Algorithm for Constrained Markov Decision Process with Linear Convergence

Egor Gladin
Humboldt University of Berlin

Maksim Lavrik-Karmazin
Moscow Institute of Physics
and Technology

Karina Zainullina
Moscow Institute of Physics
and Technology

Varvara Rudenko
Moscow Institute of Physics
and Technology,
HSE University

Alexander Gasnikov
Moscow Institute of Physics
and Technology,
ISP RAS Research Center
for Trusted Artificial Intelligence

Martin Takáč
Mohamed bin Zayed University
of Artificial Intelligence

Abstract

The problem of constrained Markov decision process is considered. An agent aims to maximize the expected accumulated discounted reward subject to multiple constraints on its costs (the number of constraints is relatively small). A new dual approach is proposed with the integration of two ingredients: entropy-regularized policy optimizer and Vaidya's dual optimizer, both of which are critical to achieve faster convergence. The finite-time error bound of the proposed approach is provided. Despite the challenge of the nonconcave objective subject to nonconcave constraints, the proposed approach is shown to converge (with linear rate) to the global optimum. The complexity expressed in terms of the optimality gap and the constraint violation significantly improves upon the existing primal-dual approaches.

1 INTRODUCTION

In this paper we consider γ-discounted infinite-horizon constrained Markov decision process (CMDP) (Altman, 1999). Such problem arises in many practical applications, such as autonomous driving (Fisac et al., 2018), robotics (Ono et al., 2015) or systems where the agent must meet safety constraints. An example of such a problem is an energy-efficient wireless communication system that aims to consume minimum power without violating any constraint on quality service (Li et al., 2016). Such Reinforcement Learning (RL) problems are often formulated as CMDP (Garcia and Fernandez, 2015).

Recently, Ying et al. (2022); Li et al. (2021); Liu et al. (2021) proposed algorithms (under various assumptions) that achieve $\tilde{\mathcal{O}}\left(1/\epsilon\right)$[1] iteration complexity to find global optimum, where ϵ characterizes optimality gap and constraint violation. Each iteration of the proposed methods has the same complexity as an iteration of the Policy Gradient (PG) methods.

Although the CMDP problem is nonconcave (CMDP problem is typically a maximization problem) in policy π (nonconcavity inherited from MDP problem, which is nonconcave even in the bandit case (Mei et al., 2020b)), the complexity $\tilde{\mathcal{O}}\left(1/\epsilon\right)$ fits lower bound for smooth concave problems with large number of constraints (Nemirovsky, 1992; Ouyang and Xu, 2021). Despite that fact, if we have only a few constraints m — that is typical for most of the practical applications — these results are not optimal and we may expect $\tilde{\mathcal{O}}\left(m\right)$ iteration complexity for concave problems with m constraints (Gasnikov et al., 2016a; Gladin et al., 2020, 2021; Xu, 2020), which corresponds to lower bound for small enough m (Nemirovsky and Yudin, 1979). In this paper we are transferring $\tilde{\mathcal{O}}\left(m\right)$ iteration complexity result to the nonconcave CMDP problem.

1.1 Related Work

There is a considerable interest in RL / MDP problems (Sutton et al., 1999; Puterman, 2014; Bertsekas, 2019) and CMDP problems (Altman, 1999). For the past ten years there was a great theoretical progress in different direc-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

[1] For clarity we skip the dependence on $1-\gamma$ and logarithmic factors.

tions. For example, given a generative model with $|\mathcal{S}|$ states and $|\mathcal{A}|$ actions, we can find ϵ-policy (ϵ is a quality in terms of cumulative reward) for γ-discounted infinite-horizon MDP problem with

$$\tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|\cdot|\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right) \tag{1}$$

samples (Sidford et al., 2018; Wainwright, 2019; Agarwal et al., 2020) (analogously for CMDP, see arXiv version of Jin and Sidford (2020)) that corresponds (up to logarithmic factors) to the lower bound from the work by Azar et al. (2012). Moreover, the dependence on ϵ can be improved to $\log(1/\epsilon)$ at the expense of dependence on $|\mathcal{S}|$. Unfortunately, in many practical applications these optimal algorithms do not work at all due to the size of $|\mathcal{S}|$.

A popular way to escape the curse of dimensionality is to use PG methods (Mnih et al., 2015; Schulman et al., 2015b; Mei et al., 2020b; Agarwal et al., 2021), where a parameterized (for example by Deep Neural Networks (Li, 2017; Wang et al., 2020)) class of policies is considered. In the core of PG-type methods for MDP problems lie gradient-type methods (Mirror Descent (Lan, 2022; Zhan et al., 2021), Natural Policy Gradient (NPG) (Khodadadian et al., 2021; Cayci et al., 2021; Ding et al., 2020; Kakade, 2001; Cen et al., 2022), etc.) in the space of parameters applied to a properly regularized (in proper proximal setup) cumulative reward maximization problem. The gradient is calculated by using policy gradient theorem (Sutton et al., 1999), which reduces gradient calculation to Q-function (value function V) estimation. Under proper choice of regularizers (proximal setups), these methods require $\tilde{\mathcal{O}}\left((1-\gamma)^{-1}\right)$ iterations (the function value and policy converge linearly) and are not sensitive to inexactness δ of Q-value estimation ($\delta \sim \epsilon$), see details in Cen et al. (2022); Lan (2022); Zhan et al. (2021) and reference therein. Given a generative model, it is possible to obtain from these results (see Azar et al. (2012); Agarwal et al. (2020)) analogs of formula (1) for sample complexity that would be worse in terms of $(1-\gamma)$ dependence (Cen et al., 2022), but can be better in terms of $|\mathcal{S}|$.

For CMDP problems, PG methods are also well-developed, see, e.g., surveys in Li et al. (2021); Liu et al. (2021) and references therein. The best (in terms of PG iterations) known complexity bounds were obtained in these works Li et al. (2021); Liu et al. (2021); Ying et al. (2022).

In Ying et al. (2022) with additional strong assumption (initial state distribution covers the entire state space) the complexity bound $\tilde{\mathcal{O}}\left(\epsilon^{-1}\right)$ was obtained for entropy-regularized CMDP (for true CMDP – $\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$). In Li et al. (2021) the complexity bound $\tilde{\mathcal{O}}\left(\epsilon^{-1}\right)$ was obtained under weaker additional assumption (Markov chain induced by any stationary policy is ergodic) by using dual approach, see Section 1.2 for the details. For both of these approaches, given a generative model, one can obtain analogs of formula (1) for sample complexity that would be worse not only in terms of $(1-\gamma)$ dependence but also in terms of ϵ (but still can be better in terms of $|\mathcal{S}|$).

In Liu et al. (2021) the complexity bound $\tilde{\mathcal{O}}\left(\epsilon^{-1}\right)$ was obtained without additional assumptions. We summarize the described results in the Table 1, where ϵ is the accuracy in terms of optimality gap and constraint violation, ξ is the Slater's parameter (could be small), ζ is a numerical constant (small $\zeta \simeq 10^{-7}$, usually much larger in practice $\zeta \simeq 10^{-1}$) in Vaidya's cutting-plane method, β is the mixing time parameter (could be close to 1), $R_{\max} \leq r_{\max}\sqrt{m}$, $r_{\max}$ is the parameter that bounds all the rewards.

1.2 Main Contributions

In the core of our approach lies the paper Li et al. (2021), where the authors introduce entropy-regularized policy optimizer and solve regularized dual problem by proper version of Nesterov's accelerated gradient method. First of all, they use the strong duality for CMDP problem, which can be derived (Paternain et al., 2019) from the fact of compactness and convexity of the set of occupation measures (Borkar, 1988) or from Linear Programming representation of CMDP problem in discounted state-action visitation distribution (Altman, 1999). The next important step is entropy policy regularization. This regularization simultaneously solves several tasks at once. First of all, it allows to estimate the gradient of the dual function using NPG method that has a linear convergence rate of policy and is robust to inexactness in Q-function evaluations (Cen et al., 2022). This is crucial since the dual accelerated method is sensitive to inexactness in gradient, which can be controlled if policy converges fast. Secondly, this regularization allows to prove smoothness (in the spirit of Nesterov (2005) and with additional nice analysis of Mitrophanov's perturbation bounds (Mitrophanov, 2005; Zou et al., 2019) for showing that visitation measure is Lipschitz w.r.t. the policy) of the dual problem. The smoothness of the dual problem allows to use Nesterov's accelerated method to solve it and to get an optimal rate. The last step is the regularization of the dual problem to obtain a linear rate of convergence for the dual accelerated method, which negates the fact that we should solve the dual problem with higher accuracy to obtain the desired accuracy for the primal problem and constraint violation (Devolder et al., 2012; Gasnikov et al., 2016b). An alternative approach, which has

[2] Samples estimate was obtained based on the results from Cen et al. (2022); Zhan et al. (2021), where it was shown that ε (accuracy of V^π – output of NPG), determines the accuracy δ of Q-function evaluation $\delta \sim (1-\gamma)^2\varepsilon$. The sample complexity of δ-value of Q-function is $|S||A|(1-\gamma)^{-3}\delta^{-2}$. So sample complexity is [# NPG method iterations] $\cdot\, |S||A|(1-\gamma)^{-7}\varepsilon^{-2}$. Note that in such a way we obtain upper bound on samples complexity. In practice such theoretical bounds turn out to be greatly overestimated (they are far from being optimal in terms of $(1-\gamma)$), rather than estimates for the number of NPG iterations.

Table 1: Complexities of State-of-the-art Methods for CMDP

Method	β-Uniform Ergodicity assumption	# NPG method iterations	Accuracy of V^π	Samples[2]
AR-CPO (Li et al., 2021)	✓	$\tilde{\mathcal{O}}\left(\frac{R_{\max}\sqrt{m}}{((1-\beta)\xi)^{1/2}(1-\gamma)^3\epsilon}\right)$	$\sim \epsilon^2$	$\sim \epsilon^{-5}$
PMD-PD (Liu et al., 2021), ARNPG (Zhou et al., 2022)	✗	$\tilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3\epsilon}\left[\frac{m}{(1-\gamma)^2}+\frac{1}{\xi}\right]\right)$	$\sim \epsilon$	$\sim \epsilon^{-3}$
This work	✓	$\tilde{\mathcal{O}}\left(\frac{m}{(1-\gamma)\zeta}\right)$	$\sim \epsilon^2$	$\sim \epsilon^{-4}$

not been realized yet, is related to the primal-dual analysis of the method, which is used for the dual problem, see Nemirovski et al. (2010) for convex problems. In this approach, it is sufficient to solve the dual problem with the same accuracy as we wish to solve the primal one. This primal-dual approach may conserve the dependence on ϵ^{-2} in (1) in the final sample complexity estimate if the method we used for the dual problem does not accumulate an error in gradient over iterations. From Nemirovski et al. (2010); Gladin et al. (2020) it is known that the Ellipsoid method is a primal-dual one and does not accumulate an error in gradient.

Our contribution is to replace the dual accelerated method in the approach described above with Vaidya's cutting-plane method (Vaidya, 1989, 1996). Vaidya's method has a linear rate of convergence (without any regularization) and outperforms the accelerated method when dealing with small dimension problems (Bubeck, 2015). Moreover, Vaidya's method does not accumulate an error in gradient value (Gladin et al., 2021) and hence is more robust than the accelerated method.

We build a new way for CMDP problems to estimate the quality of the primal solution from the dual one. To the best of our knowledge, the developed technique is also new for standard convex (concave) inequalities constrained problems and quite different from the technique that was used in Li et al. (2021). This technique can be applied to any linear convergent algorithms for the dual problem. Moreover, we improve $|\mathcal{A}|$-times the bound on the Lipschitz gradient constant of the dual function from Li et al. (2021). In Lemma 7 Li et al. (2021) the authors formulate the correct result, but in fact, they prove a $|\mathcal{A}|$-times worse result than it was formulated. We give accurate proof of Lemma 7 by using the result from Appendix of Juditsky et al. (2005).

Another important point is that our proposed method allows to obtain the final policy in a straightforward way by performing a call to the same policy optimizer which is used on every iteration. By contrast, the algorithm from Li et al. (2021) can obtain the final policy formally only in the finite setting, by calculating ν_ρ^π in $|\mathcal{S}||\mathcal{A}|$ calls to value oracle (the notation is introduced below), which loses its motivation with transition to continuous state spaces.

Similarly to Li et al. (2021), our proposed method can be applied to a wider class of nonconvex/nonconcave constrained problems with strong duality (zero duality gap) and uniqueness of the solution of the auxiliary problem, which relates the primal variables with the dual ones.

Finally, we demonstrate by numerical experiments that our proposed algorithm indeed outperforms AR-CPO from Li et al. (2021) when m is not too big.

2 PRELIMINARIES

2.1 Markov Decision Process

A Markov decision process (MDP) is determined by a five-tuple $(\mathcal{S}, \mathcal{A}, \mathrm{P}, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition kernel, r is the reward function and $\gamma \in (0,1)$ is the discount factor. Assume that $\mathcal{S}$ and $\mathcal{A}$ are finite with cardinality $|\mathcal{S}|$ and $|\mathcal{A}|$, respectively. The initial state s_0 follows a distribution ρ. At any time $t \in \mathbb{N}_+$, an agent takes an action $a_t \in \mathcal{A}$ at state $s_t \in \mathcal{S}$, after which, according to the distribution $\mathrm{P}\left(s_{t+1} \mid s_t, a_t\right)$, the environment transits to the next state and the agent receives a reward $r\left(s_t, a_t\right)$.

A stationary policy maps a state $s \in \mathcal{S}$ to a distribution $\pi(\cdot \mid s)$ over $\mathcal{A}$, which does not depend on time t. For a given policy π, its value function for any initial state $s \in \mathcal{S}$ is defined as

$$V_r^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, a_t\right) \mid s_0 = s, a_t \sim \pi\left(\cdot \mid s_t\right), \right.$$
$$\left. s_{t+1} \sim \mathrm{P}\left(\cdot \mid s_t, a_t\right)\right].$$

Next, the mathematical expectation is taken with respect to the distribution of the initial state, the expected reward is determined by following policy π as

$$V_r^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho}\left[V_r^\pi\left(s_0\right)\right].$$

The goal is to solve the problem

$$\max_\pi V_r^\pi(\rho).$$

The discounted state-action visitation distribution defined

ν_ρ^π as follows:

$$\nu_\rho^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr\Big\{s_t = s, a_t = a \mid s_0 \sim \rho, \\ a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \mathrm{P}(\cdot \mid s_t, a_t)\Big\},$$

for any $s \in \mathcal{S}, a \in \mathcal{A}$. The value function thus can be equivalently written as

$$V_r^\pi(\rho) = \frac{\sum_{s\in\mathcal{S},a\in\mathcal{A}}\nu_\rho^\pi(s,a)r(s,a)}{1-\gamma} = \frac{\langle\nu_\rho^\pi, r\rangle_{\mathcal{S}\times\mathcal{A}}}{1-\gamma},$$

where $\langle\cdot,\cdot\rangle_{\mathcal{S}\times\mathcal{A}}$ denotes the inner product over the space $\mathcal{S}\times\mathcal{A}$ by reshaping ν_ρ^π and r as $|\mathcal{S}|\times|\mathcal{A}|$-dimensional vectors, and we omit the subscripts and use $\langle\cdot,\cdot\rangle$ when there is no confusion.

