A Primal-dual Approach for Semi-Infinitely Constrained Reinforcement Learning

Di Wang[1] **Liangyu Zhang**[2,1*] **Haishan Ye**[3,1]
Guang Dai[1] **Ivor Tsang**[4]
[1]SGIT AI Lab, State Grid Corporation of China
[2]Shanghai University of Finance and Economics
[3]Xi'an Jiaotong University
[4]IAIC, Agency for Science, Technology and Research, Singapore

Abstract

We propose a primal-dual policy optimization method for reinforcement learning in semi-infinitely constrained Markov decision process (SICMDP), which extends standard constrained Markov decision processes by allowing a continuum of constraints. We call our proposed method **S**emi-**I**nfinitely **P**rimal-**D**ual **P**olicy **O**ptimization (SI-PDPO). By introducing an infinite-dimensional Lagrange multiplier, we derive the saddle-point formulation of the original SICMDP problem and prove strong duality under mild conditions. We then introduce a regularization term for the dual variable, and solve the resulting regularized saddle-point problem using first-order methods with a policy optimization subroutine (for example, NPG or PPO). We also bound the regularization bias and establish finite-time convergence guarantees when natural policy gradient is used to update the policy. Unlike existing primal-type policy optimization algorithms for semi-infinitely constrained reinforcement learning, our method does not require solving an inner-loop optimization problem, thereby mitigating computational intractability and reducing the over-conservative bias. We perform a series of numerical experiments, including a real-world power control task in wireless communications, to evaluate the performance of SI-PDPO. Empirical results show that SI-PDPO consistently outperforms existing primal-type algorithms.

1 Introduction

Reinforcement learning (RL) has achieved remarkable success in many fields, such as game-playing Silver et al. [2017], Vinyals et al. [2019], language models Achiam et al. [2023], Christiano et al. [2017], robotics Levine et al. [2016], Ibarz et al. [2021] and wireless communications Mastronarde and van der Schaar [2011], Foerster et al. [2016]. However, RL systems often struggle to address strict safety or resource constraints in real-world tasks Garcıa and Fernández [2015]. To address this limitation, a constrained Markov Decision Process (CMDP) framework Altman [2021] explicitly incorporates such constraints into the agent's decision-making process. The CMDP framework models the sequential decision-making problem as a constrained optimization task in which the agent seeks to maximize expected cumulative reward while satisfying constraints formulated as upper bounds on expected cumulative costs.

However, problems in complex real-world dynamical systems often involve infinitely many constraints (e.g., time-spatial or noise-dependent constraints), which cannot be modeled by traditional CMDP designed for handling problems with a finite set of constraints. To bridge this gap, the Semi-Infinitely Constrained Markov Decision Process (SICMDP) framework Zhang et al. [2024] generalizes the

*Corresponding author: zhangliangyu@sufe.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

CMDP framework by integrating a continuum of constraints. The constraints are no longer indexed by $i \in \{1, ..., N\}$, but a continuous variable $y \in Y \subset \mathbb{R}^m$. This framework has demonstrated practical utility in domains such as environmental science and maritime studies.

An existing policy optimization algorithm for this setting is Semi-Infinitely Constrained Policy Optimization (SI-CPO) in Zhang et al. [2024]. In each iteration, SI-CPO first determines if the policy is feasible. If the policy is feasible, it updates the policy using a policy gradient oracle to maximize the expected cumulative reward. Otherwise, it updates the policy to minimize the expected cumulative cost associated with the most violated constraint. Hence, in each iteration SI-CPO solves an inner-loop optimization problem w.r.t. the constraint index $y \in Y$ to (i) assess the feasibility of the current policy and (ii), if infeasible, identify the most violated constraint to guide the update. This introduces two drawbacks. **1)** From a computational perspective, the inner-loop optimization problem is generally intractable without additional regularity assumptions like convexity or smoothness. And we have to rely on heuristic methods like random search to solve the problem; **2)** From a statistical perspective, the constraint functions—namely, the expected cumulative costs—must be estimated from data. When the estimation error is significant, maximizing a noisy approximation of the constraint function can lead SI-CPO to overestimate constraint violations and misclassify a feasible policy as infeasible, resulting in overly conservative policies. We call this phenomenon the over-conservative bias. For infeasible policy, SI-CPO can misidentify the most violated constraint, particularly when the constraint function is multi-modal in y.

