Deterministic Policy Gradient Primal-Dual Methods for Continuous-Space Constrained MDPs

Sergio Rozada [1], **Dongsheng Ding** [2], **Antonio G. Marques** [1], **Alejandro Ribeiro** [2]

[1]Dept. of Signal Theory and Communications, King Juan Carlos University
[2]Dept. of Electrical and Systems Engineering, University of Pennsylvania
s.rozada.2019@alumnos.urjc.es, dongshed@seas.upenn.edu, antonio.garcia.marques@urjc.es, alejandro.ribeiro@seas.upenn.edu

Abstract

We study the problem of computing deterministic optimal policies for constrained Markov decision processes (MDPs) with continuous state and action spaces, which are widely encountered in constrained dynamical systems. Designing deterministic policy gradient methods in continuous state and action spaces is particularly challenging due to the lack of enumerable state-action pairs and the adoption of deterministic policies, hindering the application of existing policy gradient methods. To this end, we develop a deterministic policy gradient primal-dual method to find an optimal deterministic policy with non-asymptotic convergence. Specifically, we leverage regularization of the Lagrangian of the constrained MDP to propose a deterministic policy gradient primal-dual (D-PGPD) algorithm that updates the deterministic policy via a quadratic-regularized gradient ascent step and the dual variable via a quadratic-regularized gradient descent step. We prove that the primal-dual iterates of D-PGPD converge at a sub-linear rate to an optimal regularized primal-dual pair. We instantiate D-PGPD with function approximation and prove that the primal-dual iterates of D-PGPD converge at a sub-linear rate to an optimal regularized primal-dual pair, up to a function approximation error. Furthermore, we demonstrate the effectiveness of our method in two continuous control problems: robot navigation and fluid control. This appears to be the first work that proposes a deterministic policy search method for continuous-space constrained MDPs.

Code — https://github.com/sergiorozada12/d-pg-pd
Extended version — https://arxiv.org/abs/2408.10015

1 Introduction

Constrained Markov decision processes (MDPs) are a standard framework for incorporating system specifications into dynamical systems (Altman 2021; Brunke et al. 2022). In recent years, constrained MDPs have attracted significant attention in constrained Reinforcement Learning (RL), whose goal is to derive optimal control policies through interaction with unknown dynamical systems (Achiam et al. 2017; Tessler, Mankowitz, and Mannor 2018). Policy gradient-based constrained learning methods have become the workhorse driving recent successes across various disciplines, e.g., navigation (Paternain et al. 2022), video compression (Mandhane et al. 2022), and finance (Chow et al. 2018).

This paper is motivated by two observations. First, continuous state-action spaces are pervasive in dynamical systems, yet most methods in constrained RL are designed for discrete state and/or action spaces (Borkar 2005; Efroni, Mannor, and Pirotta 2020; Ding et al. 2022; Singh, Gupta, and Shroff 2022). Second, the literature on constrained RL largely focuses on stochastic policies. However, randomly taking actions by following a stochastic policy is often prohibitive in practice, especially in safety-critical domains (Sehnke et al. 2010; Li et al. 2022; Gao et al. 2023). Deterministic policies alleviate such concerns, but (i) they might lead to sub-optimal solutions (Ross 1989; Altman 2021); and (ii) computing them is NP-complete (Feinberg 2000; Dolgov 2005). Nevertheless, there is a rich body of constrained control literature that studies problems where optimal policies are deterministic (Posa, Kuindersma, and Tedrake 2016; Tsiamis et al. 2020; Zhao and You 2021; Ma et al. 2022). Viewing this gap, we study the problem of finding optimal *deterministic* policies for constrained MDPs with *continuous* state-action spaces.

A key consideration of this paper is the fact that deterministic policies are sub-optimal in finite state-action spaces, but sufficient for constrained MDPs with continuous state-action spaces (Feinberg and Piunovskiy 2002, 2019). This enables our formulation of a constrained RL problem with deterministic policies. To develop a tractable deterministic policy search method, we introduce a regularized Lagrangian approach that leverages proximal optimization methods. Moreover, we use function approximation to ensure scalability in continuous state-action spaces. Our main contribution is four-fold.

(i) We introduce a deterministic policy constrained RL problem for a constrained MDP with continuous state-action spaces and prove that the problem exhibits zero duality gap, despite being constrained to deterministic policies.

(ii) We propose a regularized deterministic policy gradient primal-dual (D-PGPD) algorithm that updates the primal policy via a proximal-point-type step and the dual variable via a gradient descent step, and we prove that the primal-dual iterates of D-PGPD converge to a set of regularized optimal primal-dual pairs at a sub-linear rate.

(iii) We propose an approximation for D-PGPD by including function approximation. We prove that the primal-dual

iterates of the approximated D-PGPD converge at a sub-linear rate, up to a function approximation error.

(iv) We demonstrate that D-PGPD addresses the classical constrained navigation problem involving several types of cost functions and constraints. We show that D-PGPD can solve non-linear fluid control problems under constraints.

Related work. Deterministic policy search has been studied in the context of unconstrained MDPs (Silver et al. 2014; Lillicrap et al. 2015; Kumar et al. 2020; Lan 2022). In constrained setups, however, deterministic policies have been largely restricted to occupancy measure optimization in finite state-action spaces (Dolgov 2005) or are embedded in hyper-policies (Sehnke et al. 2010; Montenegro et al. 2024a,b). This work extends deterministic policy search to constrained MDPs with continuous state-action spaces, overcoming two main roadblocks: the sub-optimality of deterministic policies and the NP-completeness of computing them (Ross 1989; Feinberg 2000; Dolgov 2005; Altman 2021; McMahan 2024). First, we show that deterministic policies are sufficient for constrained MDPs in continuous state-action spaces (Feinberg and Piunovskiy 2002, 2019), leveraging the convexity of the value image to establish strong duality in the deterministic policy space. Second, we overcome computational intractability by introducing a quadratic regularization of the reward function and proposing a regularization-based primal-dual algorithm. This algorithm exploits the structure of value functions and achieves last-iterate convergence to an optimal deterministic policy. While last-iterate convergence of primal-dual algorithms has been explored in constrained RL (Moskovitz et al. 2023; Ding et al. 2024; Ding, Huan, and Ribeiro 2024), existing methods focus on stochastic policies and finite-action spaces. In control, extensive work addresses deterministic policies in constrained setups with continuous state-action spaces (Scokaert and Rawlings 1998; Lim and Zhou 1999). However, these approaches are typically model-based and tailored to specific structured problems (Posa, Kuindersma, and Tedrake 2016; Tsiamis et al. 2020; Zhao, You, and Başar 2021; Zhao and You 2021; Ma et al. 2022). Bridging constrained control and RL has also been explored (Kakade et al. 2020; Zahavy et al. 2021; Li et al. 2023), but these methods remain model-based and focus on stochastic policies. In contrast, we propose a model-free deterministic policy search method for constrained MDPs with continuous state-action spaces.