2.2 Constrained MDP

The difference between CMDP and MDP is that the reward is an $(m+1)$-dimensional vector:

$$r(s,a) = [r_0(s,a), r_1(s,a), \ldots, r_m(s,a)]^\top.$$

Each reward function $r_i,\ i = 0,1,\ldots,m$ is positive and finite,

$$r_{i,\max} := \max_{s\in\mathcal{S},a\in\mathcal{A}}\{r_i(s,a)\},$$

for $i = 0,1,\ldots,m$; $R_{\max} := \sqrt{\sum_{i=1}^m r_{i,\max}^2}$.

Then the value function defined with respect to the i-th component of the reward vector r as follows

$$V_i^\pi(s) := \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r_i(s_t,a_t) \mid s_0 = s, a_t \sim \pi(\cdot \mid s_t), \\ s_{t+1} \sim \mathrm{P}(\cdot \mid s_t, a_t)\Big]$$

and $V_i^\pi(\rho) = \mathbb{E}_{s_0\sim\rho}[V_i^\pi(s_0)]$, for $i = 0,1,\ldots,m$. The objective of the constrained MDP is to solve the following constrained optimization problem:

$$\begin{aligned} &\max_{\pi\in\Pi} V_0^\pi(\rho) \\ &\text{s.t. } V_i^\pi(\rho) \quad \geq c_i, \quad i = 1,\ldots,m, \end{aligned} \tag{2}$$

$\Pi = \{\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \sum_{a\in\mathcal{A}}\pi(a \mid s) = 1,\ \pi(a \mid s) \geq 0, \forall(s,a) \in \mathcal{S}\times\mathcal{A}\}$ is the set of all stationary policies. Let π^* denote the optimal policy for the problem (2). The goal is to find an ϵ-optimal policy defined as follows.

Definition 2.1. A policy $\tilde{\pi}$ is ϵ-optimal if its corresponding optimality gap and the constraint violation satisfy

$$V_0^*(\rho) - V_0^{\tilde{\pi}}(\rho) \leq \epsilon; \text{ and } \left\|\left[c - V^{\tilde{\pi}}(\rho)\right]_+\right\|_2 \leq \epsilon,$$

where $V_0^*(\rho)$ is the optimal value of (2),

$$V^{\tilde{\pi}}(\rho) := \left[V_1^{\tilde{\pi}}(\rho), \ldots, V_m^{\tilde{\pi}}(\rho)\right]^\top$$

and $c := [c_1, \ldots, c_m]^\top$.

2.3 Notation

Let I_m denote $m\times m$ identity matrix, $\mathbf{1}_m \in \mathbb{R}^m$ be a vector of ones. The set of nonnegative real numbers is denoted by $\mathbb{R}_+$. Notation $\operatorname{int} P$ is used for the interior of a set $P \subseteq \mathbb{R}^m$. Given a vector $x \in \mathbb{R}^m$, let $\|x\|_p, p \in [1,\infty]$ denote the p-norm of x, $[x]_+$ be defined by $([x]_+)_i = \max\{0, x_i\}, i = 1,\ldots,m$. For two vectors $x, y \in \mathbb{R}^m$, inner product is denoted by $\langle x, y\rangle$ or $x^\top y$. Given two functions $f(\epsilon)$ and $g(\epsilon)$, we write $f(\epsilon) = \mathcal{O}(g(\epsilon))$ if there exists some constant $C > 0$, such that $f(\epsilon) \leq Cg(\epsilon)$ for small enough ϵ. $\tilde{\mathcal{O}}(\cdot)$ means $\mathcal{O}(\cdot)$ up to logarithmic factor in a small power (usually 1 or 2). Bold symbol $\boldsymbol{\pi}$ denotes the constant ($\boldsymbol{\pi} = 3.1415\ldots$) contrary to plain π which indicates policy. $\log(\cdot)$ is the natural logarithm. For two probability measures P and Q defined on a measurable space $(\Omega, \mathcal{F})$, $d_{TV}(P,Q)$ denotes the total variation distance, i.e.,

$$d_{TV}(P,Q) := \sup_{A\in\mathcal{F}}|P(A) - Q(A)|. \tag{3}$$

3 CUTTING-PLANE ALGORITHM FOR CMDP

In this section, we introduce the cutting-plane algorithm for CMDP which is presented in Algorithm 1. The algorithm assumes access to:

1. Oracles, sufficent to run natural policy gradient algorithm (see Appendix A), for our MDP with arbitrary rewards. For example, access to exact gradient of value function w.r.t. softmax policy parametrization for any vector of rewards, and exact Fisher information matrix for a softmax-parametrized policy. Or, in our finite setting, also access to soft Q-functions is enough.

2. Exact value function of a policy w.r.t to constraints' reward vectors $r_1, \ldots, r_m$.

In the algorithm, policies may be stored as full softmax parametrizations (19), or directly as vectors $\pi(a|s),\ \forall(s,a) \in \mathcal{S}\times\mathcal{A}$ (both ways are equivalent and can be calculated one from another).

The core idea is to consider the entropy-regularized Lagrange function

$$\mathcal{L}_\tau(\pi,\lambda) := V_0^\pi(\rho) + \langle\lambda, V^\pi(\rho) - c\rangle + \tau\mathcal{H}(\pi),$$

where $\lambda \in \mathbb{R}_+^m$ is the vector of dual variables,

$$V^\pi(\rho) = [V_1^\pi(\rho), \ldots, V_m^\pi(\rho)]^\top \text{ and } c = [c_1, \ldots, c_m]^\top \tag{4}$$

are respectively vectors of constraints and constraint thresholds,

$$\mathcal{H}(\pi)=-\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t\log\left(\pi\left(a_t\mid s_t\right)\right)\Big| \\ s_0\sim\rho, a_t\sim\pi\left(\cdot\mid s_t\right), s_{t+1}\sim \mathrm{P}\left(\cdot\mid s_t,a_t\right)\Big]$$

is the discounted entropy of the policy π and $\tau>0$ is the regularization coefficient. The proposed method is based on two components: Vaidya's cutting-plane method (Vaidya, 1989, 1996) for solving the dual problem

$$\min_{\lambda\in\mathbb{R}_+^m}\big\{d_\tau(\lambda):=\max_{\pi\in\Pi}\mathcal{L}_\tau(\pi,\lambda)\big\}, \tag{5}$$

and an entropy-regularized policy optimizer for solving the inner problem $\max_{\pi\in\Pi}\mathcal{L}_\tau(\pi,\lambda)$ on each iteration of the outer loop. Below is the description of both components.

3.1 Entropy-regularized Policy Optimizer

To estimate the gradient of the dual function $\nabla d_\tau(\lambda)$, one has to solve the problem $\max_{\pi\in\Pi}\mathcal{L}_\tau(\pi,\lambda)$. Note that this is equivalent to maximizing an entropy-regularized value function corresponding to a reward function $r_\lambda := r_0+\sum_{i=1}^m\lambda_i r_i$. As mentioned in the introduction, entropy regularization enables the linear rate of convergence of an NPG method. In what follows, $\mathrm{NPG}\left(r_\lambda,\tau,\delta\right)$ represents a call to NPG procedure that learns the policy $\tilde{\pi}_{\tau,\lambda}$ for the entropy-regularized MDP with the regularization coefficient τ and reward r_λ up to δ-accuracy in terms of l_∞ distance to the unique optimal regularized policy, i.e.,

$$\left\|\pi^*_{\tau,\lambda}-\tilde{\pi}_{\tau,\lambda}\right\|_\infty<\delta, \tag{6}$$

where $\pi^*_{\tau,\lambda} := \arg\max_{\pi\in\Pi}\mathcal{L}_\tau(\pi,\lambda)$ (existence and uniqueness of such optimal policy is addressed below). More details on entropy-regularized policy optimizers are provided in Appendix A.

3.2 Vaidya's Cutting-plane Method

Vaidya's cutting-plane method (Vaidya, 1989, 1996) is an algorithm for a convex optimization problem with complexity $\mathcal{O}(m\log\frac{m}{\epsilon})$, which makes it a good choice for formulations with a small or moderate dimensionality like the dual problem (5). Moreover, it has been shown that the method can be used with an inexact subgradient, and it does not accumulate the error (Gladin et al., 2021). This makes it very suitable for the problem (5), since the gradient of the dual function is computed approximately.

We will now shortly list the necessary notions used in Algorithm 1 for the problem (2). A more detailed description of those notions is placed in Appendix B. For a matrix $A\in\mathbb{R}^{k\times m}$ with rows $a_i^\top, i=1,\ldots,k$, and a vector $b\in\mathbb{R}^k$, define the polytope

$$P(A,b):=\{\lambda\in\mathbb{R}^m:\ A\lambda\geq b\},$$

the matrix

$$H(\lambda;A,b):=\sum_{i=1}^k\frac{a_ia_i^\top}{\left(a_i^\top\lambda-b_i\right)^2}, \tag{7}$$

the values

$$\sigma_i(\lambda;A,b):=\frac{a_i^\top\left(H(\lambda;A,b)\right)^{-1}a_i}{\left(a_i^\top\lambda-b_i\right)^2},\quad 1\leq i\leq k, \tag{8}$$

the volumetric barrier of the polytope $P(A,b)$

$$\mathcal{V}(\lambda;A,b):=\frac{1}{2}\log\left(\det H(\lambda;A,b)\right),$$

where $\det H(\lambda;A,b)$ denotes the determinant of $H(\lambda;A,b)$. Finally, define the volumetric center of $P(A,b)$ as

$$\operatorname{VolCenter}(A,b):=\underset{\lambda\in\operatorname{int}P(A,b)}{\arg\min}\ \mathcal{V}(\lambda;A,b). \tag{9}$$

Since $\mathcal{V}$ is a self-concordant function of x, it can be efficiently minimized with Newton-type methods. The algorithm starts with a pair $(A_0,b_0)\in\mathbb{R}^{k_0\times m}\times\mathbb{R}^{k_0}$, such that the polytope $P(A_0,b_0)$ contains the search space. We refer the reader to Appendix B for more information on the original Vaidya's method and its parameters.

4 CONVERGENCE RESULTS FOR THE PROPOSED ALGORITHM

First, we introduce technical assumptions on our CMDP instance $(\mathcal{S},\mathcal{A},P,\gamma,r_0,r_1,\ldots,r_m,\rho)$ that are widely used in reinforcement learning literature, and state the regularized optimal policy uniqueness.

Assumption 4.1 (Slater Condition)**.** There exists a constant $\xi\in\mathbb{R}_+$, and at least one policy $\pi_\xi\in\Pi$, such that for all $i=1,\ldots,m$, $V_i^{\pi_\xi}(\rho)\geq c_i+\xi$.

Slater condition asserts that there exists a strictly feasible policy. Define the set

$$\Lambda:=\{\lambda\in\mathbb{R}_+^m\mid\|\lambda\|_1\leq B_\lambda\}\ \text{ with} \\ B_\lambda:=\frac{r_{0,\max}+\log|\mathcal{A}|}{(1-\gamma)\xi}. \tag{10}$$

Proposition 4.2 (Regularized optimal policy uniqueness)**.** *For any $\tau>0,\lambda\in\Lambda$ there exists exactly one optimal policy for the problem:*

$$\max_{\pi\in\Pi}\mathcal{L}_\tau(\pi,\lambda).$$

which we call $\pi^*_{\tau,\lambda}$.

We refer the reader to the Appendix C.1 for the proof.

Algorithm 1 Cutting-plane algorithm for CMDP

```
Input: number of outer iterations $T$, NPG accuracy $\delta$ (see (6)), pair $(A_0, b_0) \in \mathbb{R}^{k_0 \times m} \times \mathbb{R}^{k_0}$, algorithm parameters
    $\eta \leq 10^{-4}$, $\zeta \leq 10^{-3} \cdot \eta$.
for $t = 0, \ldots, T-1$ do
    $\lambda_t := \mathrm{VolCenter}(A_t, b_t)$
    Compute $H_t^{-1} := (H(\lambda_t; A_t, b_t))^{-1}$ and $\{\sigma_i(\lambda_t; A_t, b_t)\}_{i=1}^{k_t}$
    $i_t := \arg\min_{1 \leq i \leq k_t} \sigma_i(\lambda_t; A_t, b_t)$
    if $\sigma_{i_t}(\lambda_t; A_t, b_t) < \zeta$ then
        Obtain $(A_{t+1}, b_{t+1})$ by removing the $i_t$-th row from $(A_t, b_t)$,
        $k_{t+1} := k_t - 1$.
    else
        if $\lambda_t \in \mathbb{R}_+^m$ then
            $\pi_t := \mathrm{NPG}\left(r_0 + \langle \lambda_t, r \rangle, \tau, \delta\right)$
            $\widehat{\nabla}_t := c - V^{\pi_t}(\rho)$  (see notation in (4)),
        else
            Define $\widehat{\nabla}_t$ as the vector with components
                $(\widehat{\nabla}_t)_i = \begin{cases} 1, & (\lambda_t)_i < 0, \\ 0, & (\lambda_t)_i \geq 0, \end{cases} \quad i = 1, \ldots, m.$
        end if
        Find such $\beta_t \in \mathbb{R}$ that $\widehat{\nabla}_t^\top \lambda_t \geq \beta_t$ from the equation
                $\frac{\widehat{\nabla}_t^\top H_t^{-1} \widehat{\nabla}_t}{(\widehat{\nabla}_t^\top \lambda_t - \beta_t)^2} = \frac{1}{2}\sqrt{\eta\zeta},$
        $A_{t+1} := \begin{pmatrix} A_t \\ \widehat{\nabla}_t^\top \end{pmatrix}, \quad b_{t+1} := \begin{pmatrix} b_t \\ \beta_t \end{pmatrix}, \quad k_{t+1} = k_t + 1.$
    end if
end for
$\lambda_T = \arg\min_{\lambda \in \{\lambda_0, \ldots, \lambda_{T-1}\}} d_\tau(\lambda)$
$\pi_T := \mathrm{NPG}\left(r_0 + \langle \lambda_T, r \rangle, \tau, \delta\right)$
Output: $\pi_T$.
```

Assumption 4.3 (Uniform Ergodicity)**.** For any $\lambda \in \Lambda$, the Markov chain induced by the policy $\pi^*_{\tau,\lambda}$ and the Markov transition kernel is uniformly ergodic, i.e., there exist constants $C_M > 0$ and $0 < \beta < 1$ such that for all $t \geq 0$,

$$\sup_{s \in \mathcal{S}} d_{TV}\left(\mathbb{P}\left(s_t \in \cdot \mid s_0 = s\right), \chi_{\pi^*_{\tau,\lambda}}\right) \leq C_M \beta^t,$$

where $\chi_{\pi^*_{\tau,\lambda}}$ is the stationary distribution of the MDP induced by policy $\pi^*_{\tau,\lambda}$, and $d_{TV}(\cdot,\cdot)$ is the total variation distance, see (3).

Convergence rate of Algorithm 1 in terms of the optimality gap $V_0^*(\rho) - V_0^{\pi_T}(\rho)$ and the constraint violation $[c - V^{\pi_T}(\rho)]_+$ is described by the following theorem. The proof can be found in Appendix D.