In this paper, we propose the first primal-dual policy optimization algorithm for RL in SICMDPs called **S**emi-**I**nfinitely **P**rimal-**D**ual **P**olicy **O**ptimization (SI-PDPO). By introducing an infinite-dimensional dual variable, we derive the dual formulation of the original SICMDP problem, transforming it into a saddle-point optimization problem. We show the saddle-point problem is equivalent to the SICMDP problem by establishing strong duality results under mild conditions. We also add a regularizer for the dual variable to stabilize the optimization process and encourage a "smoothed" dual variable that assigns positive weight to all constraints. We solve the regularized saddle-point problem with stochastic first-order methods customized for the structure of our objective. At each iteration, the gradient of the primal variable (i.e., the parameterized policy) is estimated using a combination of a policy gradient subroutine—such as Natural Policy Gradient (NPG) or Proximal Policy Optimization (PPO)—and Monte Carlo approximation. A subsequent gradient ascent step is then performed using this estimate. For the dual variable, a mirror descent update is applied, with the Kullback–Leibler (KL) divergence serving as the associated Bregman divergence.

Unlike the SI-CPO focusing on extreme constraint violations, our primal-dual method avoids solving the inner-loop optimization problem and balances between different constraints by continuously reweighting them using the smoothed infinite-dimensional dual variables. Computationally, the reweighting scheme can be efficiently approximated with Monte-Carlo techniques. Statistically, such a reweighting scheme can smooth out the estimation error and effectively reduce the overestimation bias mentioned above. And we elaborate on this issue in Remark 1.

Contributions. We summarize our main contributions as follows:

1) We propose SI-PDPO, a primal-dual algorithm for SICMDPs that avoids inner-loop maximization and mitigates the over-conservative bias of SI-CPO.

2) We prove strong duality under Slater's condition and bound the error introduced by dual regularization. When the policy is updated by natural policy gradient, we further establish finite-time bounds on the reward gap and constraint violation of SI-PDPO.

3) We validate SI-PDPO on the discharge of sewage task and a real-world power control task in communication systems Zhang and Liang [2020], Liang et al. [2023].

1.1 Related Works

Policy Optimization in Constrained Markov Decision Processes Constrained Markov Decision Processes (CMDPs) are widely used in domains like autonomous drivingFisac et al. [2018], wireless communicationsDjonin and Krishnamurthy [2007], and financeBäuerle and Rieder [2011]. Policy optimization algorithms developed for CMDPs can be broadly categorized into primal and primal-dual methods.

Primal-type methods directly optimize the objective value function, typically by modifying the objectives or the update rule to ensure feasibility. Notable examples include Interior-point Policy Optimization (IPO)Liu et al. [2020] that incorporates barrier functions into the objective to softly enforce constraints; Lyapunov-based methods Chow et al. [2018] that construct Lyapunov functions to impose local linear constraints and guarantee global safety, and Constraint-Rectified Policy Optimization (CRPO)Xu et al. [2021] that alternates between objective maximization and constraint reduction. Zhang et al. [2024] propose the SICMDP model, which generalizes the classical CMDP by allowing a continuum of constraints. They also introduce SI-CPO, a primal-type policy optimization method that adopts an alternation strategy similar to that of CRPO. Zhang et al. [2025] propose EPO, which adaptively constructs a finite working set by adding violated constraints and removing inactive ones. SI-PDPO instead updates a multiplier measure over the continuum directly.

Primal-dual methods formulate the constrained problem as a saddle-point problem via Lagrangian relaxation. Paternain et al. [2019] establish strong duality for CMDPs, and subsequent works develop convergence guarantees for primal-dual policy optimization Tessler et al. [2018], Ding et al. [2020, 2023], Li et al. [2024]. Extensions to zero-violation Bai et al. [2022], online Müller et al. [2024], Stradi et al. [2024], and offline Le et al. [2019] settings have also been studied. This work generalizes primal-dual policy optimization from CMDPs to SICMDPs, requiring the treatment of infinite-dimensional Lagrange multipliers as bounded signed measures.

