due to space restrictions, we refer the reader to Appendix C for more details about the algorithm (cf. Algorithm 3). We state the result in the following theorem.

Theorem 6.1 *Suppose that Assumptions 3.1 and 3.4 hold. When $n = 1$, for every ε_0, $\varepsilon_1 > 0$, Algorithm 3 returns a solution pair (π, λ) satisfying (31a)-(31d) in at most $\mathcal{O}\left(\log^2(1/\varepsilon_0) + \log(1/\varepsilon_1)\right)$ iterations.*

By leveraging the linear convergence, we can derive a result analogous to Corollary 5.4 for the standard CMDP, but with a linear rate.

Corollary 6.2 *Suppose that Assumptions 3.1 and 3.4 hold. Then, Algorithm 3 with the choice $\tau = \mathcal{O}(\varepsilon)$ computes a solution π for the standard CMDP satisfying (33a)-(33b), in $\mathcal{O}\left(\log^2(1/\varepsilon)\right)$ iterations.*

We refer the reader to Appendix C for the formal statements of Theorem 6.1 and Corollary 6.2 as well as their proofs.

7 CONCLUSION

In this paper, we showed that entropy regularization induces desirable properties to CMDPs from an optimization perspective. In particular, the Lagrangian dual function of CMDPs is smooth and an $\mathcal{O}(\varepsilon)$ error bound for the dual optimality gap yields an $\mathcal{O}\left(\sqrt{\varepsilon}\right)$ error bound for the primal optimality gap and the constraint violation. In addition, we proposed a novel accelerated dual-descent algorithm and proved that it achieves a global convergence with the rate $\widetilde{\mathcal{O}}(1/\varepsilon)$ for both the optimality gap and the constraint violation. It remains as an open question whether similar improved convergence results for the entropy-regularized CMDPs can be obtained with a sample-based policy gradient.

ACKNOWLEDGEMENTS

This work was supported by grants from AFOSR, ARO, ONR, NSF and C3.ai Digital Transformation Institute.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76.

Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

Bhatnagar, S. and Lakshmanan, K. (2012). An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708.

Bonnans, J. F. and Shapiro, A. (2013). *Perturbation analysis of optimization problems*. Springer Science & Business Media.

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

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2021). Fast global convergence of natural policy gradient methods with entropy regularization.

Chen, Y., Dong, J., and Wang, Z. (2021). A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. (2017). Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. (2021). Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. R. (2020). Natural policy gradient primal-dual method for constrained markov decision processes. In *Conference on Neural Information Processing Systems*.

Efroni, Y., Mannor, S., and Pirotta, M. (2020). Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*.

Fisac, J. F., Akametalu, A. K., Zeilinger, M. N., Kaynama, S., Gillula, J., and Tomlin, C. J. (2018). A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752.

Floudas, C. A. (1995). *Nonlinear and mixed-integer optimization: fundamentals and applications*. Oxford University Press.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

Kakade, S. M. (2001). A natural policy gradient. *Advances in neural information processing systems*, 14.

Mei, J., Gao, Y., Dai, B., Szepesvari, C., and Schuurmans, D. (2021). Leveraging non-uniformity in first-order non-convex optimization. *arXiv preprint arXiv:2105.06072*.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. *arXiv preprint arXiv:1702.08892*.

Necoara, I., Patrascu, A., and Glineur, F. (2019). Complexity of first-order inexact lagrangian and penalty methods for conic convex programming. *Optimization Methods and Software*, 34(2):305–335.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547.

Paternain, S., Calvo-Fullana, M., Chamon, L. F., and Ribeiro, A. (2019). Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*.

Schmidt, M., Roux, N. L., and Bach, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. *arXiv preprint arXiv:1109.2415*.

Uchibe, E. and Doya, K. (2007). Constrained reinforcement learning from intrinsic and extrinsic rewards. In *2007 IEEE 6th International Conference on Development and Learning*, pages 163–168. IEEE.

Williams, R. J. and Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

Xu, T., Liang, Y., and Lan, G. (2021). Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR.

Zang, H., Li, X., Zhang, L., Zhao, P., and Wang, M. (2020). Teac: Intergrating trust region and max entropy actor critic for continuous control.

Zhang, X., Zhang, K., Miehling, E., and Basar, T. (2019). Non-cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 32.

Ziebart, B. D. (2010). *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University.

Supplementary Material: A Dual Approach to Constrained Markov Decision Processes with Entropy Regularization

A Proofs of Results in Sections 2 and 3

Lemma A.1 (Restatement of Lemma 2.1) *For an unregularized value function $V^\pi(\rho)$ with the reward function $r(s,a) \in [0,1]$, it holds that*

$$|V^{\pi_1}(\rho) - V^{\pi_2}(\rho)| \le \ell_c \left\|\pi_1 - \pi_2\right\|_2, \tag{35}$$

for arbitrary policies π_1 and π_2, where $\ell_c = \sqrt{|\mathcal{A}|}/(1-\gamma)^2$.

Proof. It follows from the policy gradient for the direct parameterization (Lemma D.1) that

$$\frac{\partial V^\pi(\rho)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d_\rho^\pi(s) Q^\pi(s,a). \tag{36}$$

Thus, we can bound $\nabla_\pi V^\pi(\rho)$ as

$$\begin{aligned}
\left\|\nabla_\pi V^\pi(\rho)\right\|_2 &= \frac{1}{1-\gamma}\sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}} \left(d_\rho^\pi(s) Q^\pi(s,a)\right)^2} \\
&\le \frac{\max_{s\in\mathcal{S},a\in\mathcal{A}} Q^\pi(s,a)}{1-\gamma}\sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}} \left(d_\rho^\pi(s)\right)^2} \\
&\overset{(i)}{\le} \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\left\|d_\rho^\pi(\cdot)\right\|_2 \\
&\overset{(ii)}{\le} \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2} =: \ell_c,
\end{aligned} \tag{37}$$

where (i) uses $Q^\pi(s,a) \le 1/1-\gamma$ and (ii) is because of $\left\|d_\rho^\pi(\cdot)\right\|_2 \le \left\|d_\rho^\pi(\cdot)\right\|_1 = 1$. Then, (35) follows from

$$|V^{\pi_1}(\rho) - V^{\pi_2}(\rho)| \le \sup_\pi \left\{\left\|\nabla_\pi V^\pi(\rho)\right\|_2\right\} \left\|\pi_1 - \pi_2\right\|_2 \le \ell_c \left\|\pi_1 - \pi_2\right\|_2. \tag{38}$$

This completes the proof. □

Lemma A.2 (Restatement of Lemma 3.2) *Under Assumption 3.1, there exist a primal-dual pair $(\pi_\tau^\star, \boldsymbol{\lambda}^\star)$ such that $V_\tau^\star = D_\tau^\star = L(\pi_\tau^\star, \boldsymbol{\lambda}^\star)$.*

Proof. We refer the reader to (Paternain et al., 2019) for a proof of the strong duality. □

Lemma A.3 (Restatement of Lemma 3.3) *Under Assumption 3.1, it holds that*

$$0 \le \lambda_i^\star \le \frac{V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)}{\xi_i}, \quad \forall i \in [n]. \tag{39}$$

Proof. Let $C \in \mathbb{R}$. For every $\boldsymbol{\lambda} \ge 0$ such that $D(\boldsymbol{\lambda}) \le C$, one can write

$$\begin{aligned}
C \ge D(\boldsymbol{\lambda}) &\overset{(i)}{\ge} V_\tau^{\overline{\pi}}(\rho) + \sum_{i=1}^n \lambda_i \left(U_{g_i}^{\overline{\pi}}(\rho) - b_i\right) \\
&\overset{(ii)}{\ge} V_\tau^{\overline{\pi}}(\rho) + \sum_{i=1}^n \lambda_i \xi_i,
\end{aligned} \tag{40}$$

where (i) follows from the definition of $D(\boldsymbol{\lambda})$ and (ii) is due to Assumption 3.1. Since $\boldsymbol{\xi} > 0$ and $\boldsymbol{\lambda} \geq 0$, (40) gives rise to the bound $0 \leq \lambda_i \leq \left(C - V_\tau^{\overline{\pi}}(\rho)\right)/\xi_i$, for $i = 1, 2, \ldots, n$. Now, by letting $C = V_\tau^\star$, it results from the strong duality that $\{\boldsymbol{\lambda} \geq 0 \mid D(\boldsymbol{\lambda}) \leq C\}$ becomes the set of optimal dual variables. This completes the proof. □

Proposition A.4 (Restatement of Proposition 3.5) *For all policy π and $\boldsymbol{\lambda} \geq 0$, it holds that*

$$L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - L(\pi, \boldsymbol{\lambda}) \geq \frac{\tau d}{2(1-\gamma)\ln 2} \left\|\pi - \pi_{\boldsymbol{\lambda}}\right\|_2^2. \tag{41}$$