Theorem 4.4. *Suppose Assumptions 4.1 and 4.3 hold, let $T \in \mathbb{N}$ be fixed and ϵ denote the value*

$$\epsilon := \frac{2m^2 B_\lambda}{\zeta}\left(\xi + \frac{\sqrt{m}R_{max}}{1-\gamma}\right)\exp\left(\frac{\log \boldsymbol{\pi} - \zeta T}{2m}\right), \tag{11}$$

where $\boldsymbol{\pi}$ denotes the constant ($\boldsymbol{\pi} = 3.1415\ldots$). The cutting-plane algorithm for CMDP (Algorithm 1) with parameters

$$k_0 := m+1, \quad A_0 := \left[\begin{array}{c} -I_m \\ 1 \end{array}\right],$$
$$b_0 := \left[\begin{array}{c} B_\lambda \mathbf{1}_m \\ m B_\lambda \end{array}\right], \quad \tau := \min\{1, \sqrt[3]{\epsilon}\}, \tag{12}$$

number of outer iterations T and NPG accuracy $\delta > 0$ (see (6)*) provides the following convergence guarantee of the* *optimality gap and the constraint violation:*

$$V_0^*(\rho) - V_0^{\pi_T}(\rho) \le \frac{B_\lambda R_{max}\sqrt{2mL_\beta}}{1-\gamma}\sqrt{\epsilon^{2/3}+6\gamma\delta} \tag{13}$$

$$+ 2\epsilon + 18\gamma\delta\sqrt[3]{\epsilon} + \frac{\log|\mathcal{A}|}{1-\gamma}\sqrt[3]{\epsilon}$$

$$+ \sqrt{m}B_\lambda \frac{L_\beta|\mathcal{A}|R_{\max}}{(1-\gamma)}\delta, \tag{14}$$

$$\left\|[c - V^{\pi_T}(\rho)]_+\right\|_2 \le \frac{2R_{max}^2 L_\beta}{1-\gamma}(\epsilon^{2/3}+6\gamma\delta)$$

$$+ \frac{L_\beta|\mathcal{A}|R_{\max}}{(1-\gamma)}\delta, \tag{15}$$

where

$$L_\beta := \left\lceil \log_\beta\left(C_M^{-1}\right)\right\rceil + (1-\beta)^{-1} + 1.$$

The value ϵ in (11) reflects linear convergence of Algorithm 1 in terms of the dual function. As it can be seen from (13) and (15), this also implies linear convergence in terms of value function and constraint violation, if NPG provides appropriate accuracy. Thus, the algorithm results in the following complexity bound

Corollary 4.5. *Algorithm 1 outputs an ϵ-optimal policy with respect to both the optimality gap and constraint violation after*

$$T = \mathcal{O}\left(\frac{m}{\zeta}\log\frac{m\log|\mathcal{A}|}{(1-\beta)(1-\gamma)\zeta\xi\epsilon}\right) \tag{16}$$

steps. The total number of calls to the policy gradient oracle made in all NPG calls is:

$$N_{oracle} = \mathcal{O}\left(T\cdot\frac{1}{1-\gamma}\log\frac{m\log|\mathcal{A}|}{(1-\beta)(1-\gamma)\xi\epsilon}\right) =$$

$$= \mathcal{O}\Bigg(\frac{m}{(1-\gamma)\zeta}\cdot\log\frac{m\log|\mathcal{A}|}{(1-\beta)(1-\gamma)\xi\epsilon}\cdot$$

$$\cdot\log\frac{m\log|\mathcal{A}|}{(1-\beta)(1-\gamma)\zeta\xi\epsilon}\Bigg).$$

Proof of the corollary is in Appendix D.1.

Remark 1. From the proof of the Corollary 1 in Appendix Li et al. (2021), the number of policy gradient method iterations is

$$N_{\text{AR-CPO}} = \frac{R_{\max}\sqrt{m}}{((1-\beta)\xi)^{1/2}(1-\gamma)^3\epsilon},$$

where we skip not only constants, but also log-factors, which could be close to zero, $R_{\max}^2 \le r_{\max}^2 m$. Note that for the concurrent paper Liu et al. (2021)

$$N_{\text{PMD-PD}} = \frac{1}{(1-\gamma)\epsilon}\left[\frac{m}{(1-\gamma)^2}+\frac{1}{\xi}\right]$$

up to log-factors. For our approach, $N_{\text{Vaidya}} = \frac{m}{(1-\gamma)\zeta}$ up to log-factors (depending on ξ, β, γ, ϵ), ζ – small numerical parameter of Vaidya's algorithm. Thus, from these formulas we may conclude that up to a log-factor, our approach is theoretically better when $\epsilon \lesssim \zeta\cdot(1-\gamma)^{-2}$. But in reality this log-factor might be significant.

Remark 2 (Dependence on the size of state-action space). It can be shown that the total complexity estimate for the proposed method is $\sim |\mathcal{S}||\mathcal{A}|$, assuming β is fixed. However, if an appropriate policy gradient method is chosen, the present analysis doesn't require the state space to be finite. Still, a finite action space is currently assumed.

4.1 Regularization of Dual Variables

The proposed approach can also be modified in the following way: the dual problem writes as

$$\max_{\lambda\in\mathbb{R}_+^m} d_{\tau,\mu}(\lambda) := d_\tau(\lambda) + \frac{\mu}{2}\|\lambda\|_2^2, \tag{17}$$

where $\mu > 0$ is the regularization coefficient. In this case, the vector $\widehat{\nabla}_t$ in Line 11 of Algorithm 1 will be replaced with

$$\widehat{\nabla}_t := c - V^{\pi_t}(\rho) - \mu\lambda_t.$$

If μ is chosen sufficiently small, the result of Corollary 4.5 remains true, see Appendix E for details.

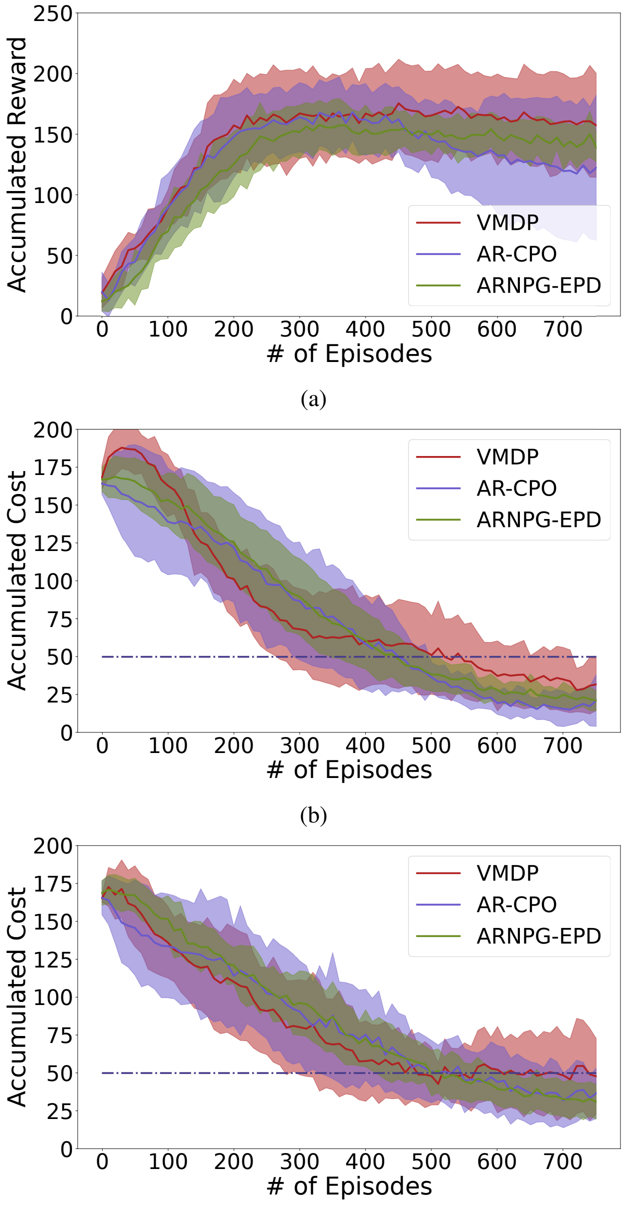
5 EXPERIMENTS

For our experiments we used Acrobot-v1, OpenAI Gym Mei et al. (2020a) environment. This environment contains two links connected linearly to form a chain, with one end of the chain fixed. The joint between the two links is actuated. The goal is to swing the end of the lower link up to a given height. Two additional constraints are implemented in order to have similar environment as in Li et al. (2021) for comparison purpose.

In Figure 1 we compare the proposed cutting-plane algorithm (VMDP) with the state-of-the-art primal-dual optimization (AR-CPO) method Li et al. (2021) and with ARNGP-EPG method Zhou et al. (2022). For a fair comparison, the same neural softmax policy and the trust region policy (Schulman et al., 2015a) optimization are used in all the algorithms.

Similarly to Li et al. (2021) we picture the average over 10 random initialized seeds and translucent error bands have the width of two standard deviations. The hyper parameters of AR-CPO algorithm are optimal from Li et al. (2021) and for ARNGP-EPG algorithm from Zhou et al. (2022). More information about experiments and parameters settings can be found in Appendix F.1 and F.2.

Figure 1a represents average total reward over episode, while Figures 1b and 1c show constraints with dashed line



(a)

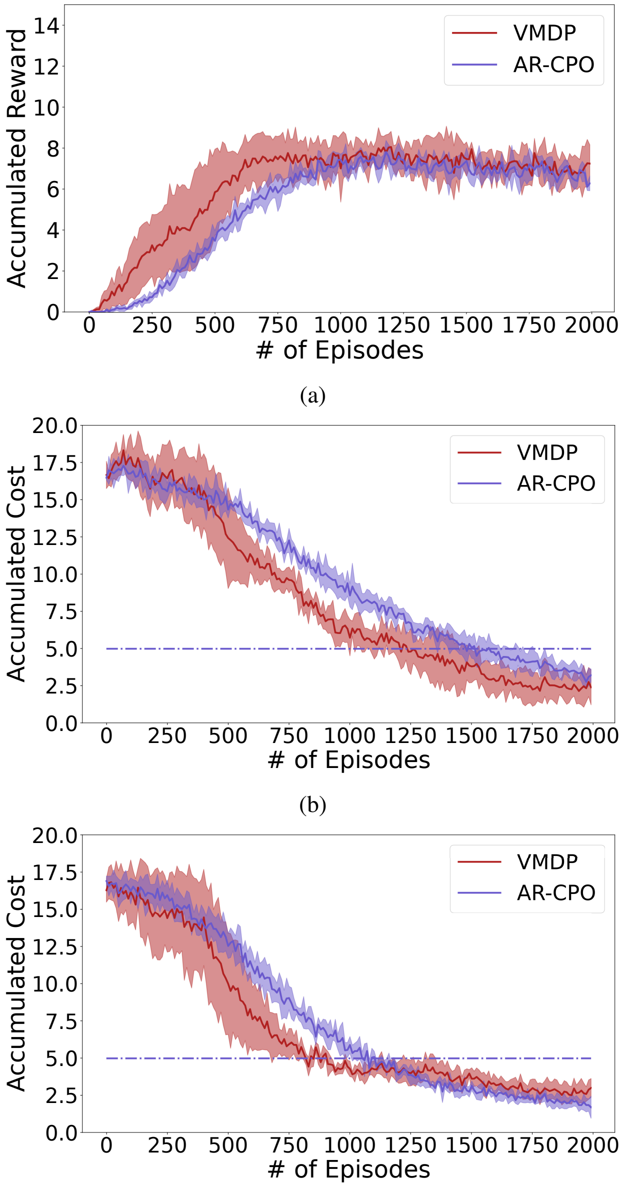
(b)

(c)

Figure 1: Average performance for VMDP, AR-CPO and ARNPG-EPD; the x-axis is training iteration.

as the constraint thresholds. We used total reward for a fair comparison with existing state-of-the-art approach. Moreover, in some cases total reward is more important in practise.

We find that VMDP algorithm achieves higher total reward with similar standard deviation. The speed of converge of both algorithms is similar. Thus, the proposed algorithm allows to achieve better performance with the same training time for an MDP task with the small number of constraints.



(a)

(b)

(c)

Figure 2: Average performance of VMDP and AR-CPO in case of discounted reward and costs; the x-axis is training iteration.

Discounted Reward Experiment Previously we considered experiments, where we calculated total reward and costs. We will now briefly describe the results of the experiments with discounted reward and costs. Considering similar environment as before, we set thresholds of 5 to the discounted costs. Figure 2 provides the comparison between AR-CPO with VMDP under their best tuned parameters provided in Appendix F.3.

In our experiment we used large number of policy optimization steps in subroutine equal to 40 for VMDP. This allowed the proposed algorithm to solve NPG subroutine with high accuracy and, as a result, to converge faster. In the same time, increasing the number of steps in subroutine did not make AR-CPO converge faster. It showed the best performance with this parameter equal to 1.

From Figure 2a, we observe that VMDP converges faster than AR-CPO, which is consistent with theory.

Thus, VMDP algorithm is useful in both discounted and total reward cases and shows better performance than AR-CPO.

6 CONCLUSION

In this paper we consider the constrained Markov decision process, where an agent aims to maximize the expected accumulated discounted reward subject to a relatively small number m of constraints on its costs. The best known algorithms achieve $\tilde{\mathcal{O}}\left(1/\epsilon\right)$ iteration complexity to find global optimum, where ϵ characterizes optimality gap and constraint violation. Each iteration of these algorithms has the same complexity as an iteration of the Policy Gradient (PG) methods. In this paper we improve (for m not too big) iteration complexity bound and obtain linear convergence $\tilde{\mathcal{O}}\left(m\right)$. Limitations of the method include the assumptions that the number of constraints m is moderate, and the action space is finite, see Remark 2.

One possible direction of improvement is to eliminate Assumption 4.3. Presumably, if Vaidya's method is shown to be primal-dual, the smoothness of the dual function (which was used to restore the solution of the direct problem) will no longer be required. Alternatively, Vaidya's method can be replaced with the ellipsoid method, which is known to be primal-dual, but has a slightly worse dependence on the number of constraints m.

Acknowledgements

The work of E. Gladin is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

The work of A. Gasnikov was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

We thank the reviewers who gave valuable comments that helped improve the present work.

References

A. Agarwal, S. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

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

E. Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

M. G. Azar, R. Munos, and B. Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

Bertsekas. *Nonlinear programming*. Athena Scientific, 1991.

D. Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.

V. S. Borkar. A convex analytic approach to markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602, 1988.

S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4): 231–357, nov 2015. ISSN 1935-8237. doi: 10.1561/2200000050. URL `https://doi.org/10.1561/2200000050`.

S. Cayci, N. He, and R. Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022. doi: 10.1287/opre.2021.2151. URL `https://doi.org/10.1287/opre.2021.2151`.

O. Devolder, F. Glineur, and Y. Nesterov. Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM Journal on Optimization*, 22(2): 702–727, 2012.

D. Ding, K. Zhang, T. Basar, and M. Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

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

J. Garcia and F. Fernandez. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

A. Gasnikov, D. Kamzolov, and M. Mendel. Universal composite prox-method for strictly convex optimization problems. *arXiv preprint arXiv:1603.07701*, 2016a.

A. V. Gasnikov, E. Gasnikova, Y. E. Nesterov, and A. Chernov. Efficient numerical methods for entropy-linear programming problems. *Computational Mathematics and Mathematical Physics*, 56(4):514–524, 2016b.

E. Gladin, I. Kuruzov, F. Stonyakin, D. Pasechnyuk, M. Alkousa, and A. Gasnikov. Solving strongly convex-concave composite saddle point problems with a small dimension of one of the variables, 2020.

E. Gladin, A. Sadiev, A. V. Gasnikov, P. E. Dvurechensky, A. Beznosikov, and M. S. Alkousa. Solving smooth min-min and min-max problems by mixed oracle algorithms. *Communications in Computer and Information Science*, 2021.

J. Ho and S. Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Y. Jin and A. Sidford. Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pages 4890–4900. PMLR, 2020.

A. B. Juditsky, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41(4):368–384, 2005.

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

S. Khodadadian, P. R. Jhunjhunwala, S. M. Varma, and S. T. Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

P. Li, Y. Jiang, W. Li, F. Zheng, and X. You. A cmdp-based approach for energy efficient power allocation in massive mimo systems. In *2016 IEEE Wireless Communications and Networking Conference*, pages 1–6. IEEE, 2016.