Semi-infinite programming Semi-infinite programming (SIP) was first introduced in Remez [1934], and has since been extensively applied in areas such as physics Carrijo et al. [2021], finance Daum and Werner [2011] and engineering Hettich and Kortanek [1993]. The problem of RL in SICMDPs can be formulated as a SIP problem. A range of numerical techniques have been developed to address SIP problems, such as discretization methods Betro [2004], Reemtsen [1991], exchange methods Campi and Garatti [2008], Zhang et al. [2010], local reduction strategies Gramlich et al. [1995] and primal-dual methods Qi et al. [2009], Wei et al. [2020]. Among these, primal-dual methods Ochoa and de Serio [2012], Wei et al. [2020] are most relevant to our work. Our algorithmic framework builds on Wei et al. [2020], but differs in two key aspects: we handle non-convex objectives arising in RL (with novel strong duality results), and we employ policy gradient subroutines instead of standard gradient descent.

2 Preliminaries

2.1 Markov Decision Process

An infinite-horizon Markov Decision Process(MDP) is defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \mu \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state space and the actions space; $\mathcal{P}\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0,1]$ is the transition dynamic, where $\mathrm{P}(s' \mid s, a)$ denotes the probability of transitioning to state s' from state s after taking action a; $r\colon \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function; $\gamma \in [0,1]$ is the discount factor; and $\mu\colon \mathcal{S} \mapsto [0,1]$ is the initial state distribution. A policy $\pi\colon \mathcal{S} \mapsto \mathcal{P}(\mathcal{A})$ maps state to a probability distribution on the action space and $\pi(a \mid s)$ denotes the probability of the agent taking action a at state s.

For a given policy π, the value function $V_r^\pi : \mathcal{S} \to \mathbb{R}$ corresponding to reward r is defined as follows: $V_r^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,|\, \pi, s_0 = s\right]$, where the expectation $\mathbb{E}$ is taken over the distribution of trajectories $\tau = \{(s_t, a_t, r_t)\}_{t=0}^{\infty}$ induced by the policy π. The objective of reinforcement learning is to solve the following optimization problem when the underlying MDP is unknown: $\max_\pi \ \mathbb{E}_{s_0 \sim \mu}\left[V_r^\pi(s_0)\right]$.

2.2 Semi-Infinitely Constrained Markov Decision Process

To incorporate constraints into sequential decision-making, the Constrained Markov Decision Process (CMDP) framework extends standard MDPs by introducing a finite set of cost functions and corresponding upper bounds. However, many real-world problems involve infinitely many constraints. To address this problem, we consider the Semi-Infinitely Constrained Markov Decision Process (SICMDP), which generalizes CMDPs by allowing a continuum of constraints indexed by a compact set. Formally, a SICMDP is defined as a tuple $\mathcal{M}_c = \langle \mathcal{S}, \mathcal{A}, Y, \mathcal{P}, r, c, u, \gamma, \mu \rangle$. $Y \subset \mathbb{R}^m$ is a compact index set of the constraints. The function $c\colon Y \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ defines a continuum of cost functions indexed by $y \in Y$, and $u\colon Y \to \mathbb{R}$ specifies the corresponding upper bounds for the

constraint function (expected discounted costs). The remaining components follow the definition in standard MDPs.

We define the cost-based value function in a similar manner to the reward-based state value function V_r^π introduced earlier. Specifically, for each $y \in Y$, the cost-based state value function $V_{c_y}^\pi(s)$ is defined as the expected discounted cumulative cost accumulated when starting from state s and following policy π, where the immediate cost is given by $c_y(s,a)$ for each state-action pair. That is, $V_{c_y}^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t\, c_y(s_t,a_t) \,\middle|\, s_0 = s\right]$.

The objective in a SICMDP is to find a policy π that maximizes the expected discounted cumulative reward while satisfying an infinite collection of cost constraints indexed by $y \in Y$. Formally, we aim to solve the following optimization problem:

$$\max_\pi \ \mathbb{E}_{s_0\sim\mu}[V_r^\pi(s_0)] \quad \text{s.t.} \quad \mathbb{E}_{s_0\sim\mu}\left[V_{c_y}^\pi(s_0)\right] \le u_y,\ \forall\, y \in Y. \tag{1}$$

For simplicity, we use $V_r^\pi(\mu) := \mathbb{E}_{s_0\sim\mu}\left[V_r^\pi(s_0)\right]$ and $V_{c_y}^\pi(\mu) := \mathbb{E}_{s_0\sim\mu}\left[V_{c_y}^\pi(s_0)\right]$. Notably, when Y is finite, the SICMDP reduces to a classical CMDP.