Proof. Recall that $L(\pi, \boldsymbol{\lambda}) = V_{\boldsymbol{\lambda}}^\pi(\rho) - \boldsymbol{\lambda}^T \mathbf{b}$, where $V_{\boldsymbol{\lambda}}^\pi(\rho) = V_\tau^\pi(\rho) + \boldsymbol{\lambda}^T U_{\mathbf{g}}^\pi(\rho)$ is the value function with the reward $r_{\boldsymbol{\lambda}}(s,a) \coloneqq r(s,a) + \boldsymbol{\lambda}^T \mathbf{g}(s,a)$ (cf. Section 2). The soft sub-optimality gap (Lemma D.4) then gives

$$L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - L(\pi, \boldsymbol{\lambda}) = V_{\boldsymbol{\lambda}}^{\pi_{\boldsymbol{\lambda}}}(\rho) - V_{\boldsymbol{\lambda}}^\pi(\rho) = \frac{\tau}{1-\gamma} \sum_{s \in \mathcal{S}} d_\rho^\pi(s) D_{\mathrm{KL}}\left[\pi(\cdot|s) \mid \pi_{\boldsymbol{\lambda}}(\cdot|s)\right], \tag{42}$$

where $D_{\mathrm{KL}}\left[P(\cdot) \mid Q(\cdot)\right] \coloneqq \sum_x P(x)\left(\log P(x) - \log Q(x)\right)$ is the KL divergence between probability distributions $P(\cdot)$ and $Q(\cdot)$. Now, we use a well-known bound relating the KL divergence to the vector 1-norm (Cover, 1999):

$$D_{\mathrm{KL}}\left[P(\cdot) \mid Q(\cdot)\right] \geq \frac{1}{2\ln 2} \|P(\cdot) - Q(\cdot)\|_1^2. \tag{43}$$

Combine (42) and (43) yields

$$L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - L(\pi, \boldsymbol{\lambda}) \geq \frac{\tau}{2(1-\gamma)\ln 2} \sum_{s \in \mathcal{S}} d_\rho^\pi(s) \|\pi(\cdot|s) - \pi_{\boldsymbol{\lambda}}(\cdot|s)\|_1^2 \overset{(i)}{\geq} \frac{\tau d}{2(1-\gamma)\ln 2} \left\|\pi - \pi_{\boldsymbol{\lambda}}\right\|_2^2, \tag{44}$$

where (i) is due to $\|\cdot\|_1 \geq \|\cdot\|_2$ and Assumption 3.4. This completes the proof. □

Proposition A.5 (Restatement of Proposition 3.6) *Under Assumptions 3.1 and 3.4, the dual function $D(\boldsymbol{\lambda})$ satisfies the following properties:*

1. *$D(\boldsymbol{\lambda})$ is differentiable and*

$$\nabla D(\boldsymbol{\lambda}) = U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - \mathbf{b} = \left(U_{g_1}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - b_1, \ldots, U_{g_n}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - b_n\right). \tag{45}$$

2. *$D(\boldsymbol{\lambda})$ is ℓ-smooth on Λ, where*

$$\ell = \frac{2 \times \ln 2 \times \left(n|\mathcal{A}| + (1-\gamma)^2\sqrt{n|\mathcal{A}|}\right)}{\tau(1-\gamma)^3 d}. \tag{46}$$

The proof of Proposition A.5 relies on the following result by Bonnans and Shapiro (2013).

Proposition A.6 (Lipschitz stability of parametric local maximizersz) *Given a set $T \subset \mathbb{R}^p$, consider a parametric optimization problem $P(t)$ with $t \in T$, stated as*

$$\max_{x \in F} \; f(x,t) \quad s.t.\; F = \{x \in \mathbb{R}^n \mid h_j(x) \leq 0,\; j = 1, 2, \ldots, m\}, \tag{47}$$

where f, h_j are twice continuously differentiable and $F \neq \varnothing$. For every $\bar{t} \in T$, if $\overline{x} = x\left(\bar{t}\right)$ is a strict local maximizer of $P(\bar{t})$ of order 2, i.e. $\nabla_{xx}^2 f(\overline{x}, \bar{t}) \succeq w_1 I_n$ for some $w_1 > 0$, then there exist ε, δ, $L > 0$ such that for all $t \in B_\varepsilon(\bar{t}) \coloneqq \{t \mid \|t - \bar{t}\|_2 < \varepsilon\}$, there exists at least one local maximizer $x(t) \in B_\delta(\overline{x})$ of $P(t)$ and for each such local maximizer we have

$$\|x(t) - \overline{x}\|_2 \leq L\|t - \bar{t}\|_2. \tag{48}$$

Especially, taking $L = w_2/w_1$ fulfills the requirement, with

$$w_2 = \max_{z \in \mathrm{cl}(B_\delta(\overline{x}))} \left[\left\|\nabla_{xt}^2 f(z, \bar{t})\right\|_F + 1\right]. \tag{49}$$

Proof of Proposition A.5. We first prove the differentiability of $D(\boldsymbol{\lambda})$. For a fixed $\boldsymbol{\lambda}$, solving for $\pi_{\boldsymbol{\lambda}} = \arg\max_{\pi\in\Pi} L(\pi,\boldsymbol{\lambda})$ is equivalent to solving an unconstrained MDP with entropy regularization (cf. Section 2):

$$\max_{\pi\in\Pi} V_{\boldsymbol{\lambda}}^{\pi}(\rho). \tag{50}$$

As shown in (Nachum et al., 2017), $\pi_{\boldsymbol{\lambda}}$ can be uniquely characterized as

$$\pi_{\boldsymbol{\lambda}}(a|s) \propto \exp\left(\frac{Q_{\boldsymbol{\lambda}}^{\pi_{\boldsymbol{\lambda}}}(s,a) - V_{\boldsymbol{\lambda}}^{\pi_{\boldsymbol{\lambda}}}(s)}{\tau}\right),\ \forall (s,a)\in\mathcal{S}\times\mathcal{A}. \tag{51}$$

Therefore, a standard result in the duality theory (Floudas, 1995) implies that $D(\boldsymbol{\lambda})$ is differentiable with the gradient

$$\nabla D(\boldsymbol{\lambda}) = U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - \mathbf{b} = \left(U_{g_1}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - b_1, \ldots, U_{g_n}^{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) - b_n\right). \tag{52}$$

Next, we show that $\nabla D(\boldsymbol{\lambda})$ is Lipschitz continuous on Λ, which implies smoothness. Consider the statements:

1. $(U_{g_i}^{\pi}(\rho) - b_i)$ is ℓ_c-Lipschitz continuous with respect to π, which is already proved in Lemmma 2.1.
2. $\pi_{\boldsymbol{\lambda}}$ is ℓ_Λ-Lipschitz continuous with respect to $\boldsymbol{\lambda}$ for some $\ell_\Lambda > 0$.

If these statements hold true, it follows that

$$\left|(U_{g_i}^{\pi_{\boldsymbol{\lambda}_1}}(\rho) - b_i) - (U_{g_i}^{\pi_{\boldsymbol{\lambda}_2}}(\rho) - b_i)\right| \le \ell_c \|\pi_{\boldsymbol{\lambda}_1} - \pi_{\boldsymbol{\lambda}_2}\|_2 \le \ell_c\ell_\Lambda \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2, \tag{53}$$

which leads to

$$\|\nabla D(\boldsymbol{\lambda}_1) - \nabla D(\boldsymbol{\lambda}_2)\|_2 \le \sqrt{n}\ell_c\ell_\Lambda \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2, \tag{54}$$

i.e. $D(\boldsymbol{\lambda})$ is $\sqrt{n}\ell_c\ell_\Lambda$-strongly smooth on Λ.

To prove Statement 2, consider the Lagrangian $L(\pi,\boldsymbol{\lambda})$, which is twice continuously differentiable on $(0,1)^{|\mathcal{S}|\times|\mathcal{A}|}\times\Lambda$. The hidden constraint for the maximization problem (12b) is linear and has the form

$$\sum_{a\in\mathcal{A}} \pi(a|s) = 1,\ \forall s\in\mathcal{S}. \tag{55}$$

By Proposition 3.5, it holds that

$$\nabla_{\pi\pi} L(\pi_{\boldsymbol{\lambda}},\boldsymbol{\lambda}) \ge \frac{\tau d}{2(1-\gamma)\ln 2} I_{|\mathcal{S}||\mathcal{A}|}, \tag{56}$$

which implies that $\pi_{\boldsymbol{\lambda}}$ is a strict global maximizer of order 2 under Assumption 3.4.

Consider $\nabla^2_{\pi\boldsymbol{\lambda}} L(\pi,\boldsymbol{\lambda})$, which is a matrix of dimension $|\mathcal{S}||\mathcal{A}|\times n$. Specifically, it holds

$$\frac{\partial^2 L(\pi,\boldsymbol{\lambda})}{\partial\pi(a|s)\partial\lambda_i} \overset{(i)}{=} \frac{\partial}{\partial\pi(a|s)}\left(U_{g_i}^{\pi}(\rho) - b_i\right) \overset{(ii)}{=} \frac{1}{1-\gamma} d_\rho^\pi(s) Q_{g_i}^\pi(s,a), \tag{57}$$

where (i) follows from definition (12a) and and (ii) is due to the policy gradient (cf. Lemma D.1).