T. Li, Z. Guan, S. Zou, T. Xu, Y. Liang, and G. Lan. Faster algorithm and sharper analysis for constrained markov decision process, 2021. URL `https://arxiv.org/abs/2110.10351`.

Y. Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian. Policy optimization for constrained mdps with provable fast global convergence. *arXiv preprint arXiv:2111.00552*, 2021.

J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *ICML*, 2020a.

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

A. Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

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

A. Nemirovski, S. Onn, and U. G. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

A. Nemirovsky and D. Yudin. Problem complexity and optimization method efficiency. *M.: Nauka (in Russian)*, 1979.

A. S. Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

M. Ono, M. Pavone, Y. Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571, 2015.

Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

B. T. Polyak. Introduction to optimization. *Inc., Publications Division, New York*, 1987.

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

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. 37:1889–1897, 07–09 Jul 2015a. URL `https://proceedings.mlr.press/v37/schulman15.html`.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015b.

A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

U. Syed, M. Bowling, and R. E. Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pages 1032–1039, 2008.

P. M. Vaidya. A new algorithm for minimizing convex functions over convex sets. In *30th Annual Symposium on Foundations of Computer Science*, pages 338–343. IEEE Computer Society, 1989.

P. M. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical programming*, 73(3):291–341, 1996.

M. J. Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

L. Wang, Q. Cai, Z. Yang, and Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BJgQfkSYDS`.

Y. Xu. First-order methods for problems with o (1) functional constraints can have almost the same convergence rate as for unconstrained problems. *arXiv preprint arXiv:2010.02282*, 2020.

D. Ying, Y. Ding, and J. Lavaei. A dual approach to constrained markov decision processes with entropy regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 1887–1909. PMLR, 2022.

W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *arXiv preprint arXiv:2105.11066*, 2021.

R. Zhou, T. Liu, D. Kalathil, P. Kumar, and C. Tian. Anchor-changing regularized natural policy gradient for multi-objective reinforcement learning. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=kY1RbKE7DWE`.

S. Zou, T. Xu, and Y. Liang. Finite-sample analysis for sarsa with linear function approximation. *Advances in neural information processing systems*, 32, 2019.

A NATURAL POLICY GRADIENT (NPG)

NPG is one of the algorithms that can efficently optimize a finite MDP with relative entropy regularization:

$$\max_{\pi\in\Pi} V_\tau^\pi(\rho) = V^\pi(\rho) + \tau\mathcal{H}(\pi), \tag{18}$$

assuming access to gradients of their soft value function and to a Fisher information matrix respective to its softmax parametrization. Specifically, policies are parametrized as follows:

$$\pi^\theta(a|s) = \frac{\exp\big(e^{\theta_{s,a}}\big)}{\sum_{a'}\exp\big(e^{\theta_{s,a'}}\big)}, \tag{19}$$

$$\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}, \tag{20}$$

and the NPG algorithm has access to functions:

$$G(\theta) = \nabla_\theta V_\tau^{\pi^\theta}(\rho), \tag{21}$$

$$\mathcal{F}_\rho^\pi(\theta) := \mathop{\mathbb{E}}_{s\sim d_\rho^{\pi^\theta}, a\sim\pi\theta(\cdot|s)}\Big[\big(\nabla_\theta \log\pi_\theta(a|s)\big)\big(\nabla_\theta\log\pi_\theta(a|s)\big)^\top\Big]. \tag{22}$$

This type of oracle is motivated by a possibility of estimating this gradient in high-dimension MDP's in applications.

A.1 Algorithm and Convergence Rates

The algorithm looks as follows:

Algorithm 2 Natural policy gradient (NPG) algorithm

```
Input: learning rate $\eta > 0$, initialization parameters $\theta_0 = 0$.
1: for $t = 0, 1, 2, \dots$ do
2:     $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta\big(\mathcal{F}_\rho^{\theta^{(t)}}\big)^\dagger \nabla_\theta V_\tau^{\pi_{\theta^{(t)}}}(\rho)$
3: end for
```

($M^\dagger$ is Moore-Penrose inverse function)

The update rule in line 2 of Algorithm 2 can be rewritten in terms of policies and soft Q-functions:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \quad \pi^{(t+1)}(a|s) = \frac{1}{Z^{(t)}(s)}\big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}}\exp\Big(\frac{\eta Q_\tau^{(t)}(s,a)}{1-\gamma}\Big), \tag{23}$$

where $Z^{(t)}(s) = \sum_{a'\in\mathcal{A}}\big(\pi^{(t)}(a'|s)\big)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta Q_\tau^{(t)}(s,a')}{1-\gamma}\big)$ is a normalizing coefficient, and soft Q-functions are defined as follows:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \qquad Q_\tau^\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V_\tau^\pi(s')\big], \tag{24}$$

So, in our finite setting we can assume we are given an oracle of soft Q-functions instead of the earlier mentioned.

In Cen et al. (2022) the following theorem is proved (in setting with $r(s,a)\in[0,1]$):

Theorem A.1 (Linear convergence of exact entropy-regularized NPG)**.** *For any learning rate* $0 < \eta \le (1-\gamma)/\tau$*, the entropy-regularized NPG updates* (23) *satisfy*

$$\big\|Q_\tau^* - Q_\tau^{(t+1)}\big\|_\infty \le C_1\gamma(1-\eta\tau)^t, \tag{25}$$

$$\big\|\log\pi_\tau^* - \log\pi^{(t+1)}\big\|_\infty \le 2C_1\tau^{-1}(1-\eta\tau)^t, \tag{26}$$

$$\big\|V_\tau^* - V_\tau^{(t+1)}\big\|_\infty \le 3C_1\gamma(1-\eta\tau)^t. \tag{27}$$

for all $t \ge 0$*, where*

$$C_1 := \big\|Q_\tau^* - Q_\tau^{(0)}\big\|_\infty + 2\tau\left(1 - \frac{\eta\tau}{1-\gamma}\right)\big\|\log\pi_\tau^* - \log\pi^{(0)}\big\|_\infty. \tag{28}$$

We will use the algorithm with $\eta = \frac{1-\gamma}{\tau}$. In this case we have convergence rates:

$$\left\|Q_\tau^* - Q_\tau^{(t+1)}\right\|_\infty \leq C_1 \gamma^{t+1}, \tag{29}$$

$$\left\|\pi_\tau^* - \pi_\tau^{(t+1)}\right\|_\infty \leq \left\|\log \pi_\tau^* - \log \pi^{(t+1)}\right\|_\infty \leq 2C_1 \tau^{-1} \gamma^t, \tag{30}$$

$$\left\|V_\tau^* - V_\tau^{(t+1)}\right\|_\infty \leq 3C_1 \gamma^{t+1}, \tag{31}$$

$$\text{with } C_1 = \left\|Q_\tau^* - Q_\tau^{(0)}\right\|_\infty. \tag{32}$$

A.2 Usage of NPG in Our Work

In our algorithm we need to solve auxiliary problems of the form:

$$\max_{\pi \in \Pi} V_{\tau,\lambda}^\pi(\rho) := V_0^\pi(\rho) + \sum_{i=1}^m \lambda_i V_i^\pi(\rho) + \tau \mathcal{H}(\pi) \tag{33}$$

with this procedure.

Since the objective is a τ-regularized value function for an MDP with rewards $r_0 + \langle \lambda, r \rangle$, we can use the NPG procedure to optimize it. However, we cannot pass our MDP with these rewards directly to this method, because it assumes $r(s,a) \in [0,1]$ in Cen et al. (2022). So, we will scale both r and τ to make rewards satisfy this condition, and run NPG with a higher accuracy.

Specifically, we define a procedure $\mathrm{NPG}(r, \tau, \delta)$ as follows.

First, define $R = \max(\max_{s,a} r(s,a), 1)$. Calculate $r'(s,a) = \frac{r(s,a)}{R}$.

Then, apply NPG algorithm to solve an MDP with rewards r' and regularization coefficient $\frac{\tau}{R}$ with accuracy $\frac{\delta}{R}$ for the policy. For that we need a number of iterations $t+1$ that satisfies:

$$2\frac{C_1}{R}\left(\frac{\tau}{R}\right)^{-1} \gamma^t < \frac{\delta}{R}, \tag{34}$$

$$\gamma^t < \frac{\delta \tau}{2C_1 R}, \tag{35}$$

$$t > \frac{\log 2C_1 R + \log \delta^{-1} + \log \tau^{-1}}{\log \gamma^{-1}}. \tag{36}$$

By this we get a δ-optimal policy in terms of l_∞ distance to the optimal policy, since it is the same after rescaling and $R \geq 1$. Also, by 34:

$$2C_1 \tau^{-1} \gamma^t < \frac{\delta}{R}, \tag{37}$$

$$3C_1 \gamma^{t+1} < \frac{6\tau\gamma}{R}\delta, \tag{38}$$

$$\left\|V_\tau^* - V_\tau^{(t+1)}\right\|_\infty \leq 6\tau\gamma\delta. \tag{39}$$

Finally, we get this statement:

Theorem A.2. *Suppose $\delta, \tau > 0$, and we have a τ-regularized MDP. Let $R = \max(\max_{s,a} r(s,a), 1)$. Then a number of NPG iterations more than:*

$$T = \frac{\log 2C_1 R + \log \delta^{-1} + \log \tau^{-1}}{\log \gamma^{-1}} + 1 \tag{40}$$

is enough for our procedure to acquire a policy $\tilde{\pi}$ that satisfies:

$$\left\|\pi_\tau^* - \tilde{\pi}\right\|_\infty < \delta, \tag{41}$$

$$\left\|V_\tau^* - V_\tau^{\tilde{\pi}}\right\|_\infty \leq 6\tau\gamma\delta. \tag{42}$$

B DESCRIPTION OF VAIDYA'S CUTTING-PLANE METHOD

Vaidya proposed a cutting-plane method from Vaidya (1989, 1996) for solving problems of the form

$$\min_{\lambda \in \Lambda} d(\lambda), \tag{43}$$

where Λ is a compact convex set with non-empty interior, $d(\lambda)$ is a continuous convex function. We will now introduce the notation and describe the algorithm. Let $P(A, b)$ denote the bounded full-dimensional polytope of the form

$$P(A, b) = \{\lambda \in \mathbb{R}^m : A\lambda \geq b\} \text{ where } A \in \mathbb{R}^{k \times m} \text{ and } b \in \mathbb{R}^k.$$

The logarithmic barrier for P is defined as

$$L(\lambda; A, b) := -\sum_{i=1}^{k} \log\left(a_i^\top \lambda - b_i\right),$$

where $a_i^\top$ is the i^{th} row of $A,\ i = 1, \ldots, k$. The Hessian of $L(\lambda)$ is given by

$$H(\lambda; A, b) = \sum_{i=1}^{k} \frac{a_i a_i^\top}{\left(a_i^\top \lambda - b_i\right)^2}, \tag{44}$$

and is positive definite for all λ in $\operatorname{int} P$ (interior of P). The *volumetric barrier* for $P(A, b)$ is defined as

$$\mathcal{V}(\lambda; A, b) = \frac{1}{2} \log\left(\det H(\lambda; A, b)\right),$$

where $\det H(\lambda; A, b)$ denotes the determinant of $H(\lambda; A, b)$. Let also $\sigma_i(\lambda; A, b)$ denote the values

$$\sigma_i(\lambda; A, b) = \frac{a_i^\top \left(H(\lambda; A, b)\right)^{-1} a_i}{\left(a_i^\top \lambda - b_i\right)^2}, \quad 1 \leq i \leq k. \tag{45}$$

Volumetric center of P is defined as the point that minimizes $\mathcal{V}(\lambda; A, b)$ over the interior of P:

$$\operatorname{VolCenter}(A, b) := \underset{\lambda \in \operatorname{int} P(A,b)}{\arg\min}\ \mathcal{V}(\lambda; A, b). \tag{46}$$

Volumetric barrier $\mathcal{V}$ is a self-concordant function and can therefore be efficiently minimized with the Newton-type methods. For more details and theoretical analysis, refer to Vaidya (1996, 1989). Consider the following version of inexact subgradient.

Definition B.1. Vector $\nu \in \mathbb{R}^m$ is called a δ-subgradient of a convex function f at $z \in \operatorname{dom} d$ (denoted $\nu \in \partial_\delta d(z)$), if

$$d(\lambda) \geq d(z) + \nu^\top (\lambda - z) - \delta \quad \forall \lambda \in \operatorname{dom} d.$$

If $\delta = 0$, this we get the usual definition of subgradient $\partial_\delta d(z) = \partial d(z)$.

It has been proved that one can use δ-subgradient instead of the exact subgradient in Vaidya's method Gladin et al. (2021). Algorithm 3 gives the version of the method using δ-subgradients. The method produces a sequence of pairs $(A_t, b_t) \in \mathbb{R}^{k_t \times m} \times \mathbb{R}^{k_t}$, such that the corresponding polytopes contain a solution of the problem (43). A simplex containing the set Λ is often taken as the initial polytope (A_0, b_0). For example, if $\|\lambda\|_2 \leq \mathcal{R}$ for any $\lambda \in \Lambda$, then a possible choice of a starting polytope is

$$P_0 = \left\{\lambda \in \mathbb{R}^m : \lambda_j \geqslant -\mathcal{R}, j = \overline{1, m}, \sum_{j=1}^{m} \lambda_j \leqslant m\mathcal{R}\right\} \supseteq \mathcal{B}_{\mathcal{R}} \supseteq \Lambda,$$

that is,

$$k_0 = m + 1, \quad b_0 = -\mathcal{R} \begin{bmatrix} \mathbf{1}_m \\ m \end{bmatrix}, \quad A_0 = \begin{bmatrix} I_m \\ -\mathbf{1}_m^\top \end{bmatrix}.$$

Algorithm 3 Vaidya's cutting-plane method with δ-subgradient

```
Input: number of outer iterations $T$, pair $(A_0, b_0) \in \mathbb{R}^{k_0 \times m} \times \mathbb{R}^{k_0}$, algorithm parameters $\eta \leq 10^{-4}$, $\zeta \leq 10^{-3} \cdot \eta$.
for $t = 0, \ldots, T-1$ do
    $\lambda_t := \mathrm{VolCenter}(A, b)$
    Compute $H_t^{-1} := (H(\lambda_t; A_t, b_t))^{-1}$ and $\{\sigma_i(\lambda_t; A_t, b_t)\}_{i=1}^{k_t}$
    $i_t := \arg\min_{1 \leq i \leq k_t} \sigma_i(\lambda_t; A_t, b_t)$
    if $\sigma_{i_t}(\lambda_t; A_t, b_t) < \zeta$ then
        Obtain $(A_{t+1}, b_{t+1})$ by removing the $i_t$-th row from $(A_t, b_t)$,
        $k_{t+1} := k_t - 1$.
    else
        if $\lambda_t \in \mathbb{R}_+^m$ then
            Take $\widehat{\nabla}_t \in -\partial_\delta d(\lambda_t)$,
        else
            Take $\widehat{\nabla}_t$ such that $\widehat{\nabla}_t^\top \lambda \geq \widehat{\nabla}_t^\top \lambda_t \; \forall \lambda \in \Lambda$.
        end if
        Find such $\beta_t \in \mathbb{R}$ that $\widehat{\nabla}_t^\top \lambda_t \geq \beta_t$ from the equation
            $$\frac{\widehat{\nabla}_t^\top H_t^{-1} \widehat{\nabla}_t}{(\widehat{\nabla}_t^\top \lambda_t - \beta_t)^2} = \frac{1}{2}\sqrt{\eta\zeta},$$
        $A_{t+1} := \begin{pmatrix} A_t \\ \widehat{\nabla}_t^\top \end{pmatrix}, \quad b_{t+1} := \begin{pmatrix} b_t \\ \beta_t \end{pmatrix}, \quad k_{t+1} = k_t + 1.$
    end if
end for
$\lambda_T = \arg\min_{\lambda \in \{\lambda_0, \ldots, \lambda_{T-1}\}} d_\tau(\lambda)$
Output: $\lambda_T$.
```

Theorem B.2. *Let $\mathcal{B}_{\mathcal{R}_{in}}$ and $\mathcal{B}_{\mathcal{R}}$ be some Euclidean balls of radii $\mathcal{R}_{in}$ and $\mathcal{R}$, respectively, such that $\mathcal{B}_{\mathcal{R}_{in}} \subseteq \Lambda \subseteq \mathcal{B}_{\mathcal{R}}$, and let a number $B > 0$ be such that $|d(\lambda) - d(\lambda')| \leq B \; \forall \lambda, \lambda' \in \Lambda$. After $T \geq \frac{2m}{\zeta} \log\left(\frac{m^{1.5}\mathcal{R}}{\gamma \mathcal{R}_{in}}\right) + \frac{1}{\gamma} \log \boldsymbol{\pi}$ iterations Vaidya's method with δ-subgradient for the problem* (43) *returns a point λ_T such that*

$$d(\lambda_T) - \min_{\lambda \in \Lambda} d(\lambda) \leq \frac{Bm^{1.5}\mathcal{R}}{\zeta \mathcal{R}_{in}} \exp\left(\frac{\log \boldsymbol{\pi} - \zeta T}{2m}\right) + \delta, \tag{47}$$

where $\zeta > 0$ is the parameter of the algorithm.