2.3 The Banach Space of Continuous Functions and Its Dual Space

We denote by $\mathcal{C}(Y)$ the Banach space of continuous functions on Y, equipped with the uniform norm $\|f\|_\infty := \sup_{y\in Y}|f(y)|, f \in \mathcal{C}(Y)$. The dual space of $\mathcal{C}(Y)$, denoted as $\mathcal{C}(Y)^*$, is the space of all bounded linear functionals $\phi : \mathcal{C}(Y) \mapsto \mathbb{R}$. The Riesz-Markov-Kakutani representation theorem Luxemburg [2000] states that every $\phi \in \mathcal{C}(Y)^*$ can be uniquely represented as: $\phi(f) = \int_Y f(y)\, d\lambda(y), \quad \forall f \in \mathcal{C}(Y)$, for some finite signed Radon measure [2] λ on Y. We define $\mathcal{M}(Y)$ to be the space of finite signed Radon measures on Y, and $\mathcal{M}_+(Y) \subset \mathcal{M}(Y)$ to be the cone of (non-negative) measures. Let $\lambda, \lambda_0 \in \mathcal{M}_+(Y)$ and $\lambda \ll \lambda_0$ (λ is absolutely continuous with respect to λ_0), the Kullback–Leibler (KL) divergence from λ to λ_0 is $\mathrm{KL}(\lambda\|\lambda_0) := \int_Y \log\left(\frac{d\lambda}{d\lambda_0}(y)\right) d\lambda(y)$, where $\frac{d\lambda}{d\lambda_0}$ denotes the Radon–Nikodym derivative of λ with respect to λ_0. In the special case where both λ and λ_0 admit Lebesgue densities ρ and ρ_0, the KL divergence reduces to $\mathrm{KL}(\rho\|\rho_0) = \int_Y \rho(y)\log\left(\frac{\rho(y)}{\rho_0(y)}\right) dy$.

3 Semi-Infinitely Primal Dual Policy Optimization

3.1 The Lagrangian Formulation

In this section, we derive the Lagrangian formulation of the SICMDP problem (1). Recall that $Y \subset \mathbb{R}^m$ is a compact set equipped with Borel σ-algebra $\mathcal{B}(Y)$. We make the following assumption:

Assumption 1. *For each policy π, the mapping $y \mapsto V_{c_y}^\pi(\mu)$ is continuous.*

This assumption is standard in the literature of semi-infinite programming, or the optimization problem will become intractable. When Assumption 1 holds, $V_{c_y}^\pi(\mu) \in \mathcal{C}(Y)$ for any fixed π. Thus we define the Lagrange multiplier $\Lambda \in \mathcal{C}(Y)^*$. As discussed in section 2.3, Λ can be represented as a finite Radon signed measure over Y, that is, $\Lambda \in \mathcal{M}(Y)$. Since a valid dual variable needs to be non-negative, we further have $\Lambda \in \mathcal{M}_+(Y)$. For ease of notations, we use $\langle \Lambda, f\rangle$ to denote the pairing between $\Lambda \in \mathcal{M}_+(Y)$ and $f \in \mathcal{C}(Y)$. Concretely, we define $\langle \Lambda, f\rangle := \int_Y f(y)\, \Lambda(dy)$.

The Lagrangian of the SICMDP problem (1) reads $\mathcal{L}(\pi,\Lambda) = V_r^\pi(\mu) - \left\langle \Lambda, V_{c_y}^\pi(\mu) - u_y \right\rangle$. We can reformulate the SICMDP problem (1) as the saddle-point problem:

$$\max_{\pi\in\Pi}\ \min_{\Lambda\in\mathcal{M}_+(Y)}\ \mathcal{L}(\pi,\Lambda). \tag{2}$$

We demonstrate the saddle-point problem (2) is equivalent to the SICMDP problem (1) by establishing strong duality. We have the following theorem:

[2] A Radon measure is a Borel measure that is locally finite and inner regular; on a compact metric sample space such as Y, every finite Borel measure is a Radon measure.

