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

Donghao Ying, Mengzi Amy Guo, Yuhao Ding, Javad Lavaei, Zuo-Jun Shen

UC Berkeley, Department of Industrial Engineering and Operations Research
{donghaoy, mengzi_guo, yuhao_ding, lavaei, maxshen}@berkeley.edu

Abstract

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

1 Introduction

Reinforcement Learning (RL) aims to learn how to map situations to actions so as to maximize the expected cumulative reward. Mathematically, this objective can be rewritten as an inner product between the state-action occupancy measure induced by the policy and a policy-independent reward for each state-action pair. However, many decision-making problems of interests take a form beyond the cumulative reward, such as apprenticeship learning (Abbeel and Ng 2004), diverse skill discovery (Eysenbach et al. 2018), pure exploration (Hazan et al. 2019), and state marginal matching (Lee et al. 2019), among others. Recently, Zhang et al. (2020) and Zahavy et al. (2021) abstract such problems as convex Markov Decision Processes (MDPs), which focus on finding a policy to maximize a concave function of the induced state-action occupancy measure.

However, in many safety-critical applications of convex MDP problems, e.g., in autonomous driving (Fisac et al. 2018), cyber-security (Zhang et al. 2019), and financial management (Abe et al. 2010), the agent is also subject to safety constraints. Nonetheless, the classical safe RL and CMDPs (Altman 1999), which assume that the objective and constraints are linear in the state-action occupancy measure, are not directly applicable to more general convex CMDP problems where the objective and the constraints can respectively be concave and convex in the occupancy measure.

In this paper, we focus on the optimization perspective of convex CMDP problems and aim to develop a principled theory for the direct policy search method. When moving beyond linear structures in the objective and the constraints, we quickly face several technical challenges. Firstly, the convex CMDP problem has a nonconcave objective and the nonconvex constraints with respect to the policy even under the simplest direct policy parameterization. Thus, the existing tools from the convex constrained optimization literature are not applicable. Secondly, as the gradient of the objective/constraint with respect to the occupancy measure becomes policy-dependent, evaluating the single-step improvement of the algorithm becomes harder without knowing the occupancy measure. Yet, evaluating the occupancy measure for a given policy can be inefficient (Tsybakov 2008). Thirdly, the performance difference lemma in Kakade and Langford (2002), which is key to the analysis of the policy-based primal-dual method for the standard CMDP in Ding et al. (2020), is no longer helpful for general convex CMDPs.

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

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

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

Table 1: We summarize our results for general convex CMDPs. Here T is the total number of iterations.

1.1 Related Work

Convex MDP Motivated by emerging applications in RL whose objectives are beyond cumulative rewards (Schaal 1996; Abbeel and Ng 2004; Ho and Ermon 2016; Hazan et al. 2019; Rosenberg and Mansour 2019; Lee et al. 2019), a series of recent works have focused on developing general approaches for convex MDPs. In particular, Zhang et al. (2020) develops a new policy gradient approach called variational policy gradient and establishes the global convergence of the gradient ascent method by exploiting the hidden convexity of the problem. The REINFORCE-based policy gradient and its variance-reduced version are studied in Zhang et al. (2021). Zahavy et al. (2021) transforms the convex MDP problem to a saddle-point problem using Fenchel duality and proposes a meta-algorithm to solve the problem with standard RL techniques. Geist et al. (2021) proves the equivalence between convex MDPs and mean-field games (MFGs) and shows that algorithms for MFGs can be used to solve convex MDPs. However, the above papers only consider the unconstrained RL problem, which may lead to undesired policies in safety-critical applications. Therefore, additional effort is required to deal with the rising safety constraints, and our work addresses this challenge.