2 Preliminaries

We consider a discounted constrained MDP, denoted by the tuple $(S, A, p, r, u, b, \gamma, \rho)$. Here, $S \subseteq \mathbb{R}^{d_s}$ and $A \subseteq \mathbb{R}^{d_a}$ are continuous state-action spaces with dimensions d_s and d_a, and bounded actions $\|a\| \leq A_{\max}$ for all $a \in A$; $p(\cdot \mid s, a)$ is a probability measure over S parametrized by the state-action pairs $(s, a) \in S \times A$; $r, u\colon S \times A \mapsto [0, 1]$ are reward/utility functions; b is a constraint threshold; $\gamma \in [0, 1)$ is a discount factor; and ρ is a probability measure that specifies an initial state. We consider the set of all deterministic policies Π in which a policy $\pi\colon S \mapsto A$ maps states to actions. The transition p, the initial state distribution ρ, and the policy π define a distribution over trajectories $\{s_t, a_t, r_t, u_t\}_{t=0}^{\infty}$, where $s_0 \sim \rho$, $a_t = \pi(s_t)$, $r_t = r(s_t, a_t)$, $u_t = u(s_t, a_t)$ and $s_{t+1} \sim p(\cdot \mid s_t, a_t)$. Given π, we define the value function $V_r^\pi\colon S \to \mathbb{R}$ as the expected sum of discounted rewards

$$V_r^\pi(s) \;:=\; \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right].$$

For the utility function, we define the corresponding value function V_u^π. Their expected values over the initial state distribution ρ are denoted as $V_r(\pi) := \mathbb{E}_\rho[V_r^\pi(s)]$ and $V_u(\pi) := \mathbb{E}_\rho[V_u^\pi(s)]$, where we drop the dependence on ρ for simplicity of notation. Boundedness of r and u leads to $V_r(\pi), V_u(\pi) \in [0, 1/(1-\gamma)]$. We introduce a discounted state visitation distribution $d_{s_0}^\pi(B) := (1-\gamma) \sum_{t=0}^{\infty} \Pr(s_t \in B \mid \pi, s_0)$ for any $B \subseteq S$ and define $d_\rho^\pi(s) := \mathbb{E}_{s_0 \sim \rho}[d_{s_0}^\pi(s)]$. For the reward function r, we define the state-action value function $Q_r^\pi\colon S \times A \to \mathbb{R}$ given an initial action a while following π,

$$Q_r^\pi(s, a) \;:=\; \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a \right].$$

We let the associated advantage function $A_r^\pi\colon S \times A \to \mathbb{R}$ be $A_r^\pi(s, a) \;:=\; Q_r^\pi(s, a) - V_r^\pi(s)$. Similarly, we define $Q_u^\pi\colon S \times A \to \mathbb{R}$ and $A_u^\pi\colon S \times A \to \mathbb{R}$ for the utility function u.

A policy is optimal for a given reward function when it maximizes the corresponding value function. However, the value functions $V_r(\pi)$ and $V_u(\pi)$ are usually in conflict, e.g., a policy that maximizes $V_r(\pi)$ is not necessary good for $V_u(\pi)$. To trade off two conflicting objectives, constrained MDP aims to find an optimal policy $\pi^\star$ that maximizes the reward value function $V_r(\pi)$ subject to an inequality constraint on the utility value function $V_u(\pi) \geq b$, where we assume $b \in (0, 1/(1-\gamma)]$ to avoid trivial solutions. We use a single constraint for the sake of simplicity, but our results extend to problems with multiple constraints. We translate the constraint $V_u(\pi) \geq b$ into the constraint $V_g(\pi) \geq 0$ for $g := u - (1-\gamma)b$, where $g\colon S \times A \mapsto [-1, 1]$ denotes the translated utility. This leads to the following problem

$$\begin{array}{ll} \max\limits_{\pi \in \Pi} & V_r(\pi) \\ \text{s.t.} & V_g(\pi) \;\geq\; 0. \end{array} \tag{1}$$

Restricting Problem (1) to deterministic policies poses several challenges. Deterministic policies can be sub-optimal in constrained MDPs with finite state-action spaces (Ross 1989; Altman 2021), and when they exist, finding them is a NP-complete (Feinberg 2000). Problem (1) is non-convex in the policy but can be reformulated as a linear program using occupancy measures with stochastic policies (Paternain et al. 2019). However, the occupancy measure representation of (1) is a *non-linear* and *non-convex* problem when only deterministic policies are considered (Dolgov 2005). Finally, multiple policies can achieve the optimal value function $V_P^{\pi^\star}$ while satisfying the constraint. We denote the set of all maximizers of (1) that attain $V_P^{\pi^\star}$ as $\Pi^\star$. To address these points, we observe that deterministic policies are sufficient in constrained MDPs with continuous state-action spaces under the following assumption (Feinberg and Piunovskiy 2002, 2019).

Assumption 1 (Non-atomicity)**.** *The MDP is non-atomic, i.e., $\rho(s) = 0$ and $p(s' \mid s, a) = 0$ for all s, $s' \in S$ and $a \in A$.*

Assumption 1 is mild in practice. Since stochastic perturbations are common in physical systems with continuous state and action spaces (Anderson and Moore 2007), the probability measures ρ and $p(\cdot \,|\, s, a)$ are normally atomless, i.e., for any measurable set $B \subseteq S$ with probability measures $\rho(B)$ and $p(B \,|\, s, a)$, there exists a measurable subset $B' \subset B$ that has smaller non-zero probability measures $\rho(B) > \rho(B') > 0$ and $p(B \,|\, s, a) > p(B' \,|\, s, a) > 0$ for any $s \in S$ and $a \in A$. In other words, the transition probability and the initial probability do not concentrate in a single state (Feinberg and Piunovskiy 2019). When a constrained MDP is non-atomic, only considering deterministic policies is sufficient (Feinberg and Piunovskiy 2019). Specifically, let $V(\pi) := [\, V_r(\pi) \;\; V_g(\pi) \,]^\top$ denote the vector of value functions for a given policy π. We define a *deterministic value image* $\mathcal{V}_D := \{V(\pi) \,|\, \pi \in \Pi\}$, which is a set of all attainable vector value functions for deterministic policies. We denote by $\mathcal{V}_T$ a *value image* for all policies. The deterministic value image $\mathcal{V}_D$ and the value image $\mathcal{V}_T$ are equivalent under Assumption 1 for discounted MDPs (see Lemmas 2 and 4 in Appendix B). Therefore, the optimal value function of a non-atomic constrained MDP is contained in the deterministic value image $\mathcal{V}_D$. Furthermore, the deterministic value image $\mathcal{V}_D$ is a convex set, even though each value function $V(\pi) \in \mathcal{V}_D$ is non-convex in policy π (see Lemmas 2 and 3 in Appendix B). These observations are summarized below.

Lemma 1 (Sufficiency of deterministic policies)**.** *For a non-atomic discounted MDP, the deterministic value image $\mathcal{V}_D$ is convex, and equals the value image $\mathcal{V}_T$, i.e., $\mathcal{V}_D = \mathcal{V}_T$.*

2.1 Zero Duality Gap

With the convexity of the deterministic value image $\mathcal{V}_D$ in hand, we next establish zero duality gap for Problem (1). We begin with a standard feasibility assumption.

Assumption 2 (Feasibility)**.** *There exists a deterministic policy $\tilde{\pi} \in \Pi$ and $\xi > 0$ such that $V_g(\tilde{\pi}) \geq \xi$.*

We dualize the constraint by introducing the dual variable $\lambda \in \mathbb{R}^+$ and the Lagrangian $L(\pi, \lambda) := V_r(\pi) + \lambda V_g(\pi)$. For a fixed λ, let $\Pi(\lambda)$ be the set of Lagrangian maximizers. The Lagrangian $L(\pi, \lambda)$ is equivalent to the value function $V_\lambda(\pi)$ associated with the combined reward/utility function $r_\lambda(s, a) = r(s, a) + \lambda g(s, a)$. The dual function $D(\lambda) := \max_{\pi \in \Pi} V_\lambda(\pi)$ is an upper bound of Problem (1), and the dual problem searches for the tightest primal upper bound

$$\min_{\lambda \,\in\, \mathbb{R}^+} \; D(\lambda). \tag{2}$$

We denote by $V_D^{\lambda^\star}$ the optimal value of the dual function, where $\lambda^\star$ is a minimizer of the dual Problem (2). Despite being non-convex in the policy, if we replace the deterministic policy space in Problem (1) with the stochastic policy space, then it is known that Problem (1) has zero duality gap (Paternain et al. 2019). The proof capitalizes on the convexity of the occupancy measure representation of (1) for stochastic policies. However, this occupancy-measure-based argument does not carry to deterministic policies, since the occupancy measure representation of Problem (1) is non-convex when only deterministic policies are used (Dolgov 2005). Instead, we leverage the convexity of the deterministic value image $\mathcal{V}_D$ to prove that strong duality holds for Problem (1); see Appendices A and C.2 for more details and the proof.