Theorem 1. *Let the following assumptions hold:*

(a) (Slater's condition) There exists a policy π_0 such that $V_{c_y}^{\pi_0}(\mu) < u_y$, $\forall y \in Y$.

(b) (Existence of the optimal policy) The primal problem (1) *admits an optimal solution π^*.*

Then there is a measure $\Lambda^* \in \mathcal{M}_+(Y)$, such that $V_r^{\pi^*}(\mu) = \sup_\pi \left\{ V_r^\pi(\mu) - \left\langle \Lambda^*, V_{c_y}^\pi(\mu) - u_y \right\rangle \right\}$ and $\left\langle \Lambda^*, V_{c_y}^{\pi^*}(\mu) - u_y \right\rangle = 0$.

The detailed proof is in Appendix A. To prove Theorem 1, we first show a graph set associated with problem (1) is convex, and then invoke the separating hyperplane theorem in Banach space to establish the result. Our proof of Theorem 1 shares a similar spirit with the approach in Paternain et al. [2019]. Both consider the convexity (concaveness) of a perturbation function or some related graph set. The main distinction is that we operate in infinite-dimensional spaces, and standard results (for example, Cor. 30.2.2 in Rockafellar [1997] as cited in Paternain et al. [2019]) do not apply directly. Consequently, we develop a self-contained analysis tailored to the infinite-dimensional setting.

With Theorem 1, we immediately have the following corollary.

Corollary 1. *Let $\pi^\star$ be an optimal solution of the primal problem and let $\Lambda^\star \in \mathcal{M}_+(Y)$ be a dual optimal measure. By strong duality proved in Theorem 1, and the existence of optimal solutions, the pair $(\pi^\star, \Lambda^\star)$ satisfies the saddle-point condition of the Lagrangian, namely*

$$\mathcal{L}(\pi, \Lambda^\star) \le \mathcal{L}(\pi^\star, \Lambda^\star) \le \mathcal{L}(\pi^\star, \Lambda), \ \forall\, \pi, \forall\, \Lambda \in \mathcal{M}_+(Y).$$

Consequently, we have

$$V_r^{\pi^*}(\mu) = \mathcal{L}(\pi^*, \Lambda^*) = \max_{\pi\in\Pi} \min_{\Lambda\in\mathcal{M}_+(Y)} \mathcal{L}(\pi, \Lambda) = \min_{\Lambda\in\mathcal{M}_+(Y)} \max_{\pi\in\Pi} \mathcal{L}(\pi, \Lambda).$$

3.2 The Regularized Saddle-Point Problem

To encourage a "smoothed" dual variable that can alleviate the over-conservative bias induced by the estimation error of the cost-based value functions, we add a regularization term to the objective in problem (2). Concretely, the regularization term is to penalize the distance between the measure-valued dual variable and a uniform measure over Y.

We extend the standard KL-divergence to a prox-function $\mathcal{B}(\Lambda, \Upsilon)$ for general nonnegative measures $\Lambda, \Upsilon \in \mathcal{M}_+(Y)$. Assuming $\Lambda \ll \Upsilon$, we define

$$B(\Lambda, \Upsilon) := \int_Y \log\left(\frac{d\Lambda}{d\Upsilon}(y)\right) \Lambda(dy) - \int_Y \Lambda(dy) + \int_Y \Upsilon(dy).$$

Otherwise, we set $B(\Lambda, \Upsilon) = +\infty$. Generally speaking, the proxy function B can be viewed as a KL divergence for general finite nonnegative measures. In particular, $B(\Lambda, \Upsilon)$ penalizes deviations of the dual measure Λ from a fixed reference measure Λ_u, defined as the uniform distribution over Y. This regularization discourages Λ from collapsing into degenerate (point-mass) measures, thereby preventing unbounded terms in the optimization and improving numerical stability. Moreover, by enforcing smoothness in the dual measure, the proposed primal-dual approach mitigates issues associated with hard-selection rules, such as those used in SI-CPO, including the overly conservative behavior observed in practice.