CMDP Our work is also pertinent to policy-based CMDP algorithms (Altman 1999; Borkar 2005; Achiam et al. 2017; Ding and Lavaei 2022; Chow et al. 2017; Efroni, Mannor, and Pirotta 2020). In particular, Ding et al. (2020) develops a natural policy gradient-based primal-dual algorithm and shows that it enjoys an $\mathcal{O}(T^{-1/2})$ global convergence rate regarding both the optimality gap and the constraint violation under the standard soft-max parameterization. Xu, Liang, and Lan (2021) considers a primal-based approach and achieves a similar global convergence rate. More recently, a line of works (Ying, Ding, and Lavaei 2021; Liu et al. 2021a; Li et al. 2021) introduce entropy regularization and obtain improved convergence rates with dual methods. Nonetheless, these papers focus on cumulative rewards/utilities and do not directly generalize to a boarder class of safe RL problems, such as safe imitation learning (Zhou and Li 2018) and safe exploration (Hazan et al. 2019). Beyond CMDPs with cumulative rewards/utilities, Bai et al. (2021) also studies the convex CMDP problem. Their algorithm is based on the randomized linear programming method proposed by Wang (2020) and it exhibits $\widetilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity. However, as their approach works directly in the space of state-action occupancy measures, it is thus not applicable to more general problems where the state-action spaces are large and a function approximation is needed. In comparison, our work addresses this issue by focusing on the policy-based primal-dual method and adopting a general soft-max policy parameterization.

1.2 Notations

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

2 Problem Formulation

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

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

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

$$Q^{\pi}(r;s,a) := \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r\left(s_k,a_k\right)\middle|\pi, s_0=s, a_0=a\right], \quad (2)$$

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

$$\lambda^{\pi}(s,a) = (1-\gamma)\sum_{k=0}^{\infty}\gamma^k\mathbb{P}\left(s_k=s,a_k=a\mid\pi,s_0\sim\rho\right). \quad (3)$$

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

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

This is known as the linear programming formulation of the CMDP (Altman 1999). Once a solution $\lambda^{\star}$ is computed, the corresponding policy can be recovered using the relation $\pi(a|s)=\lambda^{\pi}(s,a)/\sum_{a'\in\mathcal{A}}\lambda^{\pi}(s,a')$.

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

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

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

Example 2.1 (Safety-aware apprenticeship learning (AL)) *In AL, instead of maximizing the long-term reward, the agent learns to mimic an expert's demonstrations. When there are critical safety requirements, the learner will also strive to satisfy given constraints on the expected total cost (Zhou and Li 2018). This problem can be formulated as*

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

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

Example 2.2 (Feasibility constrained MDPs) *As an extension to standard CMDPs, the designer may desire to control the MDP through more general constraints described by a convex feasibility region C (Miryoosefi et al. 2019) (e.g., a single point representing a known safe policy) such that the learned policy is not too far away from C. In this case, the problem can be cast as*

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

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

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

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

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

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

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

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

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

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

The Slater's condition is a standard assumption and it holds when the feasible region has an interior point. In practice, such a point is often easy to find using prior knowledge of the problem. The following result is a direct consequence of the Slater's condition (Altman 1999).

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

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

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

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

3 Safe Policy Search beyond Cumulative Rewards/Utilities

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

$$\begin{cases} \theta^{t+1} = \mathcal{P}_\Theta\left(\theta^t + \eta_1 \nabla_\theta \mathcal{L}(\theta^t, \mu^t)\right) \\ \mu^{t+1} = \mathcal{P}_U\left(\mu^t - \eta_2 \nabla_\mu \mathcal{L}(\theta^t, \mu^t)\right) \end{cases}, \; t = 0, 1, \ldots \tag{11}$$

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

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

3.1 Gradient Evaluation of the Lagrangian Function

Computing the primal update in (11) involves evaluating the gradient of the Lagrangian with respect to θ, i.e., $\nabla_\theta \mathcal{L}(\theta, \mu) = \nabla_\theta [f(\lambda(\theta)) - \mu g(\lambda(\theta))]$. When $f(\cdot)$ and $g(\cdot)$ are linear as in the standard CMDP, i.e., $f(\lambda) = \langle r, \lambda \rangle/(1-\gamma)$ and $g(\lambda) = \langle c, \lambda \rangle/(1-\gamma)$, the Lagrangian can be viewed as a value function with reward $r - \mu c$, i.e., $\mathcal{L}(\theta, \mu) = V^{\pi_\theta}(r - \mu c)$, so the *Policy gradient* theorem proposed by Sutton et al. (1999) can be applied (cf. Lemma G.1). However, this favorable result is not applicable when $f(\cdot)$ and $g(\cdot)$ are general concave/convex functions. Instead, we present two alternative approaches below. More details are deferred to Appendix A.1.