Following the same argument as in the proof of Proposition A.1, we have

$$\left\|\nabla^2_{\pi\boldsymbol{\lambda}} L(\pi,\boldsymbol{\lambda})\right\|_F \le \frac{\sqrt{n|\mathcal{A}|}}{(1-\gamma)^2}. \tag{58}$$

Therefore, applying Proposition A.6 with

$$w_1 = \frac{\tau d}{2(1-\gamma)\ln 2},\ w_2 = \frac{\sqrt{n|\mathcal{A}|}}{(1-\gamma)^2} + 1, \tag{59}$$

we conclude that $\pi_{\boldsymbol{\lambda}}$ is locally ℓ_Λ-Lipschitz continuous with respect to $\boldsymbol{\lambda}$ for all $\boldsymbol{\lambda}\in\Lambda$, where $\ell_\Lambda = w_2/w_1$. Since ℓ_Λ is universal and does not depend on $\boldsymbol{\lambda}$, the local Lipschitz property is ready to be extended to Λ. The proof is completed by setting $\ell = \sqrt{n}\ell_c\ell_\Lambda$. □

Proposition A.7 (Restatement of Proposition 3.7) *Suppose that Assumptions 3.1 and 3.4 hold. If $\boldsymbol{\lambda} \geq 0$ is an ε-optimal multiplier, i.e. $D(\boldsymbol{\lambda}) - D_\tau^\star \leq \varepsilon$, then the associated Lagrangian maximizer $\pi_{\boldsymbol{\lambda}}$ satisfies*

$$\|\pi_{\boldsymbol{\lambda}} - \pi_\tau^\star\|_2 \leq C_1\sqrt{\varepsilon}, \tag{60a}$$

$$|V_\tau^{\pi_{\boldsymbol{\lambda}}}(\rho) - V_\tau^\star| \leq 2\varepsilon + \ell_c C_1 C_2\sqrt{\varepsilon}, \tag{60b}$$

$$\max_{i\in[n]}\left[b_i - U_{g_i}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right]_+ \leq \ell_c C_1\sqrt{\varepsilon}, \tag{60c}$$

where

$$\ell_c = \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2},\ C_1 = \sqrt{\frac{2(1-\gamma)\ln 2}{\tau d}},\ C_2 = \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n \frac{1}{\xi_i}\right). \tag{61}$$

Proof. We can write $D(\boldsymbol{\lambda}) = L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda})$ and $D_\tau^\star = L(\pi_\tau^\star, \boldsymbol{\lambda}^\star)$, where $(\pi_\tau^\star, \boldsymbol{\lambda}^\star)$ is any primal-dual pair. Then, by the strong duality (Lemma 3.2), we have

$$L(\pi_\tau^\star, \boldsymbol{\lambda}^\star) = \min_{\boldsymbol{\mu}\geq 0} L(\pi_\tau^\star, \boldsymbol{\mu}) \leq L(\pi_\tau^\star, \boldsymbol{\lambda}). \tag{62}$$

Therefore,

$$\begin{aligned}\varepsilon \geq L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - L(\pi_\tau^\star, \boldsymbol{\lambda}^\star) &= L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\mu}\geq 0} L(\pi_\tau^\star, \boldsymbol{\mu}) \\ &\geq L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - L(\pi_\tau^\star, \boldsymbol{\lambda}) \\ &\overset{(i)}{\geq} \frac{\tau d}{2(1-\gamma)\ln 2}\|\pi_{\boldsymbol{\lambda}} - \pi_\tau^\star\|_2^2,\end{aligned} \tag{63}$$

where (i) results from the quadratic lower bound given by Proposition 3.5. Then, (60a) is obtained after rearranging the terms in (63).

Next, we can use the Lipschitz continuity of the utility function (cf. Lemma 2.1) to bound the constraint violation. For every $i = 1, 2, \ldots, n$, it holds that

$$\left|U_{g_i}^{\pi_{\boldsymbol{\lambda}}}(\rho) - U_{g_i}^{\pi_\tau^\star}(\rho)\right| \leq \ell_c\|\pi_{\boldsymbol{\lambda}} - \pi_\tau^\star\|_2 \leq \ell_c C_1\sqrt{\varepsilon}. \tag{64}$$

As the optimal policy $\pi_\tau^\star$ must be feasible to (11), i.e. $U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) \geq \mathbf{b}$, we can bound the constraint violation as

$$\max_{i\in[n]}\left[b_i - U_{g_i}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right]_+ \leq \max_{i\in[n]}\left\{\left[b_i - U_{g_i}^{\pi_\tau^\star}(\rho)\right]_+ + \left|U_{g_i}^{\pi_{\boldsymbol{\lambda}}}(\rho) - U_{g_i}^{\pi_\tau^\star}(\rho)\right|\right\} \leq \ell_c C_1\sqrt{\varepsilon}. \tag{65}$$

Finally, to bound the primal optimality gap, we note that

$$0 \overset{(i)}{\leq} L(\pi_\tau^\star, \boldsymbol{\lambda}) - L(\pi_\tau^\star, \boldsymbol{\lambda}^\star) \overset{(ii)}{\leq} L(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - L(\pi_\tau^\star, \boldsymbol{\lambda}^\star) = D(\boldsymbol{\lambda}) - D(\boldsymbol{\lambda}^\star) \leq \varepsilon, \tag{66}$$

where (i) follows from the strong duality and (ii) is due to the definition of $\pi_{\boldsymbol{\lambda}}$ (cf. (14)). Thus, by expanding the Lagrangian as

$$\begin{aligned}L(\pi_\tau^\star, \boldsymbol{\lambda}) - L(\pi_\tau^\star, \boldsymbol{\lambda}^\star) &= V_\tau^{\pi_\tau^\star}(\rho) + \boldsymbol{\lambda}^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - \mathbf{b}\right) - V_\tau^{\pi_\tau^\star}(\rho) - (\boldsymbol{\lambda}^\star)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - \mathbf{b}\right) \\ &= (\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - \mathbf{b}\right),\end{aligned} \tag{67}$$

and applying the complementary slackness $(\boldsymbol{\lambda}^\star)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right) = 0$, we obtain the bound

$$0 \leq (\boldsymbol{\lambda})^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right) \leq \varepsilon. \tag{68}$$

Therefore,

$$\begin{aligned}\left|(\boldsymbol{\lambda})^T\left(U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho) - b\right)\right| &\overset{(i)}{\leq} \left|(\boldsymbol{\lambda})^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right)\right| + \left|(\boldsymbol{\lambda})^T\left(U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho) - U_{\mathbf{g}}^{\pi_\tau^\star}(\rho)\right)\right| \\ &\overset{(ii)}{\leq} \varepsilon + \ell_c C_1\left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n \frac{1}{\xi_i}\right)\sqrt{\varepsilon},\end{aligned} \tag{69}$$

where (i) is due to the triangular inequality and (ii) uses the bound (64) and the boundedness of Λ (cf. Lemma 3.3), i.e. $0 \le \lambda_i \le \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)/\xi_i$ for all $i \in [n]$ and $\boldsymbol{\lambda} \in \Lambda$. Thus, we can bound the primal optimality gap as

$$\begin{aligned}
|V_\tau^{\pi_{\boldsymbol{\lambda}}}(\rho) - V_\tau^\star| &= \left|V_\tau^{\pi_{\boldsymbol{\lambda}}}(\rho) - V_\tau^{\pi_\tau^\star}(\rho)\right| \\
&\overset{(i)}{=} \left|\left[V_\tau^{\pi_{\boldsymbol{\lambda}}}(\rho) + (\boldsymbol{\lambda})^T\left(U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho) - b\right)\right] - (\boldsymbol{\lambda})^T\left(U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho) - b\right)\right. \\
&\qquad \left. - \left[V_\tau^{\pi_\tau^\star}(\rho) + (\boldsymbol{\lambda}^\star)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right)\right]\right| \\
&\overset{(ii)}{\le} \left|L\left(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right)\right| + \left|(\boldsymbol{\lambda})^T\left(U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho) - b\right)\right| \\
&\overset{(iii)}{\le} \varepsilon + \left(\varepsilon + \ell_c C_1\left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n \frac{1}{\xi_i}\right)\sqrt{\varepsilon}\right) \\
&= 2\varepsilon + \ell_c C_1\left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n \frac{1}{\xi_i}\right)\sqrt{\varepsilon} \\
&= 2\varepsilon + \ell_c C_1 C_2\sqrt{\varepsilon},
\end{aligned} \tag{70}$$

where (i) uses the complementary slackness $(\boldsymbol{\lambda}^\star)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right) = 0$ and (ii) uses the triangular inequality and the definition of Lagrangian (12a). In (iii), we use the assumption

$$D(\boldsymbol{\lambda}) - D\left(\boldsymbol{\lambda}^\star\right) = L\left(\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right) \le \varepsilon, \tag{71}$$

and the inequality (69). This completes the proof. □

B Supplementary Materials for Sections 4 and 5

B.1 Entropy-regularized NPG

For entropy-regularized MDPs, the natural policy gradient update rule can be written as

$$\theta \leftarrow \theta + \eta\left(\mathcal{F}_\rho^\theta\right)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho), \tag{72}$$

where $\mathcal{F}_\rho^\theta$ is the Fisher information matrix, defined as

$$\mathcal{F}_\rho^\theta := \mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\left[\left(\nabla_\theta \log \pi_\theta(a|s)\right)\left(\nabla_\theta \log \pi_\theta(a|s)\right)^\top\right]. \tag{73}$$

Under the soft-max parameterization, the associated policy update has a fairly direct form (cf. (24) and (25)). We refer the reader to (Cen et al., 2021) for a detailed derivation.