Theorem 1 (Zero duality gap)**.** *Let Assumption 1 hold. Then, Problem* (1) *has zero duality gap, i.e., $V_P^{\pi^\star} = V_D^{\lambda^\star}$.*

Theorem 1 states that the optimal values of Problems (1) and (2) are equivalent, extending the zero duality gap result in (Paternain et al. 2019) to deterministic policies under the non-atomicity assumption. However, recovering an optimal policy $\pi^\star$ can be non-trivial even if an optimal dual variable $\lambda^\star$ is obtained from the dual problem (Zahavy et al. 2021). The root cause is that the maximizers of the primal problem $\Pi^\star$ and those of the Lagrangian for an optimal multiplier $\Pi(\lambda^\star)$ are different sets (Calvo-Fullana et al. 2023, Proposition 1). To address this, we employ Theorem 1 to interpret Problem (1) as a saddle point problem. Zero duality gap implies that an optimal primal-dual pair $(\pi^\star, \lambda^\star)$ is a saddle point of the Lagrangian $L(\pi, \lambda)$, and satisfies the mini-max condition

$$L(\pi, \lambda^\star) \;\leq\; L(\pi^\star, \lambda^\star) \;\leq\; L(\pi^\star, \lambda) \quad \forall (\pi, \lambda) \in \Pi \times \Lambda,$$

where λ is bounded in the interval $\Lambda := [0, \lambda_{\text{max}}]$, with $\lambda_{\text{max}} := 1/((1-\gamma)\xi)$; see Lemma 9 in Appendix B. In this paper, we refer to saddle points that satisfy the mini-max condition for all pairs $(\pi, \lambda) \in \Pi \times \Lambda$ as *global* saddle points. Our main task in Section 3 is to find a global saddle point of the Lagrangian $L(\pi, \lambda)$ that is a solution to Problem (1).

2.2 Constrained Regulation Problem

We illustrate Problem (1) using the following example

$$\max_{\pi \in \Pi} \; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(s_t^\top G_1 s_t + a_t^\top R_1 a_t\right)\right] \tag{3a}$$

$$\text{s.t.} \;\; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(s_t^\top G_2 s_t + a_t^\top R_2 a_t\right)\right] \geq b \tag{3b}$$

$$-b_s \leq C_s s_t \leq b_s, \;\; -b_a \leq C_a a_t \leq b_a \tag{3c}$$

$$s_{t+1} = B_0 s_t + B_1 a_t + \omega_t, \;\; s_0 \sim \rho \tag{3d}$$

where $B_0 \in \mathbb{R}^{d_s \times d_s}$ and $B_1 \in \mathbb{R}^{d_s \times d_a}$ denote the system dynamics, ω_t is the standard Gaussian noise, ρ is the initial state distribution, and G_1, $G_2 \in \mathbb{R}^{d_s \times d_s}$ and R_1, $R_2 \in \mathbb{R}^{d_a \times d_a}$ are negative semi-definite reward matrices. The constraint threshold is b, with $C_s \in \mathbb{R}^{d_s \times d_s}$, $C_a \in \mathbb{R}^{d_a \times d_a}$, $b_s \in \mathbb{R}^{d_s}$, and $b_a \in \mathbb{R}^{d_a}$ specifying state-action constraints, e.g., if C_s, C_a are identity matrices, b_s, b_a limit state and action ranges. Equations (3a), (3c), and (3d) describe the constrained regulation problem under Gaussian disturbances (Bemporad et al. 2002; Stathopoulos, Korda, and Jones 2016), where the optimal policy is deterministic (Scokaert and Rawlings 1998). We add a general constraint (3b). The Markovian transition dynamics (3d) are linear, and the Gaussian noise ω_t is non-atomic, rendering the transition probabilities non-atomic. If ρ is non-atomic, the underlying MDP of (3) is also non-atomic. The reward function $r(s, a) := s^\top G_1 s + a^\top R_1 a$ induces a value function $V_r(\pi)$, bounded within $[r_{\text{min}}/(1-\gamma), 0]$, with $r_{\text{min}} := b_s^\top G_1 b_s + b_a^\top R_1 b_a$. Similarly, for $u(s, a) := s^\top G_2 s + a^\top R_2 a$, the utility value V_u is also bounded. Therefore, this problem is an instance of Problem (1), assuming the state space is bounded with $\|s\| \leq S_{\text{max}}$.

3 Method and Theory

While our problem has zero duality gap, finding an optimal dual $\lambda^\star$ poses a significant challenge, due to the presence of multiple saddle points in the Lagrangian. To address it, we resort to the regularization method. More specifically, we introduce two regularizers. First, the term $h(\lambda) := \lambda^2$ promotes convexity in the Lagrange multiplier λ. Second, the term $h_a(a) := -\|a\|^2$ promotes concavity in the reward function r by penalizing large actions selected by the policy π. The associated value function is defined as $H^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t h_a(a_t) \,|\, s\right]$, and leads to the regularizer $H(\pi) := \mathbb{E}_\rho[H^\pi(s)]$. Now, we consider the problem

$$\min_{\lambda\,\in\,\Lambda}\ \max_{\pi\,\in\,\Pi}\ \ L_\tau(\pi,\lambda) := V_\lambda(\pi) + \frac{\tau}{2}H(\pi) + \frac{\tau}{2}h(\lambda), \quad (4)$$

where $\tau \geq 0$ is the regularization parameter and $L_\tau(\pi,\lambda)$ is the regularized Lagrangian. For a fixed λ, the objective of Problem (4) is equivalent to an unconstrained regularized MDP plus a regularization of the dual variable. Consider the composite regularized reward function $r_{\lambda,\tau}(s,a) := r(s,a) + \lambda g(s,a) - \frac{\tau}{2}h_a(a)$. The value function associated with the reward function $r_{\lambda,\tau}$ can be expressed as $V_{\lambda,\tau}(\pi) = V_\lambda(\pi) + \frac{\tau}{2}H(\pi)$. Then, we can reformulate the regularized Lagrangian as $L_\tau(\pi,\lambda) := V_{\lambda,\tau}(\pi) + \frac{\tau}{2}\lambda^2$. The global saddle points of the regularized Lagrangian $\Pi_\tau^\star \times \Lambda_\tau^\star$ are guaranteed to exist; see Lemma 13 in Appendix C. Moreover, a global saddle point $(\pi_\tau^\star, \lambda_\tau^\star)$ satisfies

$$V_{\lambda_\tau^\star}(\pi) + \frac{\tau}{2}H(\pi) \leq V_{\lambda_\tau^\star}(\pi_\tau^\star) \leq V_\lambda(\pi_\tau^\star) + \frac{\tau}{2}\lambda^2 \quad (5)$$

for all $(\pi,\lambda) \in \Pi\times\Lambda$. Hence, $(\pi_\tau^\star, \lambda_\tau^\star)$ is also a global saddle point of the original Lagrangian $L(\pi,\lambda)$ up to two τ-terms.