Let $\Lambda_u := \rho_0\, \phi_u \in \mathcal{M}_+(Y)$ be a uniform measure over Y, where ϕ_u is the uniform probability distribution over Y, and ρ_0 is a fixed constant. We therefore consider a regularized version of problem (2).

$$\max_{\pi\in\Pi} \min_{\Lambda\in\mathcal{M}_+(Y)} \quad L_\kappa(\pi, \Lambda) := V_r^\pi(\mu) - \left\langle \Lambda, V_{c_y}^\pi(\mu) - u_y \right\rangle + \kappa\, B(\Lambda, \Lambda_u). \tag{3}$$

where the hyperparameter $\kappa \in (0, 1]$ is the regularization coefficient. The regularization term guarantees that the regularized optimal dual variable Λ_κ^* is absolutely continuous w.r.t. the reference measure Λ_u.

The following theorem quantifies the bias introduced by the regularization.

Theorem 2 (Regularization Bias). *Under the conditions of Theorem 1, suppose $g^\pi(y) := V_{c_y}^\pi(\mu) - u_y$ is uniformly L_g-Lipschitz in y, and ϕ_u has full support on Y. Let $r_* := \|\Lambda^*\|_{\mathrm{TV}}$, $\rho_0 > 0$, and $q_Y(h) := \inf_{y\in Y} \phi_u(B(y,h) \cap Y)$, and define $p_\kappa := \inf_{\Lambda\in\mathcal{M}_+(Y)} \sup_{\pi\in\Pi} L_\kappa(\pi, \Lambda)$. For every $h > 0$, $q_Y(h) > 0$ and*

$$0 \le p_\kappa - V_r^{\pi^*}(\mu) \le r_* L_g h + \kappa \left[r_* \log \frac{r_*}{\rho_0 q_Y(h)} - r_* + \rho_0 \right], \tag{4}$$

where the logarithmic term is zero when $r_* = 0$. Thus the regularization bias vanishes as $\kappa \to 0$, including when Λ^* is atomic. If $q_Y(h) \ge q_0 h^m$ for all sufficiently small h, with $q_0 > 0$, choosing $h = \kappa$ gives an $O(\kappa \log(1/\kappa))$ bound.

The proof is deferred to Appendix A.1.

3.3 First-order Method for the Regularized Saddle-Point problem

We now solve the regularized saddle-point problem (3) by alternatively updating the primal variable π and the dual variable Λ via first-order methods. We elaborate on the primal update scheme and the dual update scheme as below.

Primal Update We consider a differentiable parametrized policy class $\{\pi_\theta \mid \theta \in \Theta\}$. Here θ represents the policy parameters like a neural network. Then at the tth iteration, we may update the primal variable via

$$\theta^{(t+1)} = \theta^{(t)} + \eta_t (\widehat{\nabla}_{\theta^{(t)}} V_r^{\pi_\theta}(\mu) - \langle \Lambda^{(t)}, \widehat{\nabla}_{\theta^{(t)}} V_{c_y}^{\pi_\theta}(\mu) \rangle).$$

Here η_t is the learning rate, and the gradient estimate $\widehat{\nabla}_\theta V_r^{\pi_\theta}(\mu)$ and $\widehat{\nabla}_\theta V_{c_y}^{\pi_\theta}(\mu)$ can be obtained by any policy gradient subroutines, such as REINFORCE, Natural Policy Gradient (NPG), or Proximal Policy Optimization (PPO). Term $\langle \Lambda^{(t)}, \widehat{\nabla}_{\theta^{(t)}} V_{c_y}^{\pi_{\theta^{(t)}}}(\mu) \rangle$ is further approximated by the Monte Carlo methods:

$$\left\langle \Lambda^{(t)}, \widehat{\nabla}_{\theta^{(t)}} V_{c_y}^{\pi_{\theta^{(t)}}}(\mu) \right\rangle = \int_Y \widehat{\nabla}_{\theta^{(t)}} V_{c_y}^{\pi_{\theta^{(t)}}}(\mu)\, \Lambda^{(t)}(dy) \approx \frac{1}{N} \sum_{i=1}^N \lambda^{(t)}(y_i)\, \widehat{\nabla}_{\theta^{(t)}} V_{c_{y_i}}^{\pi_{\theta^{(t)}}}(\mu).$$

where we sample $\{y_i\}_{i=1}^N$ uniformly from Y. Here $\lambda^{(t)} = d\Lambda^{(t)}/d\phi_u$ is the density relative to the uniform probability measure ϕ_u.