Cen et al. (2021) proved that the entropy-regularized NPG method enjoys a global linear convergence and a local quadratic convergence. We summarize the two results in Propositions B.1 and B.2, where we abuse the notations and denote the optimal unconstrained value function with entropy regularization, the corresponding Q-function, and the associated optimal policy with $V_\tau^\star$, $Q_\tau^\star$, and $\pi_\tau^\star$ respectively. Let $\mu_\tau^\star$ denote the stationary distribution over $\mathcal{S}$ of the MDP under policy $\pi_\tau^\star$[3].

Proposition B.1 (Global linear convergence) *If the step-size $\eta = (1-\gamma)/\tau$ is used, the entropy-regularized NPG algorithm (24) satisfies the error bounds:*

$$\left\|V_\tau^\star - V_\tau^{\pi^{(t+1)}}\right\|_\infty \le 3\left\|Q_\tau^\star - Q_\tau^{\pi^{(0)}}\right\|_\infty \gamma^{t+1}, \tag{74a}$$

$$\left\|\log \pi_\tau^\star - \log \pi^{(t+1)}\right\|_\infty \le 2\left\|Q_\tau^\star - Q_\tau^{\pi^{(0)}}\right\|_\infty \tau^{-1}\gamma^t, \tag{74b}$$

[3]It is straightforward to verify that $d_{\mu_\tau^\star}^{\pi_\tau^\star} = \mu_\tau^\star$.

for all $t > 0$, where

$$\left\|V_\tau^\star - V_\tau^{\pi^{(t+1)}}\right\|_\infty := \max_{s\in\mathcal{S}}\left|V_\tau^\star(s) - V_\tau^{\pi^{(t+1)}}(s)\right|, \tag{75a}$$

$$\left\|Q_\tau^\star - Q_\tau^{\pi^{(0)}}\right\|_\infty := \max_{s\in\mathcal{S},a\in\mathcal{A}}\left|Q_\tau^\star(s,a) - Q_\tau^{\pi^{(0)}}(s,a)\right|, \tag{75b}$$

$$\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty := \max_{s\in\mathcal{S},a\in\mathcal{A}}\left|\log\pi_\tau^\star(a|s) - \log\pi^{(t+1)}(a|s)\right|. \tag{75c}$$

We note that Cen et al. (2021) proved a more general result for all step-sizes $\eta \in [0, (1-\gamma)/\tau]$, whereas the fastest convergence is achieved with the maximum step-size $\eta = (1-\gamma)/\tau$.

Proposition B.2 (Local quadratic convergence) *Suppose that the entropy-regularized NPG algorithm (24) with the step-size $\eta = (1-\gamma)/\tau$ satisfies*

$$\left\|\log\pi^{(t)} - \log\pi_\tau^\star\right\|_\infty \le 1, \tag{76}$$

for all $t \ge 0$. There exist problem-dependent constants K_1 and K_2 such that

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le K_1\left(K_2\left(V_\tau^\star\left(\mu_\tau^\star\right) - V_\tau^{\pi^{(0)}}\left(\mu_\tau^\star\right)\right)\right)^{2^t}. \tag{77}$$

In our work, $V_{\boldsymbol{\lambda}}^\pi(\rho)$ is the entropy-regularized value function associated with the Lagrangian $L(\pi, \boldsymbol{\lambda})$, which has the reward function $r_{\boldsymbol{\lambda}}(s,a) = r(s,a) + \boldsymbol{\lambda}^T\mathbf{g}(s,a)$. Therefore, Proposition B.1 implies that, with step-size $\eta = (1-\gamma)/\tau$, the error bound $\left\|\log\pi_{\boldsymbol{\lambda}} - \log\pi^{(t)}\right\|_\infty \le \varepsilon$ can be achieved in

$$\frac{1}{1-\gamma}\log\left(\frac{2\left\|Q_{\boldsymbol{\lambda}}^{\pi_{\boldsymbol{\lambda}}} - Q_{\boldsymbol{\lambda}}^{\pi^{(0)}}\right\|_\infty}{\varepsilon\tau}\right), \tag{78}$$

iterations (cf. (26)). Furthermore, since $\boldsymbol{\lambda} \in \Lambda = \{\boldsymbol{\lambda} \mid 0 \le \lambda_i \le \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)/\xi_i$, for all $i \in [n]\}$, we have that $r_{\boldsymbol{\lambda}}(s,a) \in [0, 1 + C_2]$, where $C_2 = \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n 1/\xi_i\right)$. Together with the elementary entropy bound $\mathcal{H}(\rho,\pi) \in [0, \log|\mathcal{A}|/(1-\gamma)]$, it holds that

$$Q_{\boldsymbol{\lambda}}^\pi(s,a) \in \left[0, \frac{1 + C_2 + \tau\log|\mathcal{A}|}{1-\gamma}\right], \tag{79}$$

for all $\boldsymbol{\lambda} \in \Lambda$. Thus, we can drop the dependency of Q-function in (78) to obtain the following bound on the number of iterations:

$$\frac{1}{1-\gamma}\log\left(\frac{2\left(1 + C_2 + \tau\log|\mathcal{A}|\right)}{\varepsilon\tau(1-\gamma)}\right). \tag{80}$$

B.2 Accelerated Gradient Projection Method with Inexact Gradient

Gradient projection method is a feasible direction method for solving constrained optimization problems of the form:

$$\begin{aligned} &\min_x\ f(x) \\ &s.t.\ x \in X \end{aligned} \tag{81}$$

where $f(x)$ is convex and differentiable and X is convex. The general update scheme is

$$x^{(k+1)} = \mathcal{P}_X\left(x^{(k)} - \alpha^k\nabla f\left(x^{(k)}\right)\right). \tag{82}$$

When the gradient is inexact, Schmidt et al. (2011) proved the following bound for the general update (82) and the accelerated update (28).

Proposition B.3 (Convergence of inexact gradient projection method) *Assume that $f(x)$ is convex and L-smooth on X, and that we have access to a gradient oracle $h(x)$ such that $\|\nabla f(x) - h(x)\|_2 \le \delta$ for*

all $x \in X$. Let $x^\star = \arg\min_{x\in X} f(x)$. By selecting $\alpha_k = 1/L$ and $\beta_k = (k-1)/(k+2)$, then the iterates of algorithm (82) satisfy

$$f\left(\frac{1}{k}\sum_{i=1}^{k} x^{(i)}\right) - f(x^\star) \leqslant \frac{L}{2k}\left(\left\|x^{(0)} - x^\star\right\|_2 + \frac{2k\delta}{L}\right)^2 = O\left(\frac{1}{k}\right) + O\left(k^2\delta^2\right) + O(k\delta). \tag{83}$$

Moreover, for the accelerated version (28), it holds that

$$f\left(x^{(k)}\right) - f(x^\star) \leqslant \frac{2L}{(k+1)^2}\left(\left\|x^{(0)} - x^\star\right\|_2 + \frac{(k+1)k\delta}{L}\right)^2 = O\left(\frac{1}{k^2}\right) + O\left(k^2\delta^2\right) + O(\delta). \tag{84}$$

B.3 Proofs of Results in Section 5

Proposition B.4 (Restatement of Proposition 5.1) *Suppose that π is an approximate solution to (14) such that $\|\log\pi - \log\pi_{\boldsymbol{\lambda}}\|_\infty \leq \varepsilon$. The gradient estimator defined by*

$$\widetilde{\nabla} D(\boldsymbol{\lambda}) \coloneqq U_{\mathbf{g}}^{\pi}(\rho) - \mathbf{b} = \left(U_{g_1}^{\pi}(\rho) - b_1, \ldots, U_{g_n}^{\pi}(\rho) - b_n\right), \tag{85}$$

satisfies

$$\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2 \leq \frac{\sqrt{n}|\mathcal{A}|}{(1-\gamma)^2}\varepsilon. \tag{86}$$

Proof. Since $(\log x)' = 1/x \geq 1$ for all $x \in (0,1]$, it holds that $\|\pi - \pi_{\boldsymbol{\lambda}}\|_\infty \leq \|\log\pi - \log\pi_{\boldsymbol{\lambda}}\|_\infty \leq \varepsilon$. To bound the quantity $\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2 = \left\|U_{\mathbf{g}}^{\pi}(\rho) - U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right\|_2$, we can either use the Lipschitz continuity of $U_{\mathbf{g}}^{\pi}(\rho)$ (cf. Lemma 2.1) or use the performance difference lemma (cf. Lemma D.3).