3.1 Deterministic Policy Search Method

We propose a deterministic policy gradient primal-dual (D-PGPD) method for finding a global saddle point $(\pi_\tau^\star, \lambda_\tau^\star)$ of $L_\tau(\pi,\lambda)$. In the primal update, as is customary in RL, we maximize the advantage function rather than the value function directly. Specifically, we use the regularized advantage $A_{\lambda,\tau}^\pi(s,a) := Q_{\lambda,\tau}^\pi(s,a) - V_{\lambda,\tau}^\pi(s) - \frac{\tau}{2}(\|a\|^2 - \|\pi(s)\|^2)$ associated with the regularized reward $r_{\lambda,\tau}$. The primal update (6a) performs a proximal-point-type ascent step that solves a quadratic-regularized maximization sub-problem, while the dual update (6b) performs a gradient descent step that solves a quadratic-regularized minimization sub-problem

$$\pi_{t+1}(s) \;=\; \operatorname*{argmax}_{a\,\in\,A}\ A_{\lambda_t,\tau}^{\pi_t}(s,a) - \frac{1}{2\eta}\|a-\pi_t(s)\|^2 \quad (6a)$$

$$\lambda_{t+1} \;=\; \operatorname*{argmin}_{\lambda\,\in\,\Lambda}\ \lambda\,(V_g(\pi_t) + \tau\lambda_t) + \frac{1}{2\eta}\|\lambda-\lambda_t\|^2, \quad (6b)$$

where η is the step-size. D-PGPD is a single-time-scale algorithm, in the sense that the primal and the dual updates are computed concurrently in the same time-step. We remark that implementing D-PGPD is difficult in practice, and to make it tractable, we will leverage function approximation in Section 4. Before proceeding, we show that the primal-dual iterates (6) converge in the last iterate to the set of global saddle points of the regularized Lagrangian $\Pi_\tau^\star \times \Lambda_\tau^\star$.

3.2 Non-Asymptotic Convergence

Finding deterministic optimal policies is a computationally challenging problem (Feinberg 2000; Dolgov 2005). To render the problem tractable, we assume concavity and Lipschitz continuity of the regularized action value functions.

Assumption 3 (Concavity)**.** *The regularized state-action value function $Q_{\lambda,\tau}^\pi(s,a) - \tau_0\|\pi_0(s)-a\|^2$ is concave in action a for any policy π_0 and some $\tau_0 \in [0,\tau)$.*

Assumption 4 (Lipschitz continuity)**.** *The action-value functions $Q_r^\pi(s,a)$, $Q_g^\pi(s,a)$, and $H^\pi(s,a) := \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t h_a(a_t) \,|\, s_0 = s, a_0 = a\right]$ are Lipschitz in action a with Lipschitz constants L_r, L_g, and L_h, i.e.,*

$$\begin{aligned}
\|Q_r^\pi(s,a) - Q_r^\pi(s,a')\| &\leq L_r\|a-a'\| \\
\|Q_g^\pi(s,a) - Q_g^\pi(s,a')\| &\leq L_g\|a-a'\| \\
\|H^\pi(s,a) - H^\pi(s,a')\| &\leq L_h\|a-a'\|,\ \forall\, a,\, a' \in A.
\end{aligned}$$

Assumption 3 states that there exists a τ_0-strongly concave regularizer that renders $Q_{\lambda,\tau}^\pi$ concave in the action a. When $\tau_0 = 0$, $Q_{\lambda,\tau}^\pi$ is concave in the action a. An example of this is Problem (3), where the original reward and utility functions are concave and the transition dynamics are linear, leading to concavity of the associated regularized value function. Assumption 4 implies Lipschitz continuity of the reward function and the probability transition kernel, which holds for several dynamics that can be expressed as a deterministic function of the actual state-action pair and some stochastic perturbation; see Appendix D.1 for a detailed explanation over the example introduced in Section 2.2.

To show convergence of D-PGPD, we introduce first two projection operators. The operator $\mathcal{P}_{\Pi_\tau^\star}$ projects a policy into the non-empty set of optimal policies with state visitation distribution $d_\rho^\star$, and the operator $\mathcal{P}_{\Lambda_\tau^\star}$ projects a Lagrangian multiplier onto the non-empty set of optimal Lagrangian multipliers $\Lambda_\tau^\star$. Then, we characterize the convergence of the primal-dual iterates of D-PGPD using a potential function

$$\Phi_t := \frac{1}{2}\mathbb{E}_{d_\rho^\star}\left[\|\mathcal{P}_{\Pi_\tau^\star}(\pi_t(s)) - \pi_t(s)\|^2\right] + \frac{\|\mathcal{P}_{\Lambda_\tau^\star}(\lambda_t) - \lambda_t\|^2}{2(1+\eta(\tau-\tau_0))},$$

which measures the distance between a iteration pair (π_t,λ_t) of D-PGPD and the set of global saddle points of the regularized Lagrangian $\Pi_\tau^\star \times \Lambda_\tau^\star$. Theorem 2 shows that as t increases, the potential function Φ_t decreases linearly, up to an error; see Appendix C.4 for the proof.

Theorem 2 (Linear convergence)**.** *Let Assumptions 2–4 hold. For $\eta > 0$ and $\tau > \tau_0$, the primal-dual iterates* (6) *satisfy*

$$\Phi_{t+1} \;\leq\; \mathrm{e}^{-\beta_0\, t}\,\Phi_1 \;+\; \beta_1\, C_0^2, \quad \textit{where} \quad (7)$$

$$\beta_0 \;:=\; \frac{\eta(\tau-\tau_0)}{1+\eta(\tau-\tau_0)} \;\; \textit{and} \;\; \beta_1 \;:=\; \frac{\eta(1+\eta(\tau-\tau_0))}{\tau-\tau_0}$$

$$C_0 \;:=\; L_r + \lambda_{\max} L_g + \tau L_h + \tau\sqrt{d_a}A_{max} + \frac{1+\frac{\tau}{\xi}}{1-\gamma}.$$

Theorem 2 states that the primal-dual updates of D-PGPD converge to a neighborhood of the set of global saddle points of the regularized Lagrangian $\Pi_\tau^\star \times \Lambda_\tau^\star$ in a linear rate. The

size of the neighborhood depends polynomially on the parameters (L_r, L_g, L_h, $A_{\max}$, τ). When $\tau_0 = 0$, the regularization parameter τ can be arbitrarily small. Reducing the size of the convergence neighborhood can be achieved by selecting a sufficiently small η. However, a smaller the value of η leads to slower convergence. To be more specific, for $\eta = \epsilon(\tau - \tau_0)C_0^{-2}$, the size of the convergence neighborhood is $O(\epsilon)$, and when $t \geq \Omega(\epsilon^{-1}\log(\epsilon^{-1}))$, the potential function Φ_t is $O(\epsilon)$ too, where Ω encapsulates some problem-dependent constants. After $O(\epsilon^{-1})$ iterations, the primal-dual iterates (π_t, λ_t) of D-PGPD are ϵ-close to the set $\Pi_\tau^\star \times \Lambda_\tau^\star$.

The relationship between the solution to Problem (1) and the solution to the regularized Problem (4) is given by Corollary 1; see its proof in Appendix C.5.

Corollary 1 (Near-optimality)**.** *Let Assumptions 2–4 hold. If $\eta = O(\epsilon^4)$ and $\tau = O(\epsilon^2) + \tau_0$, and $t = \Omega(\epsilon^{-6}\log^2\epsilon^{-1})$, then the primal-dual iterates* (6) *satisfy*

$$\begin{aligned} V_r(\pi^\star) - V_r(\pi_t) \;&\leq\; \epsilon - \tau_0 H(\pi^\star) \\ V_g(\pi_t) \;&\geq\; -\epsilon + \tau_0 H(\pi^\star)(\lambda_{\max} - \lambda^\star)^{-1}. \end{aligned}$$

Corollary 1 highlights that the value functions corresponding to the policy iterates of D-PGPD can closely approximate the optimal solution to Problem (1). Specifically, in problems where $\tau_0 = 0$, the final policy iterate of D-PGPD achieves ϵ-optimality for Problem (1) after $\Omega(\epsilon^{-6})$ iterations. When $\tau_0 > 0$, D-PGDP converges to a saddle point of the original problem. However, the proximity of the final policy iterate to the optimal solution to Problem (1) is proportional to $H(\pi^\star)$.