Corollary B.3. *Vaidya's cutting-plane method with δ-subgradient achieves accuracy ϵ after*

$$T = \left\lceil \frac{2m}{\zeta} \log\left((\epsilon - \delta)^{-1} \frac{Bm^{1.5}\mathcal{R}}{\zeta \mathcal{R}_{in}}\right) + \frac{\log \boldsymbol{\pi}}{\zeta} \right\rceil, \tag{48}$$

provided that $\epsilon > \delta$ and $\epsilon - \delta \leq B$.

C SUPPORTING STATEMENTS

In the first part of this section, we establish the regularized optimal policy uniqueness (see Proposition 4.2). After that, we prove several lemmas and propositions used in the proof of Theorem 4.4.

C.1 Regularized Optimal Policy Uniqueness

For a policy $\pi \in \Pi$, let us denote by $\mathbb{E}_\pi$ the expectation with respect to the trajectory it generates, i.e., for some function $f(s, a)$, we write $\mathbb{E}_\pi[f(s, a)] := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right]$, where $s_0 \sim \rho, a_t \sim \pi(\cdot \mid s_t)$, and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ for $t \geq 0$.

Thus, we have

$$\begin{aligned} V_i^\pi(\rho) &\equiv \mathbb{E}_\pi[r_i(s,a)], \\ \mathcal{H}(\pi) &\equiv \mathbb{E}_\pi[-\log \pi(a \mid s)]. \end{aligned}$$

We now present the proof of Proposition 4.2 inspired by Ho and Ermon (2016). Let us transform the optimization problem over policies into a convex problem. For a policy $\pi \in \Pi$, define its occupancy measure $p_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$p_\pi(s,a) = \pi(a \mid s) \sum_{t=0}^{\infty} \gamma^t P\left(s_t = s \mid \pi\right).$$

The occupancy measure allows us to write

$$V_i^\pi(\rho) \equiv \mathbb{E}_\pi[r_i(s,a)] = \sum_{s,a} p_\pi(s,a) r_i(s,a).$$

A basic result (Puterman, 2014) is that the set of valid occupancy measures $\mathcal{D} \triangleq \{p_\pi : \pi \in \Pi\}$ can be written as a feasible set of affine constraints: if $\rho(s)$ is the distribution of starting states and $P\left(s' \mid s, a\right)$ is the dynamics model, then

$$\mathcal{D} = \Big\{ p : p \geq 0, \sum_a p(s,a) = \rho(s) + \gamma \sum_{s',a} P\left(s \mid s', a\right) p\left(s', a\right) \forall s \in \mathcal{S} \Big\}.$$

Furthermore, there is a one-to-one correspondence between Π and $\mathcal{D}$:

Proposition C.1 (Theorem 2 in Syed et al. (2008))**.** *If $p \in \mathcal{D}$, then p is the occupancy measure for $\pi_p(a \mid s) \triangleq p(s,a)/\sum_{a'} p\left(s, a'\right)$, and π_p is the only policy whose occupancy measure is p.*

We are therefore justified in writing π_p to denote the unique policy for an occupancy measure p. To show that the problem

$$\max_{\pi \in \Pi} \mathcal{L}_\tau(\pi, \lambda). \tag{49}$$

is equivalent to a maximization problem with a strictly concave objective, we will need the following lemma:

Lemma C.2 (Lemma 3.1. in Ho and Ermon (2016))**.** *Let*

$$\bar{H}(p) = -\sum_{s,a} p(s,a) \log\Big(p(s,a)/\sum_{a'} p\left(s, a'\right)\Big).$$

Then, $\bar{H}$ is strictly concave, and for all $\pi \in \Pi$ and $p \in \mathcal{D}$, we have $\mathcal{H}(\pi) = \bar{H}\left(p_\pi\right)$ and $\bar{H}(p) = \mathcal{H}\left(\pi_p\right)$.

Proposition C.1 and Lemma C.2 together allow us to switch between policies and occupancy measures. That is, if

$$\bar{\mathcal{L}}_\tau(p, \lambda) := \sum_{s,a} p(s,a) r_\lambda(s,a) - \langle \lambda, c \rangle + \tau \bar{H}(p),$$

with $r_\lambda(s,a) := r_0(s,a) + \sum_{i=1}^m \lambda_i r_i(s,a)$, then $\mathcal{L}_\tau(\pi, \lambda) = \bar{\mathcal{L}}_\tau(p_\pi, \lambda)$ for all policies $\pi \in \Pi$, and $\bar{\mathcal{L}}_\tau(p_\pi, \lambda) = \mathcal{L}_\tau(\pi_p, \lambda)$ for all occupancy measures $p \in \mathcal{D}$. In other words, the problem (49) is equivalent to the problem

$$\max_{p \in \mathcal{D}} \bar{\mathcal{L}}_\tau(p, \lambda), \tag{50}$$

where $\mathcal{D}$ is a convex set and $\bar{\mathcal{L}}_\tau(p, \lambda)$ is a strictly concave function of p. Thus, the solution p^* to (50) is unique which implies the uniqueness of the solution $\pi^* = \pi_{p^*}$ to (49).

C.2 Supporting Lemmas

From now on, we use notation introduced in Sections 2, 4. In particular, we are considering the dual problem

$$\min_{\lambda \in \mathbb{R}_+^m} \big\{ d_\tau(\lambda) := \max_{\pi \in \Pi} \mathcal{L}_\tau(\pi, \lambda) \big\}. \tag{51}$$

The first lemma establishes upper bound on the norm of a minimizer of the dual function.

Lemma C.3 (see also Li et al. (2021))**.** *Suppose Assumption 4.1 holds. Let λ^*_τ be a solution of the dual problem* (51)*. Then*

$$\|\lambda^*_\tau\|_1 \leq B_\lambda := \frac{r_{0,\max} + \log|\mathcal{A}|}{(1-\gamma)\xi}.$$

Proof. Note that $\mathcal{H}(\pi) \leq \frac{\log|\mathcal{A}|}{1-\gamma}$,

$$\begin{aligned}
&d_\tau(\lambda^*_\tau) \geq V_0^{\pi_\xi}(\rho) + \langle \lambda^*_\tau, V^{\pi_\xi}(\rho) - c\rangle + \tau\mathcal{H}(\pi) \geq \xi\|\lambda^*_\tau\|_1,\\
&d_\tau(\lambda^*_\tau) \leq d_\tau(\lambda^*) \leq d(\lambda^*) + \frac{\tau\log|\mathcal{A}|}{1-\gamma} = V_0^{\pi^*}(\rho) + \frac{\tau\log|\mathcal{A}|}{1-\gamma} \leq \frac{r_{0,max} + \tau\log|\mathcal{A}|}{1-\gamma},\\
&\|\lambda^*_\tau\|_1 \leq \frac{r_{0,\max} + \tau\log|\mathcal{A}|}{(1-\gamma)\xi}.
\end{aligned}$$

□

Recall that Λ is defined as the set

$$\Lambda := \{\lambda \in \mathbb{R}^m_+ \mid \|\lambda\|_1 \leq B_\lambda\}. \tag{52}$$

The second lemma gives an example of two Euclidean ball, one of which is contained in Λ and the other one contains Λ.

Lemma C.4. *Let $\mathcal{R} := B_\lambda$, $\mathcal{R}_{in} := \frac{B_\lambda}{m+\sqrt{m}}$, then $\mathcal{B}_{\mathcal{R}_{in}} \subseteq \Lambda \subseteq \mathcal{B}_{\mathcal{R}}$, with $\mathcal{B}_{\mathcal{R}_{in}}$ being the Euclidean ball of radius $\mathcal{R}_{in}$ centered at the point $\lambda_{in} := \mathcal{R}_{in} \cdot \mathbf{1}_m$, $\mathcal{B}_{\mathcal{R}}$ being the Euclidean ball of radius $\mathcal{R}$ centered at the origin.*

Proof. To prove the first inclusion, observe that for any $\lambda \in \mathcal{B}_{\mathcal{R}_{in}}$ it holds $\lambda \in \mathbb{R}^m_+$, which implies $\|\lambda\|_1 = \sum_{i=1}^m \lambda_i$. Maximization of this sum subject to constraint $\|\lambda - \lambda_{in}\|_2^2 \leq \mathcal{R}_{in}^2$ yields optimal point $\lambda^{(1)} := \lambda_{in} + \frac{\mathcal{R}_{in}}{\sqrt{m}}\mathbf{1}_m$. Thus, for any $\lambda \in \mathcal{B}_{\mathcal{R}_{in}}$ we have $\lambda \in \mathbb{R}^m_+$, $\|\lambda\|_1 \leq \|\lambda^{(1)}\|_1 = B_\lambda \Rightarrow \lambda \in \Lambda$. The second inclusion follows from the inequality $\|\cdot\|_1 \leq \|\cdot\|_2$. □

The following lemma bounds the range of $d_\tau(\lambda)$ on Λ.

Lemma C.5. *The dual function $d_\tau(\lambda)$ on the set Λ satisfies*

$$0 \leq d_\tau(\lambda) \leq B_d := \frac{r_{0,max} + \sqrt{m}B_\lambda R_{max} + \tau\log|\mathcal{A}|}{1-\gamma}. \tag{53}$$

Proof.

$$\begin{aligned}
0 \leq d_\tau(\lambda) &= \max_{\pi\in\Pi} V_0^\pi(\rho) + \langle\lambda, V^\pi(\rho) - c\rangle + \tau\mathcal{H}(\pi)\\
&\leq \frac{r_{0,max}}{1-\gamma} + \frac{1}{1-\gamma}\|\lambda\|_2 \cdot R_{max} + \frac{\tau\log|\mathcal{A}|}{1-\gamma}.
\end{aligned}$$

□

Now we establish the fact that the dual function $d_\tau(\lambda)$ is differentiable on Λ, and state what its gradient is.

Proposition C.6. *The function $d_\tau(\lambda)$ is differentiable for all $\lambda \in \Lambda$, and*

$$\nabla d_\tau(\lambda) = V^{\pi^*_{\tau,\lambda}}(\rho) - c, \tag{54}$$

where $\pi^*_{\tau,\lambda} := \arg\max_{\pi\in\Pi}\mathcal{L}_\tau(\pi,\lambda)$.

Proof. We will apply Danskin's theorem from Bertsekas (1991) (Proposition B.25, (a)) to $\mathcal{L}_\tau(\pi,\lambda)$, which is defined on $\Pi \times \mathbb{R}^m$. Note that Π is compact, $\mathcal{L}_\tau(\cdot,\cdot)$ is continuous, and $\mathcal{L}_\tau(\pi,\cdot)$ is linear (and hence convex and differentiable) for all $\pi \in \Pi$. Then, according to Proposition 4.2, for $\lambda \in \Lambda$, we also have that the maximizer for

$$d_\tau(\lambda) = \max_\pi \mathcal{L}_\tau(\pi,\lambda) \tag{55}$$

is unique and equal to $\pi^*_{\tau,\lambda}$. From Danskin's theorem it then follows, that $d_\tau(\lambda)$ is differentiable for all $\lambda \in \Lambda$, and

$$\nabla d_\tau(\lambda) = \nabla_\lambda\mathcal{L}_\tau(\pi^*_{\tau,\lambda},\lambda) = V^{\pi^*_{\tau,\lambda}}(\rho) - c. \tag{56}$$

□

The following two lemmas are required to prove that $d_\tau(\lambda)$ is smooth, that is, its gradient is Lipschitz continuous.

Lemma C.7. *Set any $\tau > 0$. Define the following regularized softmax function for $x \in \mathbb{R}^n$:*

$$S_\tau(x)_i = \frac{\exp(x_i/\tau)}{\sum_{j=1}^n \exp(x_j/\tau)}. \tag{57}$$

Then for any $x, x' \in \mathbb{R}^n$ it holds that:

$$\|S_\tau(x) - S_\tau(x')\|_1 \leq \frac{1}{\tau}\|x - x'\|_\infty. \tag{58}$$

Proof. First notice that $S_\tau = \nabla H_\tau$, where:

$$H_\tau(x) = \tau \log\left(\sum_{i=1}^n e^{x_i/\tau}\right). \tag{59}$$

So we can write:

$$\|S_\tau(x) - S_\tau(x')\|_1 = \|\nabla H_\tau(x) - \nabla H_\tau(x')\|_1 \overset{(i)}{\leq} \sup_{z\in\mathbb{R}^n} \left\|\nabla^2 H_\tau(z)(x' - x)\right\|_1 = \tag{60}$$

$$= \sup_{\substack{z\in\mathbb{R}^n,\\ u\in\mathbb{R}^n, \|u\|_\infty = 1}} u^\top \nabla^2 H_\tau(z)(x' - x) \leq \tag{61}$$

$$\leq \|x' - x\|_\infty \cdot \sup_{\substack{z\in\mathbb{R}^n,\\ u\in\mathbb{R}^n, \|u\|_\infty = 1\\ v\in\mathbb{R}^n, \|v\|_\infty = 1}} u^\top \nabla^2 H_\tau(z) v, \tag{62}$$

where (i) can be obtained by considering a function $W(t) = \nabla H_\tau(x(1-t) + yt)$ and applying Lagrange mean inequality for it with l_1-norm.