With the Lipschitz continuity, we have

$$\left|U_{g_i}^{\pi}(\rho) - U_{g_i}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right| \leq \ell_c\|\pi - \pi_{\boldsymbol{\lambda}}\|_2 \leq \ell_c\sqrt{n}\|\pi - \pi_{\boldsymbol{\lambda}}\|_\infty = \frac{\sqrt{n|\mathcal{A}|}}{(1-\gamma)^2}\varepsilon, \tag{87}$$

where we have used the inequality $\|\cdot\|_2 \leq \sqrt{n}\|\cdot\|_\infty$. Therefore,

$$\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2 \leq \sqrt{n}\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_\infty = \sqrt{n}\left\|U_{\mathbf{g}}^{\pi}(\rho) - U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right\|_\infty \leq \frac{n\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\varepsilon. \tag{88}$$

On the other hand, we can use the performance difference lemma to obtain

$$\begin{aligned}
\left|U_{g_i}^{\pi}(\rho) - U_{g_i}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right| &= \left|\frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d_\rho^{\pi_{\boldsymbol{\lambda}}}(s)\sum_{a\in\mathcal{A}}\left(\pi(a|s) - \pi_{\boldsymbol{\lambda}}(a|s)\right)Q_{g_i}^{\pi}(s,a)\right| \\
&\leq \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d_\rho^{\pi_{\boldsymbol{\lambda}}}(s)\sum_{a\in\mathcal{A}}\left|\pi(a|s) - \pi_{\boldsymbol{\lambda}}(a|s)\right|Q_{g_i}^{\pi}(s,a) \\
&\overset{(i)}{\leq} \frac{\varepsilon}{1-\gamma}\sum_{s\in\mathcal{S}} d_\rho^{\pi_{\boldsymbol{\lambda}}}(s)\sum_{a\in\mathcal{A}} Q_{g_i}^{\pi}(s,a) \\
&\overset{(ii)}{\leq} \frac{|\mathcal{A}|}{(1-\gamma)^2}\varepsilon,
\end{aligned} \tag{89}$$

where (i) is based on the bound $\|\pi - \pi_{\boldsymbol{\lambda}}\|_\infty \leq \varepsilon$ and (ii) is due to $Q_{g_i}^{\pi}(s,a) \leq 1/(1-\gamma)$ and the fact that $d_\rho^{\pi_{\boldsymbol{\lambda}}}(\cdot)$ is a probability distribution. Repeating (88) with the bound (89) yields that

$$\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2 \leq \sqrt{n}\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_\infty = \sqrt{n}\left\|U_{\mathbf{g}}^{\pi}(\rho) - U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}}}(\rho)\right\|_\infty \leq \frac{\sqrt{n}|\mathcal{A}|}{(1-\gamma)^2}\varepsilon. \tag{90}$$

Equations (88) and (90) give two upper bounds on the quantity $\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2$. In this work, we use the bound (90), as it has a weaker dependence on the number of constraints n. This completes the proof. □

Proposition 5.1 implies that running Algorithm 2 with the step-size $\eta = (1-\gamma)/\tau$ for

$$\frac{1}{1-\gamma}\log\left(\frac{2\sqrt{n}\,|\mathcal{A}|\,(1 + C_2 + \tau\log|\mathcal{A}|)}{\delta(1-\gamma)^3\tau}\right). \tag{91}$$

iterations, where $C_2 = \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n 1/\xi_i\right)$, guarantees a δ-accurate gradient estimation $\widetilde{\nabla} D(\boldsymbol{\lambda})$, i.e. $\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2 \le \delta$ (cf. (30)).

Theorem B.5 (Restatement of Theorem 5.2) *Suppose that Assumptions 3.1 and 3.4 hold. For every $\varepsilon_1 > 0$, Algorithm 1 with a random initialization and the parameters $\eta = (1-\gamma)/\tau$, $\alpha_k = 1/\ell$, $\beta_k = (k-1)/(k+2)$, and*

$$N_1 = T,\ N_2 = \frac{1}{1-\gamma}\log\left(\frac{2\sqrt{n}\,|\mathcal{A}|\,T(T+1)\,(1+C_2+\tau\log|\mathcal{A}|)}{(1-\gamma)^3\tau\ell}\right),\ N_3 = \frac{1}{1-\gamma}\log\left(\frac{2\sqrt{n}\,(1+C_2+\tau\log|\mathcal{A}|)}{\varepsilon_1\tau(1-\gamma)}\right), \quad (92)$$

returns a solution pair $(\pi, \boldsymbol{\lambda})$ such that

$$D(\boldsymbol{\lambda}) - D_\tau^\star \le \varepsilon_0, \quad (93a)$$

$$\|\pi - \pi_\tau^\star\|_2 \le C_1\sqrt{\varepsilon_0} + \varepsilon_1, \quad (93b)$$

$$|V_\tau^\pi(\rho) - V_\tau^\star| \le 2\varepsilon_0 + \ell_c C_1 C_2\sqrt{\varepsilon_0} + \left(\ell_c C_2 + \frac{3\gamma}{2\tau\sqrt{n}}\right)\varepsilon_1, \quad (93c)$$

$$\max_{i\in[n]}\left[b_i - U_{g_i}^\pi(\rho)\right]_+ \le \ell_c\left(C_1\sqrt{\varepsilon} + \varepsilon_1\right), \quad (93d)$$

where

$$\varepsilon_0 = \frac{2\ell}{(T+1)^2}\left(\left\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}^\star\right\|_2 + 1\right)^2,\ \ell = \frac{2\ln 2\left(n|\mathcal{A}| + (1-\gamma)^2\sqrt{n|\mathcal{A}|}\right)}{\tau(1-\gamma)^3 d}, \quad (94)$$

and

$$\ell_c = \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2},\ C_1 = \sqrt{\frac{2(1-\gamma)\ln 2}{\tau d}},\ C_2 = \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n \frac{1}{\xi_i}\right). \quad (95)$$

The total iteration complexity is $N_1 \times N_2 + N_3 = \widetilde{\mathcal{O}}(T)$ with primal error bounds $\mathcal{O}(1/T)$ given by (93b)-(93d) and a dual error bound $\mathcal{O}\left(1/T^2\right)$ given by (93a).

Proof. Under Assumptions 3.1 and 3.4, it follows from Proposition 3.6 that $D(\boldsymbol{\lambda})$ is convex, differentiable, and ℓ-smooth on Λ. Now, we fix the gradient accuracy as

$$\delta = \frac{\ell}{T(T+1)}. \quad (96)$$

Then, it follows from Proposition 5.1 and (91) that running the NPG subroutine in line 4 of Algorithm 1 for

$$N_2 = \frac{1}{1-\gamma}\log\left(\frac{2\sqrt{n}\,|\mathcal{A}|\,T(T+1)\,(1+C_2+\tau\log|\mathcal{A}|)}{(1-\gamma)^3\tau\ell}\right), \quad (97)$$

iterations guarantees obtaining an estimation $\widetilde{\nabla} D(\boldsymbol{\lambda})$ such that

$$\left\|\widetilde{\nabla} D(\boldsymbol{\lambda}) - \nabla D(\boldsymbol{\lambda})\right\|_2 \le \frac{\ell}{T(T+1)}, \quad (98)$$

where we have use the bound $\left\|Q_{\boldsymbol{\lambda}}^{\pi_{\boldsymbol{\lambda}}} - Q_{\boldsymbol{\lambda}}^{(0)}\right\|_\infty \le (1 + C_2 + \log|\mathcal{A}|)/(1-\gamma)$ for all $\boldsymbol{\lambda} \in \Lambda$ (cf. (79)). Therefore, by Proposition B.3, running the outer loop in Algorithm 1 for $N_1 = T$ iterations generates a solution $\boldsymbol{\lambda}^{(T)}$ such that

$$D\left(\boldsymbol{\lambda}^{(T)}\right) - D_\tau^\star \le \frac{2\ell}{(T+1)^2}\left(\left\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}^\star\right\|_2 + \frac{(T+1)T\delta}{\ell}\right)^2 = \frac{2\ell}{(T+1)^2}\left(\left\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}^\star\right\|_2 + 1\right)^2, \quad (99)$$

which satisfies (93a).