This work presents the first primal-dual convergence result for general constrained RL problems that directly work with *deterministic* policies and *continuous* state-action spaces. In the context of control, the convergence of different algorithms for solving constrained problems has been analyzed (Stathopoulos, Korda, and Jones 2016; Zhang et al. 2020; Garg, Arabi, and Panagou 2020). However, these analyses are limited to linear utility functions and box constraints. D-PGPD is a general algorithm that can be used for a broad range of transition dynamics and cost functions.

4 Function Approximation

To instantiate D-PGPD (6) with function approximation we begin by expanding the objective in (6a) and dropping the terms that do not depend on the action a,

$$Q^\pi_{\lambda,\tau}(s,a) + \frac{1}{\eta}\pi(s)^\top a - \left(\frac{\tau}{2} + \frac{1}{2\eta}\right)\|a\|^2.$$

The usual function approximation approach (Agarwal et al. 2021; Ding et al. 2022) is to introduce a parametric estimator of the policy π, and a compatible parametric estimator of the action value function $Q^\pi_{\lambda,\tau}$. Instead, we approximate the augmented action-value function $J^\pi(s,a) := Q^\pi_{\lambda,\tau}(s,a) + \frac{1}{\eta}\pi(s)^\top a$ using a linear estimator $\tilde{J}_\theta(s,a) = \phi(s,a)^\top\theta$ over the basis ϕ. At time t, we estimate $J^{\pi_t}(s,a)$ by computing the parameters θ_t via a mean-squared-error minimization

$$\theta_t \;:=\; \operatorname*{argmin}_\theta \mathbb{E}_{(s,a)\sim\nu}\left[\,\|\phi(s,a)^\top\theta - J^{\pi_t}(s,a)\|^2\,\right], \quad (8)$$

where ν is a pre-selected state-action distribution. Problem (8) can be easily addressed using, e.g., stochastic approximation. A subsequent policy π_{t+1} results from a primal update based on $\tilde{J}_{\theta_t}$. This leads to an approximated D-PGPD algorithm (AD-PGPD) that updates π_t and λ_t via

$$\pi_{t+1}(s) \;=\; \operatorname*{argmax}_{a\,\in\,A}\; \tilde{J}_{\theta_t}(s,a) - \left(\frac{\tau}{2} + \frac{1}{2\eta}\right)\|a\|^2 \quad (9a)$$

$$\lambda_{t+1} \;=\; \operatorname*{argmin}_{\lambda\,\in\,\Lambda}\; \lambda(V_g(\pi_t) + \tau\lambda_t) + \frac{1}{2\eta}\|\lambda - \lambda_t\|^2. \quad (9b)$$

Solving the sub-problem (9a) requires inverting the gradient of (9a) with respect to a, which is a challenge when the MDP model is unknown or the value functions cannot be computed in closed form. This is the focus of Section 5.

4.1 Non-Asymptotic Convergence

To ease the computational tractability of AD-PGPD, we assume concavity of the approximated augmented action-value function and bounded approximation error.

Assumption 5 (Concavity of approximation)**.** *The function $\tilde{J}_{\theta_t}(s,a) - \tau_0\|\pi_0(s) - a\|^2$ is concave with respect to the action a for some arbitrary policy π_0 and some $\tau_0 \in [0, \tau)$.*

Assumption 6 (Approximation error)**.** *The approximation error $\delta_{\theta_t}(s,a)$ is bounded, $\mathbb{E}_{s\sim d^\star_\rho, a\sim\mathsf{u}}[\|\delta_{\theta_t}(s,a)\|] \leq \frac{\epsilon_{\text{approx}}}{2(2A_{max})^{d_a}}$, where u is the uniform distribution and $\epsilon_{\text{approx}} \geq 0$ is a positive error constant.*

The concavity of $\tilde{J}_{\theta_t}(s,a)$ with respect to a depends on the selection of the basis function ϕ. When the augmented action-value function J^{π_t} is a concave quadratic function, it can be represented as a weighted linear combination of concave and quadratic basis functions. If these basis functions are known, J^{π_t} can be perfectly approximated, i.e., $\epsilon_{\text{approx}} = 0$. Furthermore, when J^{π_t} is concave with respect to the action a, the regularization parameter τ can be arbitrarily small.

Theorem 3 (Linear convergence)**.** *Let Assumptions 2, 4– 6 hold. If $\eta > 0$ and $\tau > \tau_0$, the primal-dual iterates* (9) *satisfy*

$$\Phi_{t+1} \;\leq\; \mathrm{e}^{-\beta_0 t}\Phi_1 \;+\; \beta_1 C_0^2 \;+\; \beta_2\epsilon_{\text{approx}}, \quad (10)$$

where β_0, β_1, and C_0 are defined in Theorem 2, and

$$\beta_2 \;:=\; \frac{1 + \eta(\tau - \tau_0)}{\tau - \tau_0}.$$

Theorem 3 shows that the primal-dual iterates of AD-PGPD converge to a neighborhood of $\Pi_\tau^\star \times \Lambda_\tau^\star$ at a linear rate. The result is similar to Theorem 2, up to an approximation error ϵ_{approx}. In fact, when $\epsilon_{\text{approx}} = 0$, Theorem 3 is equivalent to Theorem 2. Linear models can achieve $\epsilon_{\text{approx}} = 0$ when the augmented action-value function J^{π_t} can be expressed as a linear combination of the selected basis function ϕ, e.g. when J^{π_t} is convex. When the error is small, the following result relates Problem (1) to the regularized Problem (4).

Corollary 2 (Near-optimality of approximation)**.** *Let Assumptions 2 and 4–6 hold. If $\eta = O(\epsilon^4)$, $\tau = O(\epsilon^2) + \tau_0$, $\epsilon_{\text{approx}} = O(\epsilon^4)$, and $t = \Omega(\epsilon^{-6}\log^2\epsilon^{-1})$, then the primal-dual iterates* (9) *satisfy*

$$\begin{aligned} V_r(\pi^\star) - V_r(\pi_t) \;&\leq\; \epsilon - \tau_0 H(\pi^\star) \\ V_g(\pi_t) \;&\geq\; -\epsilon + \tau_0 H(\pi^\star)(\lambda_{\max} - \lambda^\star)^{-1}. \end{aligned}$$

Corollary 2 states that Corollary 1 extends to function approximation. When the approximation error is sufficiently small, i.e., $\epsilon_{\text{approx}} = O(\epsilon^4)$, the proof of Corollary 1 holds (see Appendix C.5), and the value functions corresponding to the policy iterates of AD-PGPD closely approximate an optimal solution to Problem (1). In fact, when $\tau_0 = 0$ and ϵ_{approx} are small, then the last policy iterate of AD-PGPD is an ϵ-optimal solution to Problem (1) after $\Omega(\epsilon^{-6})$ iterations.