Variational Policy Gradient (Zhang et al. 2020) By leveraging the Fenchel duality, Zhang et al. (2020) showed that the gradient $\nabla_\theta \mathcal{L}(\theta, \mu)$ can be computed by solving a stochastic saddle point problem, in particular

$$\nabla_\theta \mathcal{L}(\theta,\mu) = \lim_{\delta \to 0_+} \operatorname*{argmax}_x \inf_z \bigg\{ (1-\gamma)\big[V^{\pi_\theta}(z) + \delta \langle \nabla_\theta V^{\pi_\theta}(z), x \rangle\big] - L_*(z, \mu) - \frac{\delta}{2}\|x\|^2 \bigg\}, \tag{12}$$

where $L_*(z, \mu) := \inf_\lambda \{\langle z, \lambda \rangle - L(\lambda, \mu)\}$ is the concave conjugate of $L(\lambda, \mu)$ with respect to λ. As $L_*(z, \mu)$ is concave in z and $V^{\pi_\theta}(z) = \langle z, \lambda(\theta) \rangle$ is linear in z, the max-min problem in (12) is a concave-convex saddle point problem and can be solved by a gradient ascent-descent algorithm.

REINFORCE-Based Policy Gradient (Zhang et al. 2021) By noticing the relation $\nabla_\theta V^{\pi_\theta}(r) = [\nabla_\theta \lambda(\theta)]^\top \cdot r/(1-\gamma)$, one can view $\nabla_\theta \mathcal{L}(\theta, \mu)$ as the standard policy gradient for the value function with the reward $\nabla_\lambda L(\lambda(\theta), \mu)$ (assuming that $f(\lambda)$, $g(\lambda)$, and $\lambda(\theta)$ are all differentiable), i.e.,

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta, \mu) &= [\nabla_\theta \lambda(\theta)]^\top \cdot \nabla_\lambda L(\lambda(\theta), \mu) \\ &= (1-\gamma) \nabla_\theta V^{\pi_\theta}(\nabla_\lambda L(\lambda(\theta), \mu)), \end{aligned} \tag{13}$$

where the first equality follows from the chain rule. Thus, the gradient $\nabla_\theta \mathcal{L}(\theta, \mu)$ can be estimated with the REINFORCE algorithm proposed by Williams (1992) as long as we can choose an approximation of $\nabla_\lambda L(\lambda(\theta), \mu)$ as the reward.

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

$$g(\lambda(\theta)) = \sup_z \{(1-\gamma)V^{\pi_\theta}(z) - g^*(z)\}. \tag{14}$$

where $g^*(z) := \sup_\lambda \{\langle z, \lambda \rangle - g(\lambda)\}$ is the convex conjugate of $g(\cdot)$ and we leverage the fact that the biconjugate of a convex function equals itself, i.e., $g^{**}(\lambda) := \sup_z \{\langle z, \lambda \rangle - g^*(z)\} = g(\lambda)$. In Appendix A.1, we provide a sample-based pseudocode for algorithm (11).

3.2 Exploiting the Hidden Convexity

By itself, (10) is a nonconvex-linear maximin problem. The existing results for the analysis of the gradient ascent descent method (11) for such problems can only guarantee to find a ϵ-stationary point in $\mathcal{O}(\epsilon^{-6})$ iterations (Lin, Jin, and Jordan 2020). To obtain an improved convergence rate and achieve a global optimality, it is necessary to exploit the "hidden convexity" of (9) with respect to λ.