Below, we adopt the proof of Proposition 3.7 (cf. Appendix A). We first apply Proposition 3.7 with

$$\varepsilon_0 = \frac{2\ell}{(T+1)^2}\left(\left\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}^\star\right\|_2 + 1\right)^2 \quad (100)$$

to obtain $\|\pi_{\boldsymbol{\lambda}^{(T)}} - \pi_\tau^\star\|_2 \le C_1\sqrt{\varepsilon_0}$, where $\pi_\tau^\star$ is an optimal policy. By Propositions B.1, we can compute an approximate Lagrangian maximizer $\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}$ to (14) such that $\|\log\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}} - \log\pi_{\boldsymbol{\lambda}^{(T)}}\|_\infty \le \varepsilon_1/\sqrt{n}$, by running Algorithm 2 for

$$N_3 = \frac{1}{1-\gamma}\log\left(\frac{2\sqrt{n}\,(1+C_2+\tau\log|\mathcal{A}|)}{\varepsilon_1\tau(1-\gamma)}\right), \tag{101}$$

iterations (cf. (78)). Now, we show that $\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}$ is a solution to (11) satisfying (93b)-(93c). Firstly, we have

$$\|\log\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}} - \log\pi_{\boldsymbol{\lambda}^{(T)}}\|_2 \le \sqrt{n}\,\|\log\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}} - \log\pi_{\boldsymbol{\lambda}^{(T)}}\|_\infty \le \varepsilon_1. \tag{102}$$

By applying the triangular inequality and using the strong concavity of the logarithm function on $(0,1]$, it holds that

$$\|\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}} - \pi_\tau^\star\|_2 \le \|\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}} - \pi_{\boldsymbol{\lambda}^{(T)}}\|_2 + \|\pi_{\boldsymbol{\lambda}^{(T)}} - \pi_{\boldsymbol{\lambda}^{(T)}}\|_2 \le C_1\sqrt{\varepsilon_0} + \varepsilon_1. \tag{103}$$

Then, we bound the constraint violation. It follows from the Lipschitz continuity of the utility function (cf. (3)) that

$$\left|U_{g_i}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - U_{g_i}^{\pi_\tau^\star}(\rho)\right| \le \ell_c\,\|\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}} - \pi_\tau^\star\|_2 \le \ell_c\left(C_1\sqrt{\varepsilon_0}+\varepsilon_1\right), \quad \forall i = 1,2,\ldots,n. \tag{104}$$

As the optimal policy $\pi_\tau^\star$ must be a feasible solution to (11), i.e. $U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) \ge \mathbf{b}$, the constraint violation is bounded as

$$\max_{i\in[n]}\left[b_i - U_{g_i}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho)\right]_+ \le \max_{i\in[n]}\left\{\left[b_i - U_{g_i}^{\pi_\tau^\star}(\rho)\right]_+ + \left|U_{g_i}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - U_{g_i}^{\pi_\tau^\star}(\rho)\right|\right\} \le \ell_c\left(C_1\sqrt{\varepsilon_0}+\varepsilon_1\right). \tag{105}$$

Finally, to bound the primal optimality gap, we note that

$$0 \overset{(i)}{\le} L\left(\pi_\tau^\star, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right) \overset{(ii)}{\le} L\left(\pi_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right) = D\left(\boldsymbol{\lambda}^{(T)}\right) - D_\tau^\star \le \varepsilon_0, \tag{106}$$

where (i) follows from the strong duality and (ii) is due to the definition of $\pi_{\lambda^{(T)}}$. Thus, by expanding the Lagrangian as

$$\begin{aligned} L\left(\pi_\tau^\star, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right) &= V_\tau^{\pi_\tau^\star}(\rho) + \left(\boldsymbol{\lambda}^{(T)}\right)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - \mathbf{b}\right) - V_\tau^{\pi_\tau^\star}(\rho) - \left(\boldsymbol{\lambda}^\star\right)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - \mathbf{b}\right) \\ &= \left(\boldsymbol{\lambda}^{(T)} - \boldsymbol{\lambda}^\star\right)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - \mathbf{b}\right), \end{aligned} \tag{107}$$

and applying the complementary slackness $\left(\boldsymbol{\lambda}^\star\right)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right) = 0$, we obtain the bound

$$0 \le \left(\boldsymbol{\lambda}^{(T)}\right)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right) \le \varepsilon_0. \tag{108}$$

Therefore,

$$\begin{aligned} \left|\left(\boldsymbol{\lambda}^{(T)}\right)^T\left(U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}^{(T)}}}(\rho) - \mathbf{b}\right)\right| &\overset{(i)}{\le} \left|\left(\boldsymbol{\lambda}^{(T)}\right)^T\left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right)\right| + \left|\left(\boldsymbol{\lambda}^{(T)}\right)^T\left(U_{\mathbf{g}}^{\pi_{\boldsymbol{\lambda}^{(T)}}}(\rho) - U_{\mathbf{g}}^{\pi_\tau^\star}(\rho)\right)\right| \\ &\overset{(ii)}{\le} \varepsilon_0 + \ell_c\left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)\left(\sum_{i=1}^n \frac{1}{\xi_i}\right)\left(C_1\sqrt{\varepsilon_0}+\varepsilon_1\right) \\ &= \varepsilon_0 + \ell_c C_2\left(C_1\sqrt{\varepsilon_0}+\varepsilon_1\right), \end{aligned} \tag{109}$$

where (i) is based on the triangular inequality, and (ii) uses the bound (104) and the boundedness of Λ, i.e. $0 \le \lambda_i \le \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)/\xi_i$ for all $i \in [n]$ and $\boldsymbol{\lambda} \in \Lambda$ (cf. Lemma 3.3). Thus, we can bound the primal optimality

gap as

$$\begin{aligned}
\left|V_\tau^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - V_\tau^\star\right| &= \left|V_\tau^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - V_\tau^{\pi_\tau^\star}(\rho)\right| \\
&\overset{(i)}{=} \left|\left[V_\tau^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) + \left(\boldsymbol{\lambda}^{(T)}\right)^T \left(U_{\mathbf{g}}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - \mathbf{b}\right)\right] - \left(\boldsymbol{\lambda}^{(T)}\right)^T \left(U_{\mathbf{g}}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - \mathbf{b}\right)\right. \\
&\quad \left. - \left[V_\tau^{\pi_\tau^\star}(\rho) + (\boldsymbol{\lambda}^\star)^T \left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right)\right]\right| \\
&\overset{(ii)}{\leq} \left|L\left(\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right)\right| + \left|\left(\boldsymbol{\lambda}^{(T)}\right)^T \left(U_{\mathbf{g}}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - \mathbf{b}\right)\right| \\
&\overset{(iii)}{\leq} \left|L\left(\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right)\right| + \left|L\left(\pi_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right)\right| \\
&\quad + \left|\left(\boldsymbol{\lambda}^{(T)}\right)^T \left(U_{\mathbf{g}}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - \mathbf{b}\right)\right| \\
&\overset{(iv)}{\leq} \left(\frac{3\gamma}{2\tau\sqrt{n}}\varepsilon_1\right) + (\varepsilon_0) + \left(\varepsilon_0 + \ell_c C_2 \left(C_1\sqrt{\varepsilon_0} + \varepsilon_1\right)\right) \\
&= 2\varepsilon_0 + \ell_c C_1 C_2 \sqrt{\varepsilon_0} + \left(\ell_c C_2 + \frac{3\gamma}{2\tau\sqrt{n}}\right)\varepsilon_1,
\end{aligned} \tag{110}$$

where (i) uses the complementary slackness $(\boldsymbol{\lambda}^\star)^T \left(U_{\mathbf{g}}^{\pi_\tau^\star}(\rho) - b\right) = 0$, (ii) uses the triangular inequality and the definition of Lagrangian (12a), and (iii) uses the triangular inequality again. The inequality (iv) contains three parts where the first part uses Proposition B.1 as

$$\left|L\left(\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right)\right| = \left|V_{\boldsymbol{\lambda}^{(T)}}^{\widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}}}(\rho) - V_{\boldsymbol{\lambda}^{(T)}}^{\pi_{\boldsymbol{\lambda}^{(T)}}}(\rho)\right| \leq \frac{3\gamma}{2\tau} \left\|\log \widetilde{\pi}_{\boldsymbol{\lambda}^{(T)}} - \log \pi_{\boldsymbol{\lambda}^{(T)}}\right\|_\infty = \frac{3\gamma}{2\tau\sqrt{n}}\varepsilon_1, \tag{111}$$

the second part uses the assumption

$$D(\boldsymbol{\lambda}^{(T)}) - D_\tau^\star = L\left(\pi_{\boldsymbol{\lambda}^{(T)}}, \boldsymbol{\lambda}^{(T)}\right) - L\left(\pi_\tau^\star, \boldsymbol{\lambda}^\star\right) \leq \varepsilon_0, \tag{112}$$

and the third part uses the inequality (109). This completes the proof. □

Corollary B.6 (Restatement of Corollary 5.4) *Suppose that Assumptions 3.1 and 3.4 hold. Let*

$$\tau = \frac{(1-\gamma)\varepsilon}{4\log|\mathcal{A}|}. \tag{113}$$

Then, Algorithm 1 computes a solution π for the standard CMDP such that

$$\left|V^{\pi^\star}(\rho) - V^\pi(\rho)\right| = \mathcal{O}(\varepsilon), \tag{114a}$$

$$\max_{i\in[n]} \left[b_i - U_{g_i}^\pi(\rho)\right]_+ = \mathcal{O}(\varepsilon), \tag{114b}$$

in $\widetilde{\mathcal{O}}\left(1/\varepsilon^2\right)$ iterations, where $\pi^\star$ is an optimal policy to the standard CMDP.