5 Model-Free Algorithm

When the model of the MDP is unknown or when value-functions cannot be computed in closed form, we can leverage sample-based approaches to compute the primal and dual iterates of AD-PGPD. To that end, we assume access to a simulator of the MDP from where we can sample trajectories given a policy π. The sample-based algorithm requires modifying the policy evaluation step in (8), and the dual update in (9b). For the former, in time-step t for a given policy π_t, we have the following linear function approximation problem

$$\min_{\theta,\, \|\theta\| \leq \theta_{\max}} \; \mathbb{E}_{s,a\sim\nu}\left[\|\phi(s_n,a_n)^\top\theta - \hat{J}^{\pi_t}(s_n,a_n)\|^2\right], \quad (11)$$

where the parameters θ are bounded, i.e., $\|\theta\| \leq \theta_{\max}$, and ϕ is the basis function. The approximated augmented value-function $\hat{J}^{\pi_t} := \hat{Q}^{\pi_t}_{\lambda,\tau}(s_n,a_n) + \frac{1}{\eta}\pi(s_n)^\top a_n$ is estimated from samples, which comes down to approximating $\hat{Q}^{\pi_t}_{\lambda,\tau}(s_n,a_n)$. The dual update (9b) also requires the approximated value-function $\hat{V}_g(\pi_t)$ to be estimated. We detail how to estimate $\hat{V}_g(\pi_t)$ and $\hat{Q}^{\pi_t}_{\lambda,\tau}(s_n,a_n)$ via rollouts in Algorithms 1 and 2, which can be found in Appendix E. We use random horizon rollouts (Paternain et al. 2020; Zhang et al. 2020) to guarantee that the stochastic estimates of $\hat{Q}^{\pi_t}_{\lambda,\tau}$ and $\hat{V}_g(\pi_t)$ are unbiased. From (Paternain et al. 2020, Proposition 2), we have $Q^{\pi_t}_{\lambda,\tau}(s,a) = \mathbb{E}[\hat{Q}^{\pi_t}_{\lambda,\tau}(s,a)\,|\,s,a]$ and $V_g(\pi_t) = \mathbb{E}[\hat{V}^{\pi_t}_g(s)]$, where the expectations $\mathbb{E}$ are taken over the randomness of drawing trajectories following π_t. We solve Problem (11) at time t using projected stochastic gradient descent (SGD),

$$\begin{aligned} g_t^{(n)} &= 2\left(\phi(s_n,a_n)^\top\theta_t^{(n)} - \hat{J}^{\pi_t}(s_n,a_n)\right)\phi(s_n,a_n) \\ \theta_t^{(n+1)} &= \mathcal{P}_{\|\theta\|\leq\theta_{\max}}\left(\theta_t^{(n)} - \alpha_n\, g_t^{(n)}\right), \end{aligned} \quad (12)$$

where $n \geq 0$ is the iteration index, α_n is the step-size, $g_t^{(n)}$ is the stochastic gradient of (11), and $\mathcal{P}_{\|\theta\|\leq\theta_{\max}}$ is an operator that projects onto the domain $\|\theta\| \leq \theta_{\max}$, which is convex and bounded. Each projected SGD update (12) forms the estimate $\hat{\theta}_t$. We run N projected SGD iterations and form the weighted average $\hat{\theta}_t := \frac{2}{N(N+1)}\sum_{n=0}^{N-1}(n+1)\hat{\theta}_t$, which is the estimation of the parameters θ_t. Combining (9), the SGD rule in (12), and averaging techniques lead to a sample-based algorithm presented in Algorithm 3, in Appendix E.

The convergence analysis of Algorithm 3 has to account for the estimation error induced by the sampling process. The error $\delta_{\hat{\theta}_t}(s,a) = \tilde{J}_{\hat{\theta}_t}(s,a) - J^{\pi_t}(s,a)$ can be decomposed as $\delta_{\hat{\theta}_t}(s,a) = \delta_{\hat{\theta}_t}(s,a) - \delta_{\theta_t}(s,a) + \delta_{\theta_t}(s,a)$. The bias error term $\delta_{\theta_t}(s,a)$ is similar to the approximation error of AD-PGPD and captures how good the model approximates the true augmented value function. The term $\delta_{\hat{\theta}_t}(s,a) - \delta_{\theta_t}(s,a)$ is a statistical error that reflects the error introduced by the sampling mechanism for a given state-action pair. To deal with the randomness of the projected SGD updates, we assume that the bias error and the feature basis are bounded. We also assume that the feature covariance matrix is positive definite, and that the sampling distribution ν and the optimal state visitation frequency $d^\star_\rho$ are uniformly equivalent.

Assumption 7 (Bounded feature basis)**.** *The feature function is bounded, i.e., $\|\phi(s,a)\| \leq 1$ for all $s \in S$ and $a \in A$.*

Assumption 8 (Positive covariance)**.** *The feature covariance matrix $\Sigma_\nu = \mathbb{E}_{s,a\sim\nu}[\phi(s,a)\phi(s,a)^\top]$ is positive definite $\Sigma_\nu \geq \kappa_0 I$ for the state-action distribution ν.*

Assumption 9 (Bias error)**.** *The bias error $\delta_{\theta_t}(s,a)$ is bounded $\mathbb{E}_{s\sim d^\star_\rho, a\sim\mathsf{u}}[\|\delta_{\theta_t}(s,a)\|] \leq \frac{\epsilon_{\text{bias}}}{2(2A_{max})^{d_a}}$, where u is the uniform distribution and ϵ_{bias} is a positive error constant.*

Assumption 10 (Uniformly equivalence)**.** *The state-action distribution induced by the state-visitation frequency $d^\star_\rho$ and the uniform distribution u is uniformly equivalent to the state-action distribution ν, i.e.*

$$\frac{d^\star_\rho(s)\mathsf{u}(a)}{\nu(s,a)} \;\leq\; L_\nu \;\; \textit{for all } (s,a) \in S \times A.$$

Assumption 7 holds without loss of generality, as the basis functions are a design choice. Assumption 8 ensures that the minimizer of (11) is unique, since $\Sigma_\nu \geq \kappa_0 I$ for some $\kappa_0 > 0$. Assumption 9 states that the selected model achieves a bounded error, and Assumption 10 ensures that the sampling distribution ν is sufficiently representative of the optimal state visitation frequency $d^\star_\rho$. We characterize the convergence using the expected potential function $\mathbb{E}[\Phi_t]$, where the expectation is taken over the randomness of $\theta_t^{(n)}$. We have the following corollary; see the proof in Appendix C.7.

Corollary 3 (Linear convergence)**.** *Let Assumptions 2, 4, 5, and 7–10 hold. Then, the sample-based AD-PGPD in Algorithm 3 satisfies*

$$\mathbb{E}[\Phi_{t+1}] \leq e^{-\beta_0 t}\mathbb{E}[\Phi_1] + \beta_1 C_0^2 + \beta_2\left(\frac{C_1^2}{\eta^2(N+1)} + \epsilon_{\text{bias}}\right), \quad (13)$$

where β_0, β_1, β_2, and C_0 are given in Theorems 2 and 3, and

$$C_1 := \sqrt{2^{d_a+5}A_{max}^{d_a}L_\nu}\big(\theta_{max} + 2(1-\gamma)^{-2}\xi^{-1} + d_a A_{max}^2\big)\kappa_0^{-1}.$$

Corollary 3 is analogous to Theorem 3, but accounting for the use of sample-based estimates. The sampling effect appears as the number N of projected SGD steps performed at each time-step t. Corollary 2 holds when the bias error $\epsilon_{\text{bias}} = O(\epsilon^4)$ and the estimation error $C_1^2\eta^{-2}(N+1)^{-1} = O(\epsilon^4)$. As $\eta = O(\epsilon^4)$, the latter holds when $N = \Omega(\epsilon^{-12})$, where Ω encapsulates problem-dependent constants. Therefore, the number of rollouts required to output an ϵ-optimal policy is $tN = \Omega(\epsilon^{-18})$. While this result suggests potential improvement, it stands as the first sample-complexity result in the context of constrained MDPs with continuous spaces.

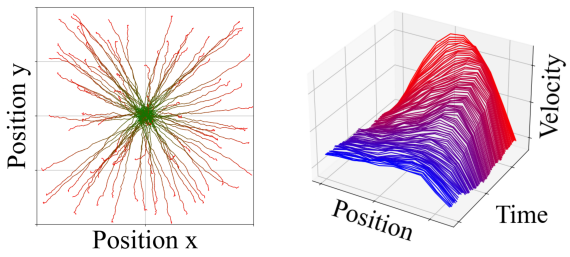

Figure 1: Navigation trajectories of an agent (Left) and velocity profile of the fluid over time (Right).