However, standard analyses based on the performance difference lemma (cf. Lemma G.4) do not apply to convex CMDPs (Agarwal et al. 2021; Ding et al. 2020). A key insight is that, due to the loss of linearity, the performance difference lemma together with concavity can only provide an upper bound for the single-step improvement with the gradient information at the current step (more details can be found in Appendix A.2). Thus, this prompts us to introduce a new analysis to bound the average performance in terms of the Lagrangian (cf. (16)).

Motivated by Zhang et al. (2020), we leverage the fact that the primal update implies the formula

$$\begin{aligned} &\mathcal{L}(\theta^{t+1}, \mu^t) - \mathcal{L}(\theta^t, \mu^t) \\ &= \max_{\theta \in \Theta} \bigg\{ (\theta - \theta^t)^\top \nabla_\theta \mathcal{L}(\theta^t, \mu^t) - \frac{1}{2\eta_1}\|\theta - \theta^t\|_2^2 \bigg\}. \end{aligned} \tag{15}$$

The basis of our analysis lies in designing a special point θ from θ^t and $\theta^\star$ to lower-bound $\mathcal{L}(\theta^{t+1}, \mu^t)$ through (15). The hidden convexity of (9) with respect to λ plays a central role in bounding the improvement $\mathcal{L}(\theta^{t+1}, \mu^t) - \mathcal{L}(\theta^t, \mu^t)$ and relating it to the sub-optimality gap $\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)$. The details are postponed to Section 4.

4 Convergence Analysis

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

First, we formally state our assumption about the parameterization (8). To avoid introducing an additional bias, it is natural to assume that the parameterization has enough expressibility to represent any policy, i.e., $\forall \pi \in \Pi$, $\exists \theta \in \Theta$ such that $\pi = \pi_\theta$. However, assuming a one-to-one correspondence between $\pi \in \Pi$ and $\pi_\theta, \theta \in \Theta$ is too restrictive. In practice, using a deep neural network to represent the policy can often arrive at an over-parameterization. Therefore, following Zhang et al. (2021), we assume that π_θ is defined such that it can represent any policy and that $\lambda(\cdot)$ is locally continuously invertible. A more detailed discussion can be found in Appendix D.

Assumption 4.1 (Parameterization) *The policy parameterization $\pi = \pi_\theta$ over-parameterizes the set of all stochastic policies and satisfies:*

(I) *For every $\theta \in \Theta$, there exists a neighborhood $\mathcal{U}_\theta \ni \theta$ such that the restriction of $\lambda(\cdot)$ to $\mathcal{U}_\theta$ is a bijection between $\mathcal{U}_\theta$ and $\mathcal{V}_{\lambda(\theta)} := \lambda(\mathcal{U}_\theta)$;*

(II) *Let $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta)}} : \mathcal{V}_{\lambda(\theta)} \to \mathcal{U}_\theta$ be the local inverse of $\lambda(\cdot)$, i.e., $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta)}}(\lambda(\theta_0)) = \theta_0$, $\forall \theta_0 \in \mathcal{U}_\theta$. Then, there exists a universal constant ℓ_Θ such that $\lambda^{-1}_{\mathcal{V}_{\lambda(\theta)}}$ is ℓ_Θ-Lipschitz continuous for all $\theta \in \Theta$;*

(III) *There exists $\bar{\varepsilon} > 0$ such that $(1-\varepsilon)\lambda(\theta) + \varepsilon\lambda(\theta^\star) \in \mathcal{V}_{\lambda(\theta)}$, $\forall \varepsilon \le \bar{\varepsilon}$, $\forall \theta \in \Theta$.*

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

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

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

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

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

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

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

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

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)\right] \\ &\le \frac{\mathcal{L}(\theta^\star, \mu^0) - \mathcal{L}(\theta^0, \mu^0)}{\varepsilon T} + 2\varepsilon \ell_L \ell_\Theta^2 + \frac{2\eta_2 M_G^2}{\varepsilon}.\end{aligned} \tag{16}$$

We remark that by choosing $\varepsilon = T^{-1/3}$ and $\eta_2 = T^{-2/3}$, the bound (16) given by Proposition 4.4 has the order of $\mathcal{O}\left(T^{-1/3}\right)$. The core idea in proving Proposition 4.4 is that one can relate the primal update (15) to the sub-optimality gap $\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)$ by leveraging the hidden convexity of (9) with respect to λ. Then, as $|\mu^{t+1} - \mu^t| = \mathcal{O}(\eta_2)$, we are able to draw a recursion between the sub-optimality gaps for two consecutive periods (cf. (38)).