Calculate $\nabla^2 H_\tau(z)$:

$$\frac{\partial^2 H}{\partial z_i \partial z_j} = \frac{\frac{1}{\tau}\exp(z_i/\tau)\delta_{ij}}{(\sum_{k=1}^n \exp(z_k)/\tau)} - \frac{\frac{1}{\tau}\exp(x_i/\tau)\exp(z_j/\tau)}{(\sum_{k=1}^n \exp(z_k)/\tau)^2}. \tag{63}$$

Fix some $z \in \mathbb{R}^n$. Let $a_i = \frac{\exp(z_i/\tau)}{\sum_{k=1}^n \exp(z_k/\tau)}$. Note that $\sum_{k=1}^n a_k = 1$, and:

$$\nabla^2 H_\tau(z)_{ij} = \frac{1}{\tau}a_i\delta_{ij} - \frac{1}{\tau}a_i a_j. \tag{64}$$

Under the supremum, knowing $\|u\|_\infty = \|v\|_\infty = 1$, we have:

$$u^\top \nabla^2 H_\tau(z) v = \frac{1}{\tau}\left(\sum_{i=1} a_i u_i v_i - \sum_{i=1}^n \sum_{j=1}^n a_i a_j u_i v_j\right) = \tag{65}$$

$$= \frac{1}{\tau}\left(\sum_{i=1}^n a_i u_i \left(v_i - \sum_{j=1}^n a_j v_j\right)\right) = \frac{1}{\tau}\left(\sum_{i=1}^n a_i u_i \left(\sum_{\substack{j=1\\ j\neq i}}^n a_j (v_i - v_j)\right)\right) \leq \tag{66}$$

$$\leq \frac{1}{\tau}\sum_{i=1}^n \sum_{\substack{j=1\\ j\neq i}}^n a_i a_j |v_i - v_j| \leq \frac{1}{\tau}. \tag{67}$$

Using this in 60, we finally get:

$$\|S_\tau(x) - S_\tau(x')\|_1 \leq \frac{1}{\tau}\|x' - x\|_\infty. \tag{68}$$

□

The next result is a corrected and enhanced version of Lemma 7 from Li et al. (2021) with a bound improved in a factor of two.

Lemma C.8 (Lemma 7 from Li et al. (2021), corrected and enhanced)**.** *The optimal policy for regularized MDP is smooth with respect to λ, i.e., for all $\lambda, \lambda' \in \mathbb{R}_+^m$, we have:*

$$\max_{s\in\mathcal{S}} \left\|\pi^*_{\tau,\lambda}(\cdot|s) - \pi^*_{\tau,\lambda'}(\cdot|s)\right\|_1 \leq \frac{R_{\max}}{(1-\gamma)\tau}\|\lambda-\lambda'\|_2. \tag{69}$$

Proof. Proof goes same as in Li et al. (2021), except we bound the promised l_1-norm on the left hand side, instead of l_∞ norm.

As was proved in Nachum et al. (2017), the regularized optimal policy can be expressed in terms of its soft Q-function:

$$\pi^*_{\tau,\lambda}(a|s) = \frac{\exp\Big(Q^*_{\tau,\lambda}(s,a)/\tau\Big)}{\sum_{a'\in\mathcal{A}}\exp\Big(Q^*_{\tau,\lambda}(s,a')/\tau\Big)}. \tag{70}$$

Fix some $s \in \mathcal{S}$. Using C.7 for $x = Q^*_{\tau,\lambda}(s|\cdot), x' = Q^*_{\tau,\lambda'}(s|\cdot)$ and τ, we get:

$$\left\|\pi^*_{\tau,\lambda}(\cdot|s) - \pi^*_{\tau,\lambda'}(\cdot|s)\right\|_1 \leq \frac{1}{\tau}\left\|Q^*_{\tau,\lambda}(s|\cdot) - Q^*_{\tau,\lambda'}(s|\cdot)\right\|_\infty \leq \frac{1}{\tau}\left\|Q^*_{\tau,\lambda} - Q^*_{\tau,\lambda'}\right\|_\infty. \tag{71}$$

Furthermore,

$$\left\|Q^*_{\tau,\lambda} - Q^*_{\tau,\lambda'}\right\|_\infty \leq \max_{s\in\mathcal{S},a\in\mathcal{A}} \left\|Q^*_{\tau,\lambda}(s,a) - Q^*_{\tau,\lambda'}(s,a)\right\|_\infty \leq \tag{72}$$

$$\leq \max_{s\in\mathcal{S},a\in\mathcal{A}}\max_{\pi\in\Pi}\left|Q^\pi_{\tau,\lambda}(s,a) - Q^\pi_{\tau,\lambda'}(s,a)\right| \overset{(i)}{\leq} \tag{73}$$

$$\leq \frac{R_{\max}}{1-\gamma}\|\lambda-\lambda'\|_2, \tag{74}$$

where (i) is due to

$$\left|Q^\pi_{\tau,\lambda}(s,a) - Q^\pi_{\tau,\lambda'}(s,a)\right| \leq |r_\lambda(s,a) - r_{\lambda'}(s,a)| + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\left|V^\pi_{\tau,\lambda}(s') - V^\pi_{\tau,\lambda'}(s')\right|\right] \leq \tag{75}$$

$$\leq R_{\max}\|\lambda-\lambda'\|_2 + \gamma\cdot\frac{1}{1-\gamma}R_{\max}\|\lambda-\lambda'\|_2 = \tag{76}$$

$$= \frac{R_{\max}\|\lambda-\lambda'\|_2}{1-\gamma}. \tag{77}$$

Substituting 74 into 71, we get the desired result. □

The next proposition specifies the smoothness coefficient for d_τ.

Proposition C.9. *Suppose assumption 4.3 holds, then $d_\tau(\lambda)$ is L_d-smooth:*

$$\|\nabla d_\tau(\lambda) - \nabla d_\tau(\lambda')\|_2 \leq L_d\|\lambda-\lambda'\|_2, \quad \forall\lambda,\lambda'\in\Lambda,$$

where $L_d = \frac{R^2_{\max}L_\beta}{(1-\gamma)^2\tau}$, $L_\beta := \left\lceil\log_\beta\left(C_M^{-1}\right)\right\rceil + (1-\beta)^{-1} + 1$.

Proof. See proof in Li et al. (2021) (Proposition 1), with $\mu = 0$. Instead of Lemma 7 in it, our version C.8 can be used, making sure the needed inequality is correct and improving L_d by a factor of 2. □

The following two lemmas provide a bound on optimality gap and constraint violation in terms of the dual function.

Lemma C.10. *Let $\lambda \in \Lambda$, then*

$$V_0^*(\rho) - V_0^{\pi^*_{\tau,\lambda}}(\rho) \leq \left\langle\lambda, \nabla d_\tau(\lambda)\right\rangle + \tau\mathcal{H}\left(\pi^*_{\tau,\lambda}\right), \tag{78}$$

$$\left\|[c - V^{\pi^*_{\tau,\lambda}}(\rho)]_+\right\|_2 = \left\|[-\nabla d_\tau(\lambda)]_+\right\|_2. \tag{79}$$

Proof.

$$V_0^{\pi^*_{\tau,\lambda}}(\rho) + \big\langle \lambda, V^{\pi^*_{\tau,\lambda}}(\rho) - c\big\rangle + \tau\mathcal{H}(\pi^*_{\tau,\lambda}) = \max_{\pi\in\Pi} \mathcal{L}_\tau(\pi,\lambda)$$
$$\geq \mathcal{L}_\tau(\pi^*,\lambda) = V_0^*(\rho) + \big\langle \lambda, V^{\pi^*}(\rho) - c\big\rangle + \tau\mathcal{H}(\pi^*).$$

The inequalities $\mathcal{H}(\pi^*) \geq 0$, $V^{\pi^*}(\rho) \geq c$, $\lambda \in \mathbb{R}^m_+$ imply $\mathcal{L}_\tau(\pi^*,\lambda) \geq V_0^*(\rho)$, hence

$$V_0^*(\rho) - V_0^{\pi^*_{\tau,\lambda}}(\rho) \leq \big\langle \lambda, V^{\pi^*_{\tau,\lambda}}(\rho) - c\big\rangle + \tau\mathcal{H}(\pi^*_{\tau,\lambda}).$$

Using the result of Proposition C.6, we get

$$\nabla d_\tau(\lambda) = V^{\pi^*_{\tau,\lambda}}(\rho) - c,$$

which finishes the proof. □

Lemma C.11. *Let $\lambda \in \Lambda$, then*

$$\big\|[-\nabla d_\tau(\lambda)]_+\big\|_2^2 \leq 2L_d(d_\tau(\lambda) - d^*_\tau), \tag{80}$$
$$\big\langle \lambda, \nabla d_\tau(\lambda)\big\rangle \leq B_\lambda\sqrt{2mL_d(d_\tau(\lambda) - d^*_\tau)} + 2(d_\tau(\lambda) - d^*_\tau), \tag{81}$$

where d^*_τ is the optimal value in the dual problem (51)*, L_d is the smoothness constant of d_τ.*

Proof. Denote $a := \nabla d_\tau(\lambda)$ and

$$N := \{1,\dots,n\},\ I := \{i \in N :\ a_i \geq 0\},\ I' := N \setminus I.$$

Moreover, for any vector $b \in \mathbb{R}^m$, define b_I to be the vector with components

$$(b_I)_i := \begin{cases} b_i & \text{if } i \in I,\\ 0 & \text{otherwise.}\end{cases}$$

Smoothness implies for any $\lambda' \in \Lambda$

$$\langle a, \lambda - \lambda'\rangle - \frac{L_d}{2}\|\lambda - \lambda'\|_2^2 \leq d_\tau(\lambda) - d_\tau(\lambda'). \tag{82}$$

Pick $\lambda' := \left[\lambda - \frac{1}{L_d}a\right]_+$, then it's sufficient to prove that

$$\frac{1}{2L_d}\big\|[-a]_+\big\|_2^2 \leq \langle a, \lambda - \lambda'\rangle - \frac{L_d}{2}\|\lambda - \lambda'\|_2^2, \tag{83}$$

and the first result of the lemma will follow from (82). Using the notation introduced above, we write

$$a = a_I + a_{I'} \text{ with } a_I, (-a_{I'}) \in \mathbb{R}^m_+, \qquad \lambda = \lambda_I + \lambda_{I'}.$$

The vector λ' can now be expressed as

$$\lambda' = \left[\lambda_I + \lambda_{I'} - \tfrac{1}{L_d}a_I - \tfrac{1}{L_d}a_{I'}\right]_+ = \left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+ + \lambda_{I'} - \frac{1}{L_d}a_{I'},$$

because $I \cap I' = \emptyset$ and $\lambda_{I'} - \frac{1}{L_d}a_{I'} \in \mathbb{R}^m_+$. The value $\lambda - \lambda'$ writes as

$$\lambda - \lambda' = \lambda_I - \left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+ + \frac{1}{L_d}a_{I'}.$$

The two terms in the right-hand side of (83) are equal to

$$\langle a, \lambda - \lambda'\rangle = \left\langle a_I, \lambda_I - \left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+\right\rangle + \frac{1}{L_d}\|a_{I'}\|_2^2,$$

and

$$\frac{L_d}{2}\|\lambda-\lambda'\|_2^2 = \frac{L_d}{2}\left\|\lambda_I - \left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+\right\|_2^2 + \frac{1}{2L_d}\|a_{I'}\|_2^2.$$

Thus, the right-hand side of (83) writes as

$$\frac{1}{2L_d}\|a_{I'}\|_2^2 + \underbrace{\left\langle a_I, \lambda_I - \left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+\right\rangle - \frac{L_d}{2}\left\|\lambda_I - \left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+\right\|_2^2}_{=:\psi}.$$

To observe that $\psi \geq 0$, consider the sets

$$J := \left\{i \in I : \lambda_i \geq \tfrac{1}{L_d}a_i\right\},\ J' := I \setminus J,$$

then

$$\left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+ = \lambda_J - \frac{1}{L_d}a_J \quad \text{and} \quad \lambda_I - \left[\lambda_I - \tfrac{1}{L_d}a_I\right]_+ = \lambda_{J'} + \frac{1}{L_d}a_J.$$

Thus,

$$\psi = \left\langle a_I - \tfrac{L_d}{2}\left(\lambda_{J'} + \tfrac{1}{L_d}a_J\right),\, \lambda_{J'} + \tfrac{1}{L_d}a_J\right\rangle = \frac{1}{2}\left\langle a_I + L_d\left(\tfrac{1}{L_d}a_{J'} - \lambda_{J'}\right),\, \lambda_{J'} + \tfrac{1}{L_d}a_J\right\rangle,$$

which is a nonnegative value as a scalar product of vectors with nonnegative components. To sum up,

$$\langle a, \lambda - \lambda'\rangle - \frac{L_d}{2}\|\lambda' - \lambda\|_2^2 = \frac{1}{2L_d}\|a_{I'}\|_2^2 + \psi \geq \frac{1}{2L_d}\|a_{I'}\|_2^2,$$

and the first result of the lemma follows from (82) since $d_\tau(\lambda') \geq d_\tau^*$.

The left-hand side of the inequality (81) equals

$$\langle \lambda, a\rangle = \langle \lambda_J, a_J\rangle + \langle \lambda_{J'}, a_{J'}\rangle + \langle \lambda_{I'}, a_{I'}\rangle. \tag{84}$$

The last term is non-positive. Let us bound the first two. Put $\lambda' := \lambda - \frac{1}{L_d}a_J$ and observe that $\lambda' \in \mathbb{R}^m_+$ due to the definition of J. Moreover, $\lambda' \in \Lambda$ since $\|\lambda'\|_1 \leq \|\lambda\|_1$. The right-hand side of (82) writes as

$$\langle a, \lambda - \lambda'\rangle - \frac{L_d}{2}\|\lambda - \lambda'\|_2^2 = \left\langle a, \tfrac{1}{L_d}a_J\right\rangle - \frac{1}{2L_d}\|a_J\|_2^2 = \frac{1}{2L_d}\|a_J\|_2^2. \tag{85}$$

Relations (82) and (85) yield

$$\|a_J\|_2 \leq \sqrt{2L_d(d_\tau(\lambda) - d_\tau(\lambda'))} \leq \sqrt{2L_d(d_\tau(\lambda) - d_\tau^*)}. \tag{86}$$

Cauchy–Bunyakovsky–Schwarz inequality, condition $\|\lambda\|_2 \leq \sqrt{m}\|\lambda\|_1 \leq \sqrt{m}B_\lambda$ and bound (86) imply

$$\langle \lambda_J, a_J\rangle \leq \|\lambda_J\|_2 \cdot \|a_J\|_2 \leq B_\lambda\sqrt{2mL_d(d_\tau(\lambda) - d_\tau^*)}. \tag{87}$$

The bound for the second term in the right-hand side of (84) can be obtained by putting $\lambda' := \lambda - \lambda_{J'}$. Indeed,

$$\begin{aligned}\langle a, \lambda - \lambda'\rangle - \frac{L_d}{2}\|\lambda - \lambda'\|_2^2 = \langle a, \lambda_{J'}\rangle - \frac{L_d}{2}\|\lambda_{J'}\|_2^2 &= \frac{1}{2}\langle a_{J'}, \lambda_{J'}\rangle + \frac{L_d}{2}\left(\frac{1}{L_d}\langle a_{J'}, \lambda_{J'}\rangle - \|\lambda_{J'}\|_2^2\right)\\ &= \frac{1}{2}\langle a_{J'}, \lambda_{J'}\rangle + \frac{L_d}{2}\left\langle \tfrac{1}{L_d}a_{J'} - \lambda_{J'}, \lambda_{J'}\right\rangle \geq \frac{1}{2}\langle a_{J'}, \lambda_{J'}\rangle,\end{aligned}$$

where the last inequality is due to the definition of J'. Thus,

$$\langle a_{J'}, \lambda_{J'}\rangle \leq 2(d_\tau(\lambda) - d_\tau^*), \tag{88}$$

and the second result of the Lemma follows from (84), (87) and (88).