Dual Update For the dual variable $\Lambda \in \mathcal{M}_+(Y)$, we adopt a mirror-descent-style update scheme inspired by Wei et al. [2020]. In particular, we perform the following update:

$$\Lambda^{(t+1)} \in \arg\min_{\Lambda\in\mathcal{V}} \Big[\alpha_t \langle \Lambda^{(t)} - \Lambda, \widehat{V}_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - u_y \rangle + \alpha_t \kappa B(\Lambda, \Lambda_u) + B(\Lambda, \Lambda^{(t)}) \Big].$$

Here α_t can be viewed as a surrogate of the learning rate, and $\widehat{V}_{c_y}^{\pi_{\theta^{(t)}}}(\mu)$ can be obtained from any policy evaluation subroutine like Monte Carlo methods or various TD-learning methods. To stabilize the dual update, the dual update is confined to a bounded set $\mathcal{V} := \{\Lambda \in \mathcal{M}_+(Y) : \|\Lambda\|_{\mathrm{TV}} \le \bar{\rho}\}$. According to Theorem 4.1 in Wei et al. [2020], let $\lambda^{(t)}$ be the density of $\Lambda^{(t)}$ relative to ϕ_u, we have a closed-form update equation:

$$\lambda^{(t+1)}(y) = \mathrm{Proj}_{\mathcal{V}} \Big\{ \big(\lambda^{(t)}(y)\big)^{\frac{1}{1+\alpha_t\kappa}} \rho_0^{\frac{\alpha_t\kappa}{1+\alpha_t\kappa}} \exp\left(\frac{\alpha_t (\widehat{V}_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - u_y)}{1+\alpha_t\kappa} \right) \Big\}. \tag{5}$$

The dual update step also relies on a Monte Carlo approximation technique. At the t-th iteration, we draw $\{y_i\}_{i=1}^N$ independently from the uniform distribution over Y. We then compute the updated dual variables $(\lambda^{(t+1)}(y_1), \ldots, \lambda^{(t+1)}(y_N))$ using Equation (5). Since $\lambda^{(t)}$ is only known at the sample points $\{y_i'\}_{i=1}^N$ from the previous iteration, we interpolate to estimate $\lambda^{(t)}(y_i)$ for each y_i. A detailed description of the algorithm is provided in Appendix D.

Remark 1 (Comparisons with the Inner-loop Optimization Approach). *As discussed in Section 1, primal-type methods such as SI-CPO require solving an inner-loop problem $\max_{y\in Y} \widehat{V}_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - u_y$ at each iteration, which is often intractable and can introduce over-conservative bias. Our regularized primal-dual method avoids this inner-loop and can be interpreted as implicitly solving it*

via optimization-by-sampling Ma et al. [2019] or simulated annealing Bertsimas and Tsitsiklis [1993]: rather than computing a single-point solution $y^{(t)}$, we maintain a dual measure $\Lambda^{(t)}$ approximately proportional to the constraint violations, which smoothly reweights different constraints during the primal update and alleviates the over-conservative bias. Even when the value estimates are pointwise unbiased, maximizing them over y can introduce an upward bias. Smooth reweighting avoids selecting a single noisy maximum.

3.4 Convergence Analysis

We analyze SI-PDPO using natural policy gradient for policy updates, following Ding et al. [2020].

For $\pi_t = \pi_{\theta^{(t)}}$, let $\nu_t = \nu_\mu^{\pi_t}$ be its normalized discounted state–action occupancy and let $\psi_t = \nabla_\theta \log \pi_{\theta^{(t)}}$. Write $\nu^* = \nu_\mu^{\pi^*}$, and let A_f^π denote the advantage function for signal f.