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

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)\right] \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] + \frac{1}{T}\sum_{t=0}^{T-1}\mu^t\left[G(\theta^t) - G(\theta^\star)\right].\end{aligned} \tag{17}$$

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

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

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[F(\theta^\star) - F(\theta^t)\right] \le \frac{2M_F + M_G^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2M_G^2}{T^{1/3}}, \tag{18a}$$

$$\frac{1}{T}\left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ \le \frac{2M_F + M_G^2/2}{T^{2/3}} + \frac{2\ell_L\ell_\Theta^2 + 2M_G^2 + C_0^2/2}{T^{1/3}}. \tag{18b}$$

Theorem 4.5 shows that algorithm (11) achieves a global convergence in the average sense such that the optimality gap and the constraint violation decay to zero with the rate $\mathcal{O}\left(T^{-1/3}\right)$. In other words, to obtain an $\mathcal{O}(\epsilon)$-accurate solution, the iteration complexity is $\mathcal{O}(\epsilon^{-3})$. When $f(\cdot)$ and $g(\cdot)$ are linear functions as in standard CMDPs (cf. (1)), Theorem 4.5 matches the rate of the natural policy gradient primal-dual algorithm under the general parameterization (Ding et al. 2020, Theorem 2)[1].

[1]Under the general parameterization, the convergence rate presented in Ding et al. (2020, Theorem 2) is $\left(\mathcal{O}\left(T^{-1/2}\right), \mathcal{O}\left(T^{-1/4}\right)\right)$ under the parameters $\eta_1 = \eta_2 = \mathcal{O}(T^{-1/2})$. This rate can be converted to $\left(\mathcal{O}\left(T^{-1/3}\right), \mathcal{O}\left(T^{-1/3}\right)\right)$ by choosing the parameters $\eta_1 = \eta_2 = \mathcal{O}(T^{-1/3})$.

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

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

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

$$\begin{aligned} &\frac{1}{T} \sum_{t=0}^{T-1} [\mathcal{L}(\theta^\star, \mu^t) - \mathcal{L}(\theta^t, \mu^t)] \\ &\le \frac{\mathcal{L}(\theta^T, \mu^{T-1}) - \mathcal{L}(\theta^0, \mu^0)}{\tilde{\varepsilon} T} + \frac{\eta_2 M_G^2}{\tilde{\varepsilon}}, \end{aligned} \tag{19}$$

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

Different from the general concave case (16), the bound (19) does not contain the constant error term $\mathcal{O}(\varepsilon)$. Thus, by choosing $\eta_2 = T^{-1/2}$, the average performance has the order $\mathcal{O}\left(T^{-1/2}\right)$. In a similar manner as in Theorem 4.5, we can decouple the average performance to bound the optimality gap and constraint violation.

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

$$\frac{1}{T} \sum_{t=0}^{T-1} \left[F(\theta^\star) - F(\theta^t)\right] \le \frac{M_L + M_F}{\tilde{\varepsilon} T} + \left(\frac{M_G^2}{\tilde{\varepsilon}} + \frac{M_G^2}{2}\right) \frac{1}{\sqrt{T}}, \tag{20a}$$

$$\frac{1}{T} \left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ \le \frac{M_L + M_F}{\tilde{\varepsilon} T} + \left(\frac{M_G^2}{\tilde{\varepsilon}} + \frac{M_G^2 + C_0^2}{2}\right) \frac{1}{\sqrt{T}}, \tag{20b}$$