6 Computational Experiments

We test D-PGPD on constrained robot navigation and fluid control problems (Figure 1). See Appendix F for more details.

Navigation Problem. An agent moves in a horizontal plane following some linearized dynamics with zero-mean Gaussian noise (Shimizu et al. 2020; Ma et al. 2022). We aim to drive the agent to the origin while constraining its velocity. When the dynamics are known and the reward function linearly weights quadratic penalties on position and action, this problem is an instance of the constrained linear regulation problem (Scokaert and Rawlings 1998), which has closed-form solution. Hence, we can directly apply D-PGPD (6) and AD-PGPD (9) (See Appendix F.1). However, we consider the dynamics to be unknown, and we leverage our sample-based implementation of AD-PGPD. Furthermore, we use absolute value penalties instead of quadratic ones, as the latter can result in unstable behavior in sample-based scenarios (Engel and Babuška 2014). Conventional methods do not solve this problem straightforwardly. We compare our sample-based AD-PGPD with PGDual, a dual method with linear function approximation (Zhao and You 2021; Brunke et al. 2022). Figure 2 shows the value functions of the policy iterates generated by AD-PGPD and PGDual over $40,000$ iterations. The oscillations of AD-PGPD are damped over time, and it converges to a feasible solution with low variance in reward and utility, indicating a near-deterministic behavior without constraint violation. In contrast, PGDual exhibits large variance, indicating that the resultant policy violates the constraint. Nevertheless, the final primal return performance of PGDual is similar to that of AD-PGPD on average.

Fluid Velocity Control. We apply D-PGPD (6) to the control of the velocity of an incompressible Newtonian fluid described by the one-dimensional Burgers' equation (Baker, Armaou, and Christofides 2000), a non-linear stochastic control problem. The velocity profile of the fluid z varies in a one-dimensional space $x \in [0, 1]$ and time $t \in [0, 1]$, and the goal is to drive the velocity of the fluid towards zero via the control action a, e.g., injection of polymers. By discretizing Burgers' equation, we have a non-linear system $s_{t+1} = B_0 s_t + B_1 a_t + B_2 s_t^2 + \omega_t$, where $s_t \in \mathbb{R}^d$ is the state, s_t^2 is the element-wise squared state vector, $a_t \in \mathbb{R}^d$ is the control input, and B_0, B_1, $B_2 \in \mathbb{R}^{d \times d}$ are matrices representing the discretized spatial operators and non-linear

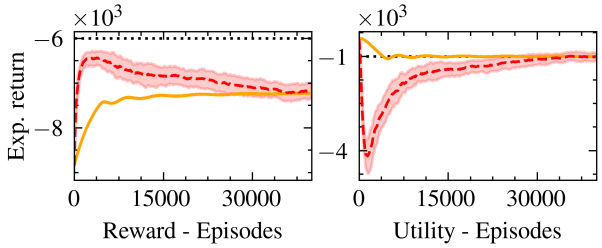

Figure 2: Avg. reward/utility value functions of AD-PGPD (—) and PGDual (– –) iterates in the navigation problem.

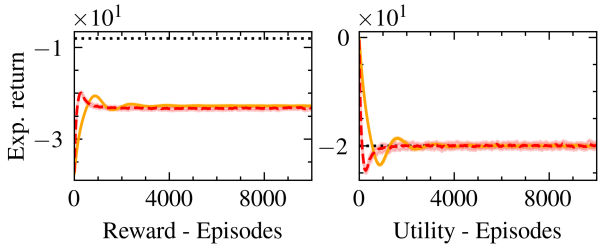

Figure 3: Avg. reward/utility value functions of AD-PGPD (—) and PGDual (– –) iterates in a fluid velocity control.

terms (Borggaard and Zietsman 2020). The details can be found in Appendix F. We consider a reward function that penalizes the state quadratically, and a budget constraint that limits the total control action. We compare our sample-based AD-PGPD with PGDual. Figure 3 shows the value functions of the policy iterates generated by AD-PGPD and PGDual over $10,000$ iterations. The results are consistent with those of the navigation problem. The AD-PGPD algorithm successfully mitigates oscillations and converges to a feasible solution with low return variance. In contrast, although PGDual achieves similar objective value, it does not dampens oscillations, as indicated by the variance of the solution. This implies that PGDual violates the constraint in the last iterate.

7 Concluding Remarks

We have presented a deterministic policy gradient primal-dual method for continuous state-action constrained MDPs with non-asymptotic convergence guarantees. We have leveraged function approximation to make the implementation practical and developed a sample-based algorithm. Furthermore, we have shown the effectiveness of the proposed method in navigation and non-linear fluid constrained control problems. Our work opens new avenues for constrained MDPs with continuous state-action spaces, such as (i) minimal assumption on value functions; (ii) online exploration; (iii) optimal sample complexity; and (iv) general function approximation.

Appendix

All the theoretical proofs and additional materials referenced in this paper, the supplementary experiments and introductions to key concepts are included in the extended version of the paper, available at https://arxiv.org/abs/2408.10015.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work has been partially supported by the Spanish NSF (AEI/10.13039 /501100011033) grants TED2021-130347B-I00 and PID2022-136887NB-I00, and the Community of Madrid via the Ellis Madrid Unit and grant TEC-2024/COM-89.

References

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained Policy Optimization. In *International Conference on Machine Learning*, 22–31.

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. 2021. *Constrained Markov Decision Processes*. Routledge.

Anderson, B. D.; and Moore, J. B. 2007. *Optimal Control: Linear Quadratic Methods*. Courier Corporation.

Baker, J.; Armaou, A.; and Christofides, P. D. 2000. Nonlinear Control of Incompressible Fluid Flow: Application to Burgers' Equation and 2D Channel Flow. *Journal of Mathematical Analysis and Applications*, 252(1): 230–255.

Bemporad, A.; Morari, M.; Dua, V.; and Pistikopoulos, E. N. 2002. The Explicit Linear Quadratic Regulator for Constrained Systems. *Automatica*, 38(1): 3–20.

Borggaard, J.; and Zietsman, L. 2020. The Quadratic-Quadratic Regulator Problem: Approximating Feedback Controls for Quadratic-in-State Nonlinear Systems. In *American Control Conference*, 818–823.

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

Brunke, L.; Greeff, M.; Hall, A. W.; Yuan, Z.; Zhou, S.; Panerati, J.; and Schoellig, A. P. 2022. Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5: 411–444.

Calvo-Fullana, M.; Paternain, S.; Chamon, L. F.; and Ribeiro, A. 2023. State Augmented Constrained Reinforcement Learning: Overcoming the Limitations of Learning with Rewards. *IEEE Transactions on Automatic Control*.

Chow, Y.; Ghavamzadeh, M.; Janson, L.; and Pavone, M. 2018. Risk-Constrained Reinforcement Learning with Percentile Risk Criteria. *Journal of Machine Learning Research*, 18(167): 1–51.

Ding, D.; Huan, Z.; and Ribeiro, A. 2024. Resilient Constrained Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 3412–3420.

Ding, D.; Wei, C.-Y.; Zhang, K.; and Ribeiro, A. 2024. Last-Iterate Convergent Policy Gradient Primal-Dual Methods for Constrained MDPs. *Advances in Neural Information Processing Systems*, 36.

Ding, D.; Zhang, K.; Duan, J.; Başar, T.; and Jovanović, M. R. 2022. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. *arXiv preprint arXiv:2206.02346*.

Dolgov, D. 2005. Stationary Deterministic Policies for Constrained MDPs with Multiple Rewards, Costs, and Discount Factors. In *International Joint Conference on Artificial Intelligence*.