Proof. Since the total iteration complexity of Algorithm 1 is dominated by $N_1 \times N_2$ and the error bounds (31b)-(31d) are dominated by $\sqrt{\varepsilon_0}$ (cf. (94)), we ignore the effect of ε_0, ε_1 and only focus on $\sqrt{\varepsilon_0}$ terms in (31b)-(31d) through the analysis below.

Firstly, by invoking the optimality of $\pi_\tau^\star$ with respect to the entropy-regularized CMDP and the elementary entropy bound $0 \leq \mathcal{H}(\rho, \pi) \leq \log|\mathcal{A}|/(1-\gamma)$, we obtain

$$V^{\pi_\tau^\star}(\rho) + \frac{\tau}{1-\gamma}\log|\mathcal{A}| \geq V^{\pi_\tau^\star}(\rho) + \tau\mathcal{H}\left(\rho, \pi_\tau^\star\right) = V_\tau^\star(\rho) \geq V_\tau^{\pi^\star}(\rho) \geq V^{\pi^\star}(\rho), \tag{115}$$

which implies the sandwich bound (34). Now, we choose T in such a way that $\ell_c C_1 C_2 \sqrt{\varepsilon_0} = \varepsilon/2$, where ℓ_c, C_1, and C_2 are specified in Theorem B.5. Then, Theorem B.5 implies that running Algorithm 1 with

$$N_1 = T,\ N_2 = \frac{1}{1-\gamma}\log\left(\frac{2\sqrt{n}\,|\mathcal{A}|\,T(T+1)\left(1 + C_2 + \tau\log|\mathcal{A}|\right)}{(1-\gamma)^3\tau\ell}\right) \tag{116}$$

returns a solution π such that $|V_\tau^\star - V_\tau^\pi(\rho)| = \mathcal{O}(\varepsilon/2)$ (cf. (93c)). It then follows that

$$\left|V^{\pi^\star}(\rho) - V^\pi(\rho)\right| \overset{(i)}{\leq} |V^\star(\rho) - V_\tau^\star(\rho)| + |V_\tau^\star(\rho) - V_\tau^\pi(\rho)| + |V_\tau^\pi(\rho) - V^\pi(\rho)| \overset{(ii)}{\leq} \frac{2\tau \log|\mathcal{A}|}{1-\gamma} + \mathcal{O}\left(\frac{\varepsilon}{2}\right) = \mathcal{O}(\varepsilon), \quad (117)$$

where (i) is due to the triangular inequality and (ii) uses the bound (115) (cf. (34)). The $\mathcal{O}(\varepsilon)$-constraint violation follows directly from Theorem 5.2, since it enjoys the same order of convergence as the primal optimality gap.

We note that $\ell_c C_1 C_2 \sqrt{\varepsilon_0}$ can be written as $C/(\tau(T+1))$, where

$$C = \ell_c (C_1 \cdot \sqrt{\tau}) C_2 \left(\sqrt{2\ell \cdot \tau}\right) \left(\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}^\star\|_2 + 1\right) \quad (118)$$

is a constant that does not depend on T and ε. Thus, the choice $\tau = [(1-\gamma)\varepsilon] / (4\log|\mathcal{A}|)$ implies that

$$\frac{C}{T+1} = \frac{(1-\gamma)\varepsilon}{4\log|\mathcal{A}|} \times \frac{\varepsilon}{2}, \quad (119)$$

i.e. $T = \mathcal{O}(1/\varepsilon^2)$. Since the total iteration complexity is $N_1 \times N_2 = T \times \mathcal{O}(\log T)$, we obtain the $\widetilde{\mathcal{O}}\left(1/\sqrt{T}\right)$ error bound for the primal optimality gap and the constraint violation. This completes the proof. □

C Supplementary Materals for Section 6

In this subsection, we consider the special situation where there is a single constraint ($n = 1$). In particular, we use the non-bold notations to emphasize that the associated notations denote numbers instead of vectors used in the previous sections, e.g. multiplier λ, constraint $U_g^\pi(\rho) \geq b$, Slater condition $V_\tau^{\overline{\pi}}(\rho) - b \geq \xi$.

Since the feasible region $\Lambda = [0, C_2]$, where $C_2 = \left(V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)\right)/\xi$, is bounded and $D(\lambda)$ is convex, an approximate stationary point is also an approximate optimal solution. Specifically, if $|\nabla D(\lambda)| < \varepsilon$, then

$$D(\lambda) - D_\tau^\star \leq |\nabla D(\lambda)| \times |\lambda - \lambda^\star| < C_2 \varepsilon. \quad (120)$$

Additionally, if $\operatorname{sign}(\nabla D(0)) = \operatorname{sign}(\nabla D(C_2)) = 1$, then $D(\lambda)$ attains the optimum at $\lambda = 0$ due to the convexity. Similarly, $D(\lambda)$ attains the optimum at $\lambda = C_2$ if $\operatorname{sign}(\nabla D(0)) = \operatorname{sign}(\nabla D(C_2)) = -1$. Therefore, it only remains to consider the case where $D(0) < 0$ and $D(C_2) > 0$.

The proposed method aims to find an approximate stationary point with the bisection scheme. For a given search interval, it computes the gradient at the middle point. If the gradient is greater than ε, it shrinks the search interval to the left-half interval; if the gradient is smaller than $-\varepsilon$, it shrinks the search interval to the right-half interval. The iterates terminate when it finds a point λ such that $\left|\widetilde{\nabla} D(\lambda)\right| < \varepsilon$, where $\widetilde{\nabla} D(\lambda)$ is the approximate gradient. We summarize the proposed method in Algorithm 3, where we separately define the gradient estimator (cf. lines 4 and 5 in Algorithm 1) as a new subroutine Grad_{Sub} (cf. Algorithm 4) for the ease of presentation. We also assume that the initialization is non-trivial in the sense that $\nabla D(0) < 0$ and $\nabla D(C_2) > 0$, since otherwise we can return the optimal solution $\lambda^\star$ as 0 or C_2.

Below, we give a formal statement for the convergence result of Algorithm 3 (cf. Theorem 6.1).

Theorem C.1 (Restatement of Theorem 6.1) *Suppose that Assumptions 3.1 and 3.4 hold. When $n = 1$, for every ε, $\varepsilon_1 > 0$, Algorithm 3 with the parameters*

$$\eta = \frac{(1-\gamma)}{\tau}, \ N_1 = \frac{1}{1-\gamma} \log\left(\frac{4|\mathcal{A}|\,(1 + C_2 + \tau\log|\mathcal{A}|)}{(1-\gamma)^3 \tau \varepsilon}\right), N_2 = \frac{1}{1-\gamma} \log\left(\frac{2\,(1 + C_2 + \tau\log|\mathcal{A}|)}{\varepsilon_1 \tau (1-\gamma)}\right), \quad (121)$$

returns a solution (π, λ) in at most $\log_2(\ell C_2/\varepsilon)$ outer loops, such that

$$D(\lambda) - D_\tau^\star \leq \varepsilon_0, \quad (122a)$$

$$\|\pi - \pi_\tau^\star\|_2 \leq C_1\sqrt{\varepsilon_0} + \varepsilon_1, \quad (122b)$$

$$|V_\tau^\pi(\rho) - V_\tau^\star| \leq 2\varepsilon_0 + \ell_c C_1 C_2 \sqrt{\varepsilon_0} + \left(\ell_c C_2 + \frac{3\gamma}{2\tau}\right)\varepsilon_1, \quad (122c)$$

$$\left[b - U_g^\pi(\rho)\right]_+ \leq \ell_c \left(C_1\sqrt{\varepsilon} + \varepsilon_1\right), \quad (122d)$$

Algorithm 3 Bisection Method with NPG Subroutine

1: **Input:** Initialization π, $p_0 = 0$, $q_0 = C_2$; step-size η; maximum number of iterations N_1, N_2; threshold ε.
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Let $\lambda = (p_t + q_t)/2$.
4: **if** $|\mathrm{Grad}_{Sub}(\lambda, \pi, \eta, N_1)| < \varepsilon$ **then**
5: **break**
6: **else**
7: **if** $\mathrm{Grad}_{Sub}(\lambda, \pi, \eta, N_1) \geq \varepsilon$ **then**
8: Let $p_{t+1} \leftarrow p_t$ and $q_{t+1} \leftarrow \lambda$.
9: **else**
10: Let $p_{t+1} \leftarrow \lambda$ and $q_{t+1} \leftarrow q_t$.
11: **end if**
12: **end if**
13: **end for**
14: Recover the policy from the dual variable: $\widetilde{\pi}_\lambda \leftarrow \mathrm{NPG}_{Sub}(\lambda, \pi, \eta, N_2)$.