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

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

Remark 4.8 (Direct parameterization) *As a special case, the direct parameterization satisfies Assumption 4.1 once there is a universal positive lower bound for the state occupancy measure defined as $d^\pi(s) = \sum_a \lambda^\pi(s,a), \forall s$. Under the direct parameterization, it can be shown that the primal update (15) also enjoys the so-called variational gradient dominance property for standard MDPs (see, e.g., Agarwal et al. (2021, Lemma 4.1)). This evidence gives a clearer intuition of how the hidden convexity enables us to prove the global convergence of algorithm (11). We refer the reader to Appendix F for a detailed discussion.*

5 Zero Constraint Violation

In safety-critical systems where violating the constraint may induce an unexpected cost, having a zero constraint violation is of great importance. Motivated by the recent works (Liu et al. 2021b,a), we will show that a zero constraint violation can be achieved while maintaining the same order of convergence rate for the optimality gap. Consider the pessimistic counterpart of (9):

$$\max_{\theta \in \Theta} F(\theta) = f(\lambda(\theta)) \;\text{ s.t. }\; G_\delta(\theta) := g(\lambda(\theta)) + \delta \le 0, \tag{21}$$

where $\delta > 0$ is the pessimistic term to be determined. Suppose that $\delta < \xi$, i.e., the pessimistic term is smaller than the strict feasibility of the Slater point, so that the Slater's condition still holds for problem (21). Consider applying algorithm (11) to problem (21), i.e.,

$$\begin{cases} \theta^{t+1} = \mathcal{P}_\Theta\left(\theta^t + \eta_1 \nabla_\theta \mathcal{L}_\delta(\theta^t, \mu^t)\right) \\ \mu^{t+1} = \mathcal{P}_U\left(\mu^t - \eta_2 \nabla_\mu \mathcal{L}_\delta(\theta^t, \mu^t)\right) \end{cases}, \; t = 0, 1, \ldots \tag{22}$$

where $\mathcal{L}_\delta(\theta, \mu) := F(\theta) - \mu G_\delta(\theta)$ is the Lagrangian function for (21). We refer to this variate as the Primal-dual Policy Gradient-Zero Algorithm (PDPG-0). The following theorem states that, with a carefully chosen δ, the constraint violation of the iterates generated by algorithm (22) will be zero for the original problem (9) when T is reasonably large and the optimality gap remains the same order as before. Due to the limit of space, we present an informal version of the theorem below and direct the reader to Appendix C for a detailed statement as well as the proof.

Theorem 5.1 *Let Assumptions 2.1, 4.1, and 4.2 hold.*

(I) *For every reasonably large $T > 0$, the sequence $\left\{(\theta^t, \mu^t)\right\}_{t=0}^{T-1}$ generated by the PDPG-0 algorithm with $\delta = \mathcal{O}(T^{-1/3})$ satisfies*

$$\frac{1}{T} \sum_{t=0}^{T-1} \left[F(\theta^\star) - F(\theta^t)\right] = \mathcal{O}\left(T^{-1/3}\right), \; \frac{1}{T} \left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = 0. \tag{23}$$

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

$$\frac{1}{T} \sum_{t=0}^{T-1} \left[F(\theta^\star) - F(\theta^t)\right] = \mathcal{O}(T^{-1/2}), \; \frac{1}{T} \left[\sum_{t=0}^{T-1} G(\theta^t)\right]_+ = 0. \tag{24}$$

We briefly introduce the ideas behind Theorem 5.1 here. Adding the pessimistic term δ would shift the optimal solution from $\theta^\star$ to another point $\theta_\delta^\star$. By leveraging the Slater's condition, we can upper-bound the sub-optimality gap $|F(\theta^\star) - F(\theta_\delta^\star)|$ by $\mathcal{O}(\delta)$. Since the orders of convergence rates are the same for optimality gap and constraint violation in Theorems 4.5 and 4.7, we can choose δ to have the same order and then offset the constraint violation for the pessimistic problem (21). As a result, the constraint violation becomes zero for the original problem (9) and the optimality gap preserves its previous order.

6 Numerical Experiment

In this section, we validate algorithm (11) in a feasibility constrained MDP problem (cf. Example 2.2). The experiment is performed on the single agent version of OpenAI Particle environment (Lowe et al. 2017) as illustrated in Figure 1a.