□

D PROOF OF THEOREM 4.4

In this section, we will write for the shortness of notation $V_0^\pi := V_0^\pi(\rho), V^\pi := V^\pi(\rho)$, and so on, still taking the dependence on ρ into account. Recall that π_T is the output of the Algorithm 1 after T iterations, V_0^* and d_τ^* are optimal values in the primal (2) and dual (5) problems, respectively, and the following notation is used

$$\begin{aligned} V_{\tau,\lambda}^\pi &:= V_0^\pi + \sum_{i=1}^m \lambda_i V_i^\pi + \tau \mathcal{H}(\pi), \\ V_{\tau,\lambda}^* &:= \max_{\pi\in\Pi} V_{\tau,\lambda}^\pi, \\ \pi_{\tau,\lambda}^* &:= \arg\max_{\pi\in\Pi} \mathcal{L}_\tau(\pi,\lambda) \equiv \arg\max_{\pi\in\Pi} V_{\tau,\lambda}^\pi. \end{aligned} \tag{89}$$

Plan of the proof is as follows.

1. We describe how to apply Theorem B.2 about convergence of Vaidya's method with δ-subgradient (Algorithm 3) to our proposed Algorithm 1 for the dual problem.
2. We express the values $V_0^* - V_0^{\pi_{\tau,\lambda_T}^*}$ and $\left\|[c - V^{\pi_{\tau,\lambda_T}^*}]_+\right\|_2$ through the optimality gap of the dual problem $d_\tau(\lambda_T) - d_\tau^*$. That is, we estimate the optimality gap and constraint violation as if the NPG could solve the problem (89) exactly for the last iterate λ_T.
3. We estimate the values $V_0^* - V_0^{\pi_T}$ and $\left\|[c - V^{\pi_T}]_+\right\|_2$ using the results from the previous step and the convergence rate of Vaidya's algorithm.

To begin part 1 of the proof, observe that the proposed Algorithm 1 is a special case of Vaidya's method with δ-subgradient (Algorithm 3) if the value $-\widehat{\nabla}_t := V^{\pi_t} - c$ from line 11 of Algorithm 1 is a $\tilde{\delta}$-subgradient (Definition B.1) for some $\tilde{\delta} > 0$ which depends on the parameter δ of NPG. This is the case due to the lemma from page 132 of Polyak (1987) which we give below keeping notation consistent with the rest of the paper.

Lemma D.1. *Let*

$$d_\tau(\lambda) := \max_{\pi\in\Pi} \mathcal{L}_\tau(\pi,\lambda),$$

where Π is a compact set, $\mathcal{L}_\tau(\pi,\lambda)$ is continuous in π,λ and convex in λ. Let $\tilde{\pi}$ satisfy for a fixed λ the inequality $d(\lambda) - \mathcal{L}_\tau(\pi,\lambda) \le \tilde{\delta}$, then $\partial_\lambda \mathcal{L}_\tau(\hat{\pi},\lambda) \in \partial_{\tilde{\delta}} d_\tau(\lambda)$.

The given lemma shows that $\tilde{\delta}$-optimal policy (in terms of optimality gap of regularized Lagrangian $\mathcal{L}_\tau$) gives us a $\tilde{\delta}$-subgradient for the dual function $d_\tau(\lambda)$. According to Theorem A.2, the call $\mathrm{NPG}\,(r_0 + \langle\lambda_t, r\rangle, \tau, \delta)$ in line 10 of Algorithm 1 ensures accuracy $\tilde{\delta} = 6\tau\gamma\delta$. Additionally, note that the vector $\widehat{\nabla}_t$ from line 13 of the Algorithm 1 satisfies the inequality from line 12 of the Algorithm 3. Indeed, the value $\widehat{\nabla}_t^\top \lambda_t$ is negative as the sum of negative components of λ_t, while $\widehat{\nabla}_t^\top \lambda$ is nonnegative for all $\lambda \in \Lambda$ as a scalar product of two vectors with nonnegative elements.

Before we can apply Theorem B.2 to the proposed algorithm, we need to replace the dual problem $\min_{\lambda\in\mathbb{R}_+^m} d_\tau(\lambda)$ with the equivalent one, but on a compact set. This is possible due to Lemma C.3 which states that the solution λ_τ^* of the dual problem satisfies

$$\|\lambda_\tau^*\|_1 \le B_\lambda := \frac{r_{0,\max} + \log|\mathcal{A}|}{(1-\gamma)\xi}. \tag{90}$$

Thus, the equivalent formulation is

$$\min_{\lambda\in\Lambda} d_\tau(\lambda) \quad \text{with } \Lambda := \{\lambda \in \mathbb{R}_+^m \mid \|\lambda\|_1 \le B_\lambda\}. \tag{91}$$

The only thing left to do is to find the values $\mathcal{R}, \mathcal{R}_{in}$ and B_d such that $\mathcal{B}_{\mathcal{R}_{in}} \subseteq \Lambda \subseteq \mathcal{B}_{\mathcal{R}}$ and $|d_\tau(\lambda) - d_\tau(\lambda')| \le B_d \; \forall \lambda, \lambda' \in \Lambda$. Such values are given by Lemmas C.4 and C.5:

$$\mathcal{R} := B_\lambda, \tag{92}$$

$$\mathcal{R}_{in} := \frac{B_\lambda}{m + \sqrt{m}}, \tag{93}$$

$$B_d := \frac{r_{0,max} + \sqrt{m} B_\lambda R_{max} + \tau \log|\mathcal{A}|}{1-\gamma} \overset{(90)}{\le} \left(\xi + \frac{\sqrt{m} R_{max}}{1-\gamma}\right) B_\lambda. \tag{94}$$

We can put

$$A_0 := \begin{bmatrix} -I_m \\ 1 \end{bmatrix}, \quad b_0 := \begin{bmatrix} B_\lambda \mathbf{1}_m \\ m B_\lambda \end{bmatrix},$$

which will correspond to the initial simplex

$$P(A_0, b_0) = \left\{ \lambda \in \mathbb{R}^m : \lambda_j \geqslant -\mathcal{R}, j = 1, \ldots, m, \sum_{j=1}^{m} \lambda_j \leqslant m\mathcal{R} \right\} \supseteq \mathcal{B}_{\mathcal{R}}.$$

Thus, Theorem B.2 applied to the proposed algorithm yields the following convergence estimate for the dual problem (91):

$$\begin{aligned} d_\tau(\lambda_T) - d_\tau^* &\le \frac{m^2(1+\sqrt{m})B_\lambda}{\zeta} \left(\xi + \frac{\sqrt{m}R_{max}}{1-\gamma} \right) \exp\left(\frac{\log \boldsymbol{\pi} - \zeta T}{2m} \right) + 6\tau\gamma\delta && (95) \\ &=: \epsilon + 6\tau\gamma\delta, && (96) \end{aligned}$$

where ϵ denotes the first term of the estimate (95).

Part 2 of the proof goes as follows. First, we use Proposition C.9 to state that $d_\tau(\lambda)$ is L_d-smooth with

$$L_d = \frac{R_{\max}^2 L_\beta}{(1-\gamma)^2 \tau}, \quad \text{where } L_\beta := \left\lceil \log_\beta \left(C_M^{-1} \right) \right\rceil + (1-\beta)^{-1} + 1. \tag{97}$$

Second, we refer to Lemmas C.10 and C.11 which provide the following bounds:

$$\begin{aligned} V_0^* - V_0^{\pi^*_{\tau,\lambda_T}} &\le \left\langle \lambda_T, \nabla d_\tau(\lambda_T) \right\rangle + \tau \mathcal{H}(\pi^*_{\tau,\lambda_T}) && (98) \\ &\le B_\lambda \sqrt{2mL_d(\epsilon + 6\tau\gamma\delta)} + 2(\epsilon + 6\tau\gamma\delta) + \tau\mathcal{H}(\pi^*_{\tau,\lambda_T}), && (99) \\ \left\| [c - V^{\pi^*_{\tau,\lambda_T}}]_+ \right\|_2 = \left\| [-\nabla d_\tau(\lambda_T)]_+ \right\|_2 &\le 2L_d(\epsilon + 6\tau\gamma\delta). && (100) \end{aligned}$$

Let us begin part 3 of the proof. First, we bound the value $V_0^{\pi^*_{\tau,\lambda_T}} - V_0^{\pi_T}$. Recall that $\pi_T := \mathrm{NPG}\left(r_0 + \langle \lambda_T, r \rangle, \tau, \delta\right)$. Therefore, according to Theorem A.2, π_T satisfies

$$V^*_{\tau,\lambda_T} - V^{\pi_T}_{\tau,\lambda_T} \le 6\tau\gamma\delta. \tag{101}$$

Furthermore,

$$\begin{aligned} V_0^{\pi^*_{\tau,\lambda_T}} - V_0^{\pi_T} &= \underbrace{V_0^{\pi^*_{\tau,\lambda_T}} + \left\langle \lambda_T, V^{\pi^*_{\tau,\lambda_T}} \right\rangle + \tau\mathcal{H}(\pi^*_{\tau,\lambda_T})}_{V^*_{\tau,\lambda_T}} - \underbrace{\left(V_0^{\pi_T} + \left\langle \lambda_T, V^{\pi_T} \right\rangle + \tau\mathcal{H}(\pi_T) \right)}_{V^{\pi_T}_{\tau,\lambda_T}} \\ &\quad + \left\langle \lambda_T, V^{\pi_T} - V^{\pi^*_{\tau,\lambda_T}} \right\rangle + \tau \left(\mathcal{H}(\pi_T) - \mathcal{H}(\pi^*_{\tau,\lambda_T}) \right) \\ &\overset{(101)}{\le} 6\tau\gamma\delta + \left\langle \lambda_T, V^{\pi_T} - V^{\pi^*_{\tau,\lambda_T}} \right\rangle + \tau \left(\mathcal{H}(\pi_T) - \mathcal{H}(\pi^*_{\tau,\lambda_T}) \right). && (102) \end{aligned}$$

The scalar product is bounded by

$$\left\langle \lambda_T, V^{\pi_T} - V^{\pi^*_{\tau,\lambda_T}} \right\rangle \le \|\lambda_T\|_2 \cdot \left\| V^{\pi_T} - V^{\pi^*_{\tau,\lambda_T}} \right\|_2 \le \sqrt{m} B_\lambda \|\hat{\delta}\|_2.$$

where $\hat{\delta} := V^{\pi_T}(\rho) - V^{\pi^*_{\tau,\lambda_T}}(\rho)$ is a value controlled by the NPG parameter δ. Now, the optimality gap can be estimated as follows:

$$\begin{aligned} V_0^* - V_0^{\pi_T} &= V_0^* - V_0^{\pi^*_{\tau,\lambda_T}} + V_0^{\pi^*_{\tau,\lambda_T}} - V_0^{\pi_T} \\ &\overset{(99)}{\le} B_\lambda \sqrt{2mL_d(\epsilon + 6\tau\gamma\delta)} + 2(\epsilon + 6\tau\gamma\delta) + \tau\mathcal{H}(\pi^*_{\tau,\lambda_T}) + V_0^{\pi^*_{\tau,\lambda_T}} - V_0^{\pi_T} \\ &\overset{(102)}{\le} B_\lambda \sqrt{2mL_d(\epsilon + 6\tau\gamma\delta)} + 2(\epsilon + 9\tau\gamma\delta) + \tau\mathcal{H}(\pi_T) + \sqrt{m}B_\lambda \|\hat{\delta}\|_2 \\ &\overset{(97)}{=} \frac{B_\lambda R_{max} \sqrt{2mL_\beta}}{1-\gamma} \sqrt{\epsilon/\tau + 6\gamma\delta} + 2(\epsilon + 9\tau\gamma\delta) + \tau\mathcal{H}(\pi_T) + \sqrt{m}B_\lambda \|\hat{\delta}\|_2. \end{aligned}$$

Note that $\mathcal{H}(\pi_T) \leq \frac{\log|\mathcal{A}|}{1-\gamma}$. It is reasonable to balance the terms that are proportional to $\sqrt{\epsilon/\tau}$ and τ by taking $\tau := \min(1, \sqrt[3]{\epsilon})$.

Also, we can bound $\left\|\hat{\delta}\right\|_2$ via knowing that NPG approximated π^*_{τ,λ_T} by π_T with accuracy δ, as stated in A.2:

$$\left\|\pi_T - \pi^*_{\tau,\lambda_T}\right\|_\infty < \delta. \tag{103}$$

By Lemma 6 from Li et al. (2021), which can be proved for any two policies, we have:

$$\left\|\nu_\rho^{\pi_T} - \nu_\rho^{\pi^*_{\tau,\lambda_T}}\right\|_1 \leq L_\beta \max_{s\in\mathcal{S}} \left\|\pi_T(\cdot|s) - \pi^*_{\tau,\lambda_T}(\cdot|s)\right\|_1 \leq L_\beta|\mathcal{A}|\delta. \tag{104}$$

Now, we can rewrite $\|\hat{\delta}\|_2$:

$$\|\hat{\delta}\|_2^2 = \left\|V^{\pi_T}(\rho) - V^{\pi^*_{\tau,\lambda_T}}(\rho)\right\|_2^2 = \sum_{i=1}^m \left(\frac{1}{1-\gamma}\langle r_i, \nu_\rho^{\pi_T}\rangle - \frac{1}{1-\gamma}\langle r_i, \nu_\rho^{\pi^*_{\tau,\lambda_T}}\rangle\right)^2 \leq \tag{105}$$

$$\leq \frac{1}{(1-\gamma)^2}\sum_{i=1}^m \langle r_i, \nu_\rho^{\pi_T} - \nu_\rho^{\pi^*_{\tau,\lambda_T}}\rangle^2 \leq \frac{1}{(1-\gamma)^2}\sum_{i=1}^m \left(r_{i,\max}\left\|\nu_\rho^{\pi_T} - \nu_\rho^{\pi^*_{\tau,\lambda_T}}\right\|_1\right)^2 = \tag{106}$$

$$= \frac{\left\|\nu_\rho^{\pi_T} - \nu_\rho^{\pi^*_{\tau,\lambda_T}}\right\|_1^2}{(1-\gamma)^2} R_{\max}^2. \tag{107}$$

So, we get:

$$\|\hat{\delta}\|_2 \leq \frac{\left\|\nu_\rho^{\pi_T} - \nu_\rho^{\pi^*_{\tau,\lambda_T}}\right\|_1}{(1-\gamma)} R_{\max} \leq \frac{L_\beta|\mathcal{A}|R_{\max}}{(1-\gamma)}\delta. \tag{108}$$

Plugging this estimate and the choice of $\tau = \min(1, \sqrt[3]{\epsilon})$. into (97), we get the desired result for optimality gap 13.

The second part of the result, 15, we achieve as follows:

$$\begin{aligned}
c - V^{\pi_T} &= c - V^{\pi^*_{\tau,\lambda_T}} + V^{\pi^*_{\tau,\lambda_T}} - V^{\pi_T}, \\
&\leq \left\|[c - V^{\pi^*_{\tau,\lambda_T}}]_+\right\|_2 + \left\|V^{\pi_T} - V^{\pi^*_{\tau,\lambda_T}}\right\|_2 \\
&\overset{(100)}{\leq} 2L_d(\epsilon + 6\tau\gamma\delta) + \|\hat{\delta}\|_2 \\
&\overset{(97)}{=} \frac{2R_{max}L_\beta}{1-\gamma}(\epsilon^{2/3} + 6\gamma\delta) + \|\hat{\delta}\|_2 \leq \\
&\leq \frac{2R_{max}L_\beta}{1-\gamma}(\epsilon^{2/3} + 6\gamma\delta) + \frac{L_\beta|\mathcal{A}|R_{\max}}{(1-\gamma)}\delta.
\end{aligned}$$

where $\|\hat{\delta}\|_2$ is bounded in the same way as earlier.

Having achieved both bounds 13, 15, we conclude the proof.