Algorithm 4 Gradient Estimator (Grad_{Sub})

1: **Input:** Target point λ, initialization π, step-size η, maximum number of iterations N.
2: Estimate the optimal policy π_λ for problem (14) through the natural policy gradient subroutine: $\widetilde{\pi}_\lambda \leftarrow \mathrm{NPG}_{Sub}(\lambda, \pi, \eta, N)$.
3: Compute and output the approximate gradient at λ: $\widetilde{\nabla} D(\lambda) \coloneqq U_g^{\widetilde{\pi}_\lambda}(\rho) - b$.

where

$$\varepsilon_0 = \frac{3C_2}{2}\varepsilon,\ \ell = \frac{2\ln 2\left(|\mathcal{A}| + (1-\gamma)^2\sqrt{|\mathcal{A}|}\right)}{\tau(1-\gamma)^3 d},\ \ell_c = \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2},\ C_1 = \sqrt{\frac{2(1-\gamma)\ln 2}{\tau d}},\ C_2 = \frac{V_\tau^\star - V_\tau^{\overline{\pi}}(\rho)}{\xi}. \tag{123}$$

The total iteration complexity is $\log_2(\ell C_2/\varepsilon) \times N_2 + N_3 = \mathcal{O}\left(\log^2(1/\varepsilon) + \log(1/\varepsilon_1)\right)$ *with primal error bounds* $\mathcal{O}\left(\sqrt{\varepsilon} + \varepsilon_1\right)$ *given by (122b) - (122d) and a dual error bound* $\mathcal{O}(\varepsilon)$ *given by (122a).*

We remark that the constants ℓ, ℓ_c, C_1, and C_2 used in Theorem C.1 coincide with those used in previous sections, except that they correspond to the 1-dimensional situation. The maximum number of iterations N_1 and N_2 in Theorem C.1, respectively, correspond to the N_2 and N_3 in Theorem B.5.

Proof. Under the assumption that $\nabla D(0) < 0$ and $\nabla D(\xi) > 0$, the optimal dual variable $\lambda^\star$ must belong to the interval $(0, C_2)$ and $\nabla D(\lambda^\star) = 0$. Denote $\widetilde{\nabla} D(\lambda) = \mathrm{Grad}_{Sub}(\lambda, \pi, \eta, N_1)$, i.e. the output of the gradient estimator. For a given threshold ε, when

$$N_1 = \frac{1}{1-\gamma}\log\left(\frac{4|\mathcal{A}|\left(1 + C_2 + \tau\log|\mathcal{A}|\right)}{(1-\gamma)^3\tau\varepsilon}\right), \tag{124}$$

it follows from the proof of Theorem 5.2 (cf. (98)) that

$$|\widetilde{\nabla} D(\lambda) - \nabla D(\lambda)| \leq \frac{\varepsilon}{2}. \tag{125}$$

Thus, if $\widetilde{\nabla} D(\lambda) \geq \varepsilon$, we have that $\nabla D(\lambda) \geq \varepsilon/2$. Similarly, if $\widetilde{\nabla} D(\lambda) \leq -\varepsilon$, we have that $\nabla D(\lambda) \leq -\varepsilon/2$. Therefore, the lines 7-11 in Algorithm 3 shrink by a factor of 2 the search region that contains the optimal solution $\lambda^\star$.

By leveraging the triangular inequality, we have that

$$|\widetilde{\nabla} D(\lambda)| \leq |\widetilde{\nabla} D(\lambda) - \nabla D(\lambda)| + |\nabla D(\lambda) - \nabla D(\lambda^\star)| \leq \frac{\varepsilon}{2} + \ell|\lambda - \lambda^\star|. \tag{126}$$

where we apply the smoothness of the dual function (cf. Proposition 3.6). Thus, Algorithm 3 terminates in at most $t = \log_2(\ell C_2/\varepsilon)$ iterations with an ε-optimal stationary point due to

$$|\widetilde{\nabla} D(\lambda)| \leq \frac{\varepsilon}{2} + \ell|\lambda - \lambda^\star| \leq \frac{\varepsilon}{2} + \ell\left(\frac{1}{2}\right)^t C_2 \leq \varepsilon, \tag{127}$$

where λ denotes the midpoint $(p_t + q_t)/2$ generated in the t-th iteration in line 3.

Now, suppose that λ is the output solution with $|\widetilde{\nabla} D(\lambda)| \le \varepsilon$. It holds that

$$D(\lambda) - D_\tau^\star \le |\nabla D(\lambda)| \times |\lambda - \lambda^\star| \le \left(|\widetilde{\nabla} D(\lambda)| + \frac{\varepsilon}{2}\right) C_2 = \frac{3C_2}{2}\varepsilon, \tag{128}$$

where we have used (120) and (127). By substituting (128) into the proof of Theorem 5.2 and letting $\varepsilon_0 = 3C_2\varepsilon/2$, the desired bounds follow. The convergence is linear and the total iteration complexity is upper-bounded by $\log_2(\ell C_2/\varepsilon) \times N_2 + N_3 = \mathcal{O}(\log^2(1/\varepsilon) + \log(1/\varepsilon_1))$. □

Corollary C.2 (Restatement of Corollary 6.2) *Suppose that Assumptions 3.1 and 3.4 hold. Let*

$$\tau = \frac{(1-\gamma)\varepsilon}{4\log|\mathcal{A}|}. \tag{129}$$

Then, Algorithm 3 computes a solution π for the standard CMDP such that

$$\left|V^{\pi^\star}(\rho) - V^{\pi}(\rho)\right| = \mathcal{O}(\varepsilon), \tag{130a}$$

$$\left[b - U_g^\pi(\rho)\right]_+ = \mathcal{O}(\varepsilon), \tag{130b}$$

in $\mathcal{O}\left(\log^2(1/\varepsilon)\right)$ iterations, where $\pi^\star$ is an optimal policy to the standard CMDP.

Proof. The proof can be fully adopted from that of Corollary B.6 The main difference lies in the outer-loop complexity. To have $\ell_c C_1 C_2 \sqrt{\varepsilon_0} = \varepsilon/2$, compared to the previous $\widetilde{\mathcal{O}}\left(1/\varepsilon^2\right)$ total iterations, it only requires $\mathcal{O}\left(\log^2(1/\varepsilon)\right)$ total iterations for Algorithm 3 when there is a single constraint. □

D Supporting Lemmas

The followings are standard results about unregularized and entropy-regularized MDPs. We refer the reader to (Agarwal et al., 2021; Mei et al., 2020) for the proofs.

Lemma D.1 (Policy gradient for direct parameterization) *Suppose that $V^\pi(\rho)$ is an unregularized value function. For the direct policy parameterization where $\theta(s,a) = \pi_\theta(a|s)$, the gradient is*

$$\frac{\partial V^\pi(\rho)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d_\rho^\pi(s) Q^\pi(s,a). \tag{131}$$

Lemma D.2 (Policy gradient for soft-max parameterization) *Suppose that $V^\pi(\rho)$ is an unregularized value function. For the soft-max policy parameterization (cf. (6)), the gradient is*

$$\frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s,a). \tag{132}$$

Lemma D.3 (Performance difference) *Suppose that $V^\pi(\rho)$ is an unregularized value function. For all policies π and π', it holds that*

$$V^{\pi'}(\rho) - V^\pi(\rho) = \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_\rho^\pi(s) \sum_{a\in\mathcal{A}} (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi'}(s,a). \tag{133}$$

Lemma D.4 (Soft sub-optimality) *Suppose that $V_\tau^\pi(\rho)$ is an entropy-regularized value function and $\pi_\tau^\star$ is the optimal policy. For every policy π, it holds that*

$$V_\tau^{\pi^\star}(\rho) - V_\tau^\pi(\rho) = \frac{\tau}{1-\gamma} \sum_{s\in\mathcal{S}} d_\rho^\pi(s) D_{\mathrm{KL}}\left[\pi(\cdot|s) \mid \pi_\tau^\star(\cdot|s)\right], \tag{134}$$

where $D_{KL}[P(\cdot) \mid Q(\cdot)] := \sum_x P(x)(\log P(x) - \log Q(x))$ is the KL divergence between probability distributions $P(\cdot)$ and $Q(\cdot)$.