The state space is continuous and formed by the two-dimensional location of the agent. At each state, there are five actions available to the agent, corresponding to moving leftward, rightward, upward, downward, or staying at the current place. The goal of the agent is to safely navigate from the starting point "S" to the target point "T". In particular, they will be rewarded as they approach the target and will be penalized when they step onto the unsafe region in the middle. It can be observed from Figure 1a that there are two high-reward routes available for the agent as highlighted by the blue dashed lines. Let π_0 be the deterministic policy corresponding to the red curve in Figure 1a and λ_0 be the associated state-action occupancy measure. We enforce the feasibility constraint $\|\lambda^\pi - \lambda_0\|_2 \leq d_0$, where $d_0 > 0$ is the allowable deviation. This constraint means that we want the deviation of the learned policy from the demonstration π_0 is within a threshold d_0. In summary, this problem can be formulated as:

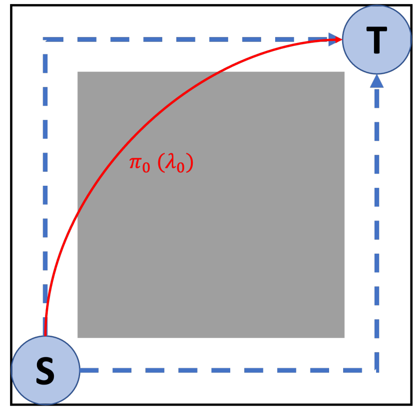
$$\max_{\pi\in\Pi} \frac{1}{1-\gamma} \langle \lambda^\pi, r \rangle, \quad \text{s.t. } \|\lambda^\pi - \lambda_0\|_2 \leq d_0, \tag{25}$$

where r integrates both the reward and penalty signals.

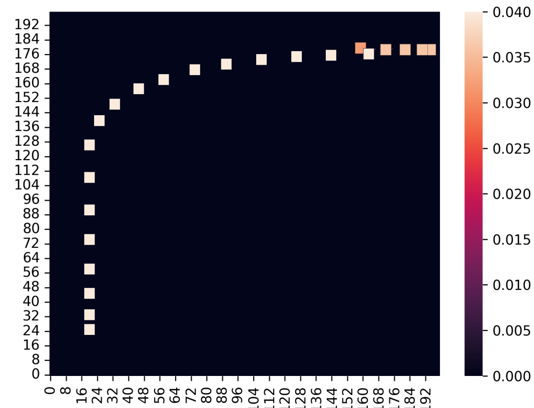
To solve the problem, we discretize the state space into a 200×200 size grid. The policy is parameterized by a two-layer fully-connected neural network with 64 neurons in each layer and ReLU activations. We estimate the policy gradient $\nabla_\theta \mathcal{L}(\theta, \mu)$ through the REINFORCE-based method (Zhang et al. 2021) with $n = 10$ and $K = 25$ (see Algorithm 1 in Appendix A.1). The feasibility constraint has a threshold of $d_0 = 0.2$. Figure 1b visualizes the state-action occupancy measure λ^π of the learned policy π. We observe that the learned policy chooses one of the high-reward routes that is closer to the demonstration π_0. The learning curve of the algorithm in terms of the cumulative reward and constraint violation is provided in the online appendix.

7 Conclusion

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



(a) Environment.



(b) Occupancy measure.

Figure 1: **(a)** The goal of the agent is to navigate from "S" to "T" while stepping onto the shaded region would incur a penalty. The two paths in blue dashed lines represent the two high-reward routes available. A feasibility constraint is enforced for the agent such that $\|\lambda^\pi - \lambda_0\| \leq d_0$, where λ_0 is the given demonstration corresponding to the red curve. **(b** Heatmap of the state-action occupancy measure λ^π corresponding to the learned policy π. It shows that the learned policy chooses one of the high-reward routes which is closed to the given demonstration λ_0.