D.1 Proof of Corollary 4.5

Suppose we need the resulting accuracy to be $\kappa > 0$:

$$V_0^*(\rho) - V_0^{\pi_T}(\rho) \leq \kappa, \tag{109}$$

$$\left\|[c - V^{\pi_T}(\rho)]_+\right\|_2 \leq \kappa. \tag{110}$$

We will find some T, δ to use for the algorithm, so that by setting other parameters as in 4.4, we will get κ-optimal solution by its results. It is enough to satisfy these inequalities:

$$\frac{B_\lambda R_{max}\sqrt{2mL_\beta}}{1-\gamma}\sqrt{\epsilon^{2/3}+6\gamma\delta} < \kappa/5, \tag{111}$$

$$2\epsilon < \kappa/5, \tag{112}$$

$$18\gamma\delta\sqrt[3]{\epsilon} < \kappa/5, \tag{113}$$

$$\frac{\log|\mathcal{A}|}{1-\gamma}\sqrt[3]{\epsilon} < \kappa/5, \tag{114}$$

$$\sqrt{m}B_\lambda\frac{L_\beta|\mathcal{A}|R_{\max}}{1-\gamma}\delta < \kappa/5, \tag{115}$$

$$\frac{2R_{max}^2L_\beta}{1-\gamma}(\epsilon^{2/3}+6\gamma\delta) < \kappa/2, \tag{116}$$

$$\frac{L_\beta|\mathcal{A}|R_{\max}}{(1-\gamma)}\delta < \kappa/2. \tag{117}$$

To satisfy them, it is enough to satisfy these:

$$\epsilon^{2/3}+6\gamma\delta < \frac{(1-\gamma)^2\kappa^2}{50B_\lambda^2R_{max}^2mL_\beta}, \tag{118}$$

$$\epsilon < \kappa/10, \tag{119}$$

$$\delta\sqrt[3]{\epsilon} < \kappa/(90\gamma), \tag{120}$$

$$\epsilon < \frac{\kappa^3(1-\gamma)^3}{125\log^3|\mathcal{A}|}, \tag{121}$$

$$\delta < \frac{\kappa(1-\gamma)}{5\sqrt{m}B_\lambda L_\beta|\mathcal{A}|R_{\max}}, \tag{122}$$

$$\epsilon^{2/3}+6\gamma\delta < \frac{\kappa(1-\gamma)}{4R_{max}^2L_\beta}, \tag{123}$$

$$\delta < \frac{(1-\gamma)\kappa}{2L_\beta|\mathcal{A}|R_{\max}}. \tag{124}$$

To satisfy them, it is enough to satisfy these:

$$\epsilon < \frac{(1-\gamma)^3\kappa^3}{1000B_\lambda^3R_{max}^3m^{3/2}L_\beta^{3/2}}, \tag{125}$$

$$\delta < \frac{(1-\gamma)^2\kappa^2}{600\gamma B_\lambda^2R_{max}^2mL_\beta}, \tag{126}$$

$$\epsilon < \kappa/10, \tag{127}$$

$$\delta < \kappa^{2/3}/(90\gamma)^{2/3}, \tag{128}$$

$$\epsilon < \kappa/(90\gamma), \tag{129}$$

$$\epsilon < \frac{\kappa^3(1-\gamma)^3}{125\log^3|\mathcal{A}|}, \tag{130}$$

$$\delta < \frac{\kappa(1-\gamma)}{5\sqrt{m}B_\lambda L_\beta|\mathcal{A}|R_{\max}}, \tag{131}$$

$$\epsilon < \frac{\kappa^{3/2}(1-\gamma)^{3/2}}{16\sqrt{2}R_{max}^3L_\beta^{3/2}}, \tag{132}$$

$$\delta < \frac{\kappa(1-\gamma)}{48\gamma R_{max}^2L_\beta}, \tag{133}$$

$$\delta < \frac{(1-\gamma)\kappa}{2L_\beta|\mathcal{A}|R_{\max}}. \tag{134}$$

Or, rewritten shorter,

$$\epsilon < \min\left(\frac{\kappa^3(1-\gamma)^3}{1000B_\lambda^3R_{max}^3m^{3/2}L_\beta^{3/2}}, \frac{\kappa}{10}, \frac{\kappa}{90\gamma}, \frac{\kappa^3(1-\gamma)^3}{125\log^3|\mathcal{A}|}, \frac{\kappa^{3/2}(1-\gamma)^{3/2}}{16\sqrt{2}R_{max}^3L_\beta^{3/2}}\right) =: C_\epsilon, \tag{135}$$

$$\delta < \min\left(\frac{\kappa^2(1-\gamma)^2}{600\gamma B_\lambda^2R_{max}^2mL_\beta}, \frac{\kappa^{2/3}}{(90\gamma)^{2/3}}, \frac{\kappa(1-\gamma)}{5\sqrt{m}B_\lambda L_\beta|\mathcal{A}|R_{\max}}, \frac{\kappa(1-\gamma)}{48\gamma R_{max}^2L_\beta}, \frac{\kappa(1-\gamma)}{2L_\beta|\mathcal{A}|R_{\max}}\right) =: C_\delta. \tag{136}$$

Now, knowing that ϵ depends on T as in 11, we need to choose T, such that $\epsilon < C_\epsilon$:

$$\epsilon = \frac{2m^2B_\lambda}{\zeta}\left(\xi + \frac{\sqrt{m}R_{max}}{1-\gamma}\right)\exp\left(\frac{\log\boldsymbol{\pi} - \zeta T}{2m}\right) < C_\epsilon, \tag{137}$$

$$\exp\left(\frac{\log\boldsymbol{\pi} - \zeta T}{2m}\right) < \frac{\zeta C_\epsilon}{2m^2B_\lambda\left(\xi + \frac{\sqrt{m}R_{max}}{1-\gamma}\right)}, \tag{138}$$

$$\frac{\log\boldsymbol{\pi} - \zeta T}{2m} < \log\left(\frac{\zeta C_\epsilon}{2m^2B_\lambda\left(\xi + \frac{\sqrt{m}R_{max}}{1-\gamma}\right)}\right), \tag{139}$$

$$T > \frac{\log\boldsymbol{\pi}}{\zeta} + \frac{2m}{\zeta}\log\left(\frac{2m^2B_\lambda\left(\xi + \frac{\sqrt{m}R_{max}}{1-\gamma}\right)}{\zeta C_\epsilon}\right). \tag{140}$$

And a sufficient number of NPG iterations needed to achieve C_δ accuracy of each NPG call can be determined using A.2:

$$N_{NPG} \approx \frac{\log 2C_1R + \log C_\delta^{-1} + \max(\frac{1}{3}\log C_\epsilon^{-1}, 0)}{\log\gamma^{-1}}, \tag{141}$$

where $C_1 \geq \left\|Q_\tau^*(\rho) - Q_\tau^{(0)}(\rho)\right\|_\infty$, $R \geq \max_{s,a} r(s,a)$ for any MDP on which NPG might be called throughout execution of the algorithm.

It can be seen that asymptotic is

$$T = \mathcal{O}\left(\frac{m}{\zeta}\log\left(\frac{m\log|\mathcal{A}|}{\zeta\xi(1-\gamma)(1-\beta)\kappa}\right)\right), \tag{142}$$

$$N_{NPG} = \mathcal{O}\left(\frac{1}{1-\gamma}\log\left(\frac{m\log|\mathcal{A}|}{(1-\gamma)\xi(1-\beta)\kappa}\right)\right), \tag{143}$$

which gives us the result (accuracy κ is ϵ in the statement).

E LEMMAS FOR THE CASE OF REGULARIZED DUAL VARIABLES

Consider the regularized dual problem:

$$\max_{\lambda\in\mathbb{R}_+^m} d_{\tau,\mu}(\lambda) := d_\tau(\lambda) + \frac{\mu}{2}\|\lambda\|_2^2. \tag{144}$$

The objective is $L_{d,\mu}$-smooth with $L_{d,\mu} := L_d + \mu$. The proof of the convergence mimics the proof from Appendix D but with a better bounds on the value $\langle\nabla d_\tau(\lambda), \lambda\rangle$ and on the norm of dual variable derived below.

Lemma E.1. *It holds*

$$\frac{1}{2L_{d,\mu}}\sum_{i:\lambda_i > \frac{1}{L_{d,\mu}}\frac{\partial d_{\tau,\mu}}{\partial\lambda_i}}\left(\frac{\partial d_{\tau,\mu}}{\partial\lambda_i}\right)^2 + \frac{1}{2}\sum_{i:\lambda_i \leq \frac{1}{L_{d,\mu}}\frac{\partial d_{\tau,\mu}}{\partial\lambda_i}}\frac{\partial d_{\tau,\mu}}{\partial\lambda_i}\lambda_i \leq d_{\tau,\mu}(\lambda) - d_{\tau,\mu}(\lambda_*).$$

Proof. Define $\widetilde{\lambda} := \left[\lambda - \frac{1}{L_{d,\mu}}\nabla d_{\tau,\mu}(\lambda)\right]_+$. Since $d_{\tau,\mu}$ has a Lipschitz continuous gradient on Λ, the following implications hold:

$$\begin{aligned}
d_{\tau,\mu}(\widetilde{\lambda}) &\leq d_{\tau,\mu}(\lambda) + \left\langle \nabla d_{\tau,\mu}, \widetilde{\lambda} - \lambda \right\rangle + \frac{L_{d,\mu}}{2} \left\| \widetilde{\lambda} - \lambda \right\|_2^2 = \\
&= d_{\tau,\mu}(\lambda) - \sum_{i:\lambda_i > \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i} - \sum_{i:\lambda_i \leq \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \lambda_i + \\
&\quad + \sum_{i:\lambda_i > \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \frac{L_{d,\mu}}{2} \left(\frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \right)^2 + \sum_{i:\lambda_i \leq \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \frac{L_{d,\mu}}{2} \lambda_i^2 \qquad (145) \\
&\leq d_{\tau,\mu}(\lambda) - \frac{1}{2L_{d,\mu}} \sum_{i:\lambda_i > \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \left(\frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \right)^2 - \frac{1}{2} \sum_{i:\lambda_i \leq \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \lambda_i .
\end{aligned}$$

The statement now follows from $d_{\tau,\mu}(\widetilde{\lambda}) \geq d_{\tau,\mu}(\lambda_*)$. □

Lemma E.2.

$$\langle \lambda, \nabla d_\tau(\lambda) \rangle \leq \frac{L_{d,\mu}}{\mu} (d_{\tau,\mu}(\lambda) - d_{\tau,\mu}(\lambda_*)). \qquad (146)$$

Proof. From (145) we have:

$$d_{\tau,\mu}(\lambda) - d_{\tau,\mu}(\lambda_*) \geq \frac{1}{2L_{d,\mu}} \sum_{i:\lambda_i > \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \left(\frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \right)^2 + \frac{1}{2} \sum_{i:\lambda_i \leq \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \lambda_i . \qquad (147)$$

Note that

$$\left(\frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \right)^2 = \left(\frac{\partial d_\tau}{\partial \lambda_i} + \mu \lambda_i \right)^2 \geq 2\mu \frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \lambda_i, \qquad (148)$$

$$\frac{\partial d_{\tau,\mu}}{\partial \lambda_i} \lambda_i \geq \frac{\partial d}{\partial \lambda_i} \lambda_i, \ \lambda_i \geq 0. \qquad (149)$$

Then, from (148),(149)

$$d_{\tau,\mu}(\lambda) - d_{\tau,\mu}(\lambda_*) \geq \frac{2\mu}{2L_{d,\mu}} \sum_{i:\lambda_i > \frac{1}{L_{d,\mu}} \frac{\partial d_{\tau,\mu}}{\partial \lambda_i}} \frac{\partial d}{\partial \lambda_i} \lambda_i + \frac{1}{2} \sum_{i:\lambda_i \leq \frac{1}{L_{d,\mu}} \frac{\partial d_{\mu,\tau}}{\partial \lambda_i}} \frac{\partial d}{\partial \lambda_i} \lambda_i \geq \frac{\mu}{L_{d,\mu}} \left\langle \nabla d_\tau(\lambda), \lambda \right\rangle, \qquad (150)$$

and we get:

$$\langle \nabla d_\tau(\lambda), \lambda \rangle \leq \frac{L_{d,\mu}}{\mu} (d_{\tau,\mu}(\lambda) - d_{\tau,\mu}(\lambda_*)). \qquad (151)$$

□

Regularization by μ:

Lemma E.3.

$$\|\lambda_*^\mu\|_2^2 \leqslant \frac{2}{\mu} \left(d_\tau(0) - d_\tau\left(\lambda_*\right) \right), \qquad (152)$$

where λ_*^μ is the solution of (144).

Proof. From $\frac{\mu}{2} \left\| \lambda - \lambda_* \right\|_2^2 \leq d_\tau(\lambda) - d_\tau(\lambda_*)$, we get

$$d_{\tau,\mu}\left(\lambda_*^\mu\right) = d_\tau\left(\lambda_*^\mu\right) + \frac{\mu}{2} \left\|\lambda_*^\mu\right\|_2^2 \geqslant d_\tau\left(\lambda_*\right) + \frac{\mu}{2} \left\|\lambda_*^\mu\right\|_2^2 \geqslant d_\tau\left(\lambda_*\right). \qquad (153)$$

Then

$$\frac{\mu}{2} \left\| \lambda_\mu^* \right\|_2^2 = \frac{\mu}{2} \left\| 0 - \lambda_\mu^* \right\|_2^2 \leqslant d_{\tau,\mu}(0) - d_{\tau,\mu}\left(\lambda_*^\mu\right) \leqslant d_\tau(0) - d_\tau\left(\lambda_*\right), \qquad (154)$$

where we used that $d_{\tau,\mu}(0) = d_\tau(0)$. □

F EXPERIMENTAL PARAMETERS

F.1 Environment

The reward and cost functions in our environment are the same as in Li et al. (2021). The agent receives a reward +1 for the end of the lower link being at a height of 0.5 and cost one of 1 when the first link swings at a anticlockwise direction and the agent applies a +1 torque to the actuating joint; it also receives a cost two of 1 when the second link swings at a anticlockwise direction with respect to the first link and the agent applies a +1 torque to the actuating joint. The cost thresholds are 50.

F.2 Algorithm Parameters

The policy networks for all experiments have two hidden layers of sizes 128 with ReLu activation function. We also have value networks with the same architecture and activation functions as the policy networks. Table 2 summarizes the hyperparameters used in our experiments.

Table 2: Hyperparameters in total reward case.

Hyperparameter	VMDP	AR-CPO
Batch size	5100	5100
Discount γ	0.98	0.98
Maximum episode length	500	500
Learning rate	1	1
The number of policy optimization steps in NPG subroutine	4	1
max_KL: the parameter that controls the NPG updates	0.01	0.01
η: the parameter in the update of λ from Li et al. (2021)	N/A	0.0003
s from Li et al. (2021)	N/A	1
H from Li et al. (2021)	N/A	45
Entropy regularisation constant τ	0.01	0
Regularization coefficient μ	0	0
Optimization set radius for λ	1	N/A
η from Algorithm 1	1000	N/A
ζ from Algorithm 1	10^{-1}	N/A

F.3 Algorithm Parameters for Discounted Case

The most of the parameters are similar to according ones in Table 2. Values, which differ, are represented in Table 3.

Table 3: Changed hyperparameters in discounted case.

Hyperparameter	VMDP	AR-CPO
The number of policy optimization steps in NPG subroutine	40	1
max_KL: the parameter that controls the NPG updates	0.01	0.001
Entropy regularisation constant τ	0.001	0
Regularization coefficient μ	0.01	0.001