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

Engel, J.-M.; and Babuška, R. 2014. On-Line Reinforcement Learning for Nonlinear Motion Control: Quadratic and Non-Quadratic Reward Functions. *IFAC Proceedings Volumes*, 47(3): 7043–7048.

Feinberg, E. A. 2000. Constrained Discounted Markov Decision Processes and Hamiltonian Cycles. *Mathematics of Operations Research*, 25(1): 130–140.

Feinberg, E. A.; and Piunovskiy, A. 2019. Sufficiency of Deterministic Policies for Atomless Discounted and Uniformly Absorbing MDPs with Multiple Criteria. *SIAM Journal on Control and Optimization*, 57(1): 163–191.

Feinberg, E. A.; and Piunovskiy, A. B. 2002. Nonatomic Total Rewards Markov Decision Processes with Multiple Criteria. *Journal of Mathematical Analysis and Applications*, 273(1): 93–111.

Gao, X.; Yan, L.; Li, Z.; Wang, G.; and Chen, I.-M. 2023. Improved Deep Deterministic Policy Gradient for Dynamic Obstacle Avoidance of Mobile Robot. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(6): 3675–3682.

Garg, K.; Arabi, E.; and Panagou, D. 2020. Prescribed-Time Convergence with Input Constraints: A Control Lyapunov Function Based Approach. In *American Control Conference*, 962–967.

Kakade, S.; Krishnamurthy, A.; Lowrey, K.; Ohnishi, M.; and Sun, W. 2020. Information Theoretic Regret Bounds for Online Nonlinear Control. *Advances in Neural Information Processing Systems*, 33: 15312–15325.

Kumar, H.; Kalogerias, D. S.; Pappas, G. J.; and Ribeiro, A. 2020. Zeroth-Order Deterministic Policy Gradient. *arXiv preprint arXiv:2006.07314*.

Lan, G. 2022. Policy Optimization over General State and Action Spaces. *arXiv preprint arXiv:2211.16715*.

Li, G.; Li, S.; Li, S.; and Qu, X. 2022. Continuous Decision-Making for Autonomous Driving at Intersections Using Deep Deterministic Policy Gradient. *IET Intelligent Transport Systems*, 16(12): 1669–1681.

Li, Z.; Liu, B.; Yang, Z.; Wang, Z.; and Wang, M. 2023. Double Duality: Variational Primal-Dual Policy Optimization for Constrained Reinforcement Learning. *Journal of Machine Learning Research*, 24(385): 1–43.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous Control with Deep Reinforcement Learning. *arXiv preprint arXiv:1509.02971*.

Lim, A. E.; and Zhou, X. Y. 1999. Stochastic Optimal LQR Control with Integral Quadratic Constraints and Indefinite Control Weights. *IEEE Transactions on Automatic Control*, 44(7): 1359–1369.

Ma, J.; Cheng, Z.; Zhang, X.; Tomizuka, M.; and Lee, T. H. 2022. Alternating Direction Method of Multipliers for Constrained Iterative LQR in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 23031–23042.

Mandhane, A.; Zhernov, A.; Rauh, M.; Gu, C.; Wang, M.; Xue, F.; Shang, W.; Pang, D.; Claus, R.; Chiang, C.-H.; et al. 2022. MuZero with Self-Competition for Rate Control in VP9 Video Compression. *arXiv preprint arXiv:2202.06626*.

McMahan, J. 2024. Deterministic Policies for Constrained Reinforcement Learning in Polynomial-Time. *arXiv preprint arXiv:2405.14183*.

Montenegro, A.; Mussi, M.; Metelli, A. M.; and Papini, M. 2024a. Learning Optimal Deterministic Policies with Stochastic Policy Gradients. In *International Conference on Machine Learning*.

Montenegro, A.; Mussi, M.; Papini, M.; and Metelli, A. M. 2024b. Last-iterate global convergence of policy gradients for constrained reinforcement learning. *arXiv preprint arXiv:2407.10775*.

Moskovitz, T.; O'Donoghue, B.; Veeriah, V.; Flennerhag, S.; Singh, S.; and Zahavy, T. 2023. Reload: Reinforcement Learning with Optimistic Ascent-Descent for Last-Iterate Convergence in Constrained MDPs. In *International Conference on Machine Learning*, 25303–25336.

Paternain, S.; Bazerque, J. A.; Small, A.; and Ribeiro, A. 2020. Stochastic Policy Gradient Ascent in Reproducing Kernel Hilbert Spaces. *IEEE Transactions on Automatic Control*, 66(8): 3429–3444.

Paternain, S.; Calvo-Fullana, M.; Chamon, L. F.; and Ribeiro, A. 2022. Safe Policies for Reinforcement Learning via Primal-Dual Methods. *IEEE Transactions on Automatic Control*, 68(3): 1321–1336.

Paternain, S.; Chamon, L.; Calvo-Fullana, M.; and Ribeiro, A. 2019. Constrained Reinforcement Learning Has Zero Duality Gap. *Advances in Neural Information Processing Systems*, 32.

Posa, M.; Kuindersma, S.; and Tedrake, R. 2016. Optimization and Stabilization of Trajectories for Constrained Dynamical Systems. In *IEEE International Conference on Robotics and Automation*, 1366–1373.

Ross, K. W. 1989. Randomized and Past-Dependent Policies for Markov Decision Processes with Multiple Constraints. *Operations Research*, 37(3): 474–477.

Scokaert, P. O.; and Rawlings, J. B. 1998. Constrained Linear Quadratic Regulation. *IEEE Transactions on Automatic Control*, 43(8): 1163–1169.

Sehnke, F.; Osendorfer, C.; Rückstieß, T.; Graves, A.; Peters, J.; and Schmidhuber, J. 2010. Parameter-exploring policy gradients. *Neural Networks*, 23(4): 551–559.

Shimizu, Y.; Zhan, W.; Sun, L.; Chen, J.; Kato, S.; and Tomizuka, M. 2020. Motion Planning for Autonomous Driving with Extended Constrained Iterative LQR. In *Dynamic Systems and Control Conference*, volume 84270, V001T12A001.

Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; and Riedmiller, M. 2014. Deterministic Policy Gradient Algorithms. In *International Conference on Machine Learning*, 387–395.

Singh, R.; Gupta, A.; and Shroff, N. B. 2022. Learning in constrained Markov decision processes. *IEEE Transactions on Control of Network Systems*, 10(1): 441–453.

Stathopoulos, G.; Korda, M.; and Jones, C. N. 2016. Solving the Infinite-Horizon Constrained LQR Problem Using Accelerated Dual Proximal Methods. *IEEE Transactions on Automatic Control*, 62(4): 1752–1767.

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward Constrained Policy Optimization. *arXiv preprint arXiv:1805.11074*.

Tsiamis, A.; Kalogerias, D. S.; Chamon, L. F.; Ribeiro, A.; and Pappas, G. J. 2020. Risk-Constrained Linear-Quadratic Regulators. In *IEEE Conference on Decision and Control*, 3040–3047. IEEE.

Zahavy, T.; O'Donoghue, B.; Desjardins, G.; and Singh, S. 2021. Reward is Enough for Convex MDPs. *Advances in Neural Information Processing Systems*, 34: 25746–25759.

Zhang, K.; Koppel, A.; Zhu, H.; and Basar, T. 2020. Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612.

Zhao, F.; and You, K. 2021. Primal-Dual Learning for the Model-Free Risk-Constrained Linear Quadratic Regulator. In *Learning for Dynamics and Control*, 702–714.

Zhao, F.; You, K.; and Başar, T. 2021. Infinite-Horizon Risk-Constrained Linear Quadratic Regulator with Average Cost. In *IEEE Conference on Decision and Control*, 390–395.