One important direction of future work lies in establishing a lower bound for convex CMDP problems under a general soft-max parameterization to verify the optimality of our upper bounds. Also, an extension to this work is studying the sample complexity of the PDPG method. Furthermore, it is interesting to study whether geometric structures, such as entropy regularization (Ying, Ding, and Lavaei 2021; Liu et al. 2021a; Li et al. 2021) or policy mirror descent (Xiao 2022), can be exploited to accelerate the convergence.

Acknowledgments

This work was supported by grants from ARO, AFOSR, ONR and NSF.

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1.

Abe, N.; Melville, P.; Pendus, C.; Reddy, C. K.; Jensen, D. L.; Thomas, V. P.; Bennett, J. J.; Anderson, G. F.; Cooley, B. R.; Kowalczyk, M.; et al. 2010. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 75–84.

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained policy optimization. In *International conference on machine learning*, 22–31. PMLR.

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

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

Bai, Q.; Bedi, A. S.; Agarwal, M.; Koppel, A.; and Aggarwal, V. 2021. Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Primal-Dual Approach. *arXiv preprint arXiv:2109.06332v2*.

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

Boyd, S.; Boyd, S. P.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

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

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

Ding, Y.; and Lavaei, J. 2022. Provably Efficient Primal-Dual Reinforcement Learning for CMDPs with Non-stationary Objectives and Constraints. *arXiv preprint arXiv:2201.11965*.

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

Eysenbach, B.; Gupta, A.; Ibarz, J.; and Levine, S. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

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

Geist, M.; Pérolat, J.; Laurière, M.; Elie, R.; Perrin, S.; Bachem, O.; Munos, R.; and Pietquin, O. 2021. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*.

Hazan, E.; Kakade, S.; Singh, K.; and Van Soest, A. 2019. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2681–2691. PMLR.

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

Kakade, S.; and Langford, J. 2002. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer.

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

Lee, L.; Eysenbach, B.; Parisotto, E.; Xing, E.; Levine, S.; and Salakhutdinov, R. 2019. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*.

Li, T.; Guan, Z.; Zou, S.; Xu, T.; Liang, Y.; and Lan, G. 2021. Faster Algorithm and Sharper Analysis for Constrained Markov Decision Process. *arXiv preprint arXiv:2110.10351*.

Lin, T.; Jin, C.; and Jordan, M. 2020. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, 6083–6093. PMLR.

Liu, T.; Zhou, R.; Kalathil, D.; Kumar, P.; and Tian, C. 2021a. Fast Global Convergence of Policy Optimization for Constrained MDPs. *arXiv preprint arXiv:2111.00552*.

Liu, T.; Zhou, R.; Kalathil, D.; Kumar, P.; and Tian, C. 2021b. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Neural Information Processing Systems (NIPS)*.

Miryoosefi, S.; Brantley, K.; Daume III, H.; Dudik, M.; and Schapire, R. E. 2019. Reinforcement learning with convex constraints. *Advances in Neural Information Processing Systems*, 32.

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

Rosenberg, A.; and Mansour, Y. 2019. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, 5478–5486. PMLR.

Schaal, S. 1996. Learning from demonstration. *Advances in neural information processing systems*, 9.

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

Tsybakov, A. B. 2008. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition. ISBN 0387790519.

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

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

Xiao, L. 2022. On the Convergence Rates of Policy Gradient Methods. *arXiv preprint arXiv:2201.07443*.

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

Ying, D.; Ding, Y.; and Lavaei, J. 2021. A Dual Approach to Constrained Markov Decision Processes with Entropy Regularization. *arXiv preprint arXiv:2110.08923*.

Zahavy, T.; O'Donoghue, B.; Desjardins, G.; and Singh, S. 2021. Reward is enough for convex MDPs. *Advances in Neural Information Processing Systems*, 34.

Zhang, J.; Koppel, A.; Bedi, A. S.; Szepesvari, C.; and Wang, M. 2020. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33: 4572–4583.

Zhang, J.; Ni, C.; Szepesvari, C.; Wang, M.; et al. 2021. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34.

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

Zhou, W.; and Li, W. 2018. Safety-aware apprenticeship learning. In *International Conference on Computer Aided Verification*, 662–680. Springer.