Assumption 2 (SICMDP regularity and Slater's condition)**.** *The state and action spaces are finite, $\gamma \in (0,1)$, and Y is compact. Rewards, costs, and thresholds are bounded, with $|r| \le R_{\max}$ and $\sup_\pi \|g^\pi\|_\infty \le G$. The function $g^\pi(y) = V_{c_y}^\pi(\mu) - u_y$ is uniformly L_g-Lipschitz, and an optimal policy π^* exists. Slater's condition holds: there exist a policy π_{sl} and a constant $\xi > 0$ such that $\sup_{y\in Y} g^{\pi_{\mathrm{sl}}}(y) \le -\xi$. The reference probability measure ϕ_u has full support on Y.*

Assumption 3 (Policy regularity)**.** *The policies are positive, their scores are bounded and β-Lipschitz along the update segments, and the Fisher matrices satisfy $F_t = \mathbb{E}_{\nu_t}[\psi_t \psi_t^\top] \succeq \mu_F I$ for some $\mu_F > 0$.*

Assumption 4 (Compatible function approximation)**.** *For $f = r$ or c_y, let $w_{f,t}$ minimize $\mathbb{E}_{\nu_t}[(A_f^{\pi_t} - w^\top \psi_t)^2]$. These coefficients satisfy $\|w_{r,t}\|_2 \vee \sup_y \|w_{c_y,t}\|_2 \le W_0$ and*

$$\mathbb{E}_{\nu^*}[(A_r^{\pi_t} - w_{r,t}^\top \psi_t)^2] \vee \sup_y \mathbb{E}_{\nu^*}[(A_{c_y}^{\pi_t} - w_{c_y,t}^\top \psi_t)^2] \le \varepsilon_{\mathrm{bias}}.$$

The following result bounds the reward gap and constraint violation of the averaged policy $\bar\pi_T$, which selects one of the first T iterates uniformly at the start of each episode and follows it throughout.

Theorem 3 (Convergence of SI-PDPO with NPG)**.** *Under Assumptions 2–4, initialize $\Lambda^{(0)} = \Lambda_u$ and use the exact dual update and $\theta^{(t+1)} = \theta^{(t)} + \eta w_t$, where $w_t = w_{r,t} - \int_Y w_{c_y,t}\, \Lambda^{(t)}(dy)$. Choose $\bar\rho > \rho_* := 2R_{\max}/((1-\gamma)\xi)$, $0 < \rho_0 \le \bar\rho$, and $\eta = \alpha = T^{-1/2}$. Let $\Psi_0 = \mathbb{E}_{s\sim d_\mu^{\pi^*}} D_{\mathrm{KL}}(\pi^*(\cdot \mid s) \| \pi_0(\cdot \mid s))$, where $d_\mu^{\pi^*}$ is the state marginal of ν^*. Define*

$$W = (1+\bar\rho)W_0, \quad C_\theta = \frac{\Psi_0 + \beta W^2/2}{1-\gamma}, \quad \delta = \frac{(1+\bar\rho)\sqrt{\varepsilon_{\mathrm{bias}}}}{1-\gamma}, \quad B_h = \bar\rho \log \frac{\bar\rho}{\rho_0 q_Y(h)} - \bar\rho + \rho_0.$$

For every $h > 0$,

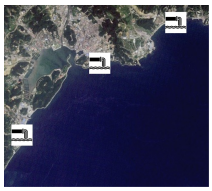
$$V_r^{\pi^*}(\mu) - V_r^{\bar\pi_T}(\mu) \le \frac{C_\theta + \rho_0 + \bar\rho G^2/2}{\sqrt{T}} + \delta + \kappa\rho_0, \tag{6}$$

$$\sup_{y\in Y} [V_{c_y}^{\bar\pi_T}(\mu) - u_y]_+ \le \left[\frac{G}{\bar\rho - \rho_*}\left\{\frac{C_\theta + B_h + \bar\rho G^2/2}{\sqrt{T}} + \delta + \kappa B_h + \bar\rho L_g h\right\}\right]^{1/2}. \tag{7}$$

For fixed h, the iteration-dependent terms decrease at rates $O(T^{-1/2})$ for the reward gap and $O(T^{-1/4})$ for constraint violation. The remaining terms arise from function approximation, regularization, and smoothing. The complete proof is in Appendix A.2.

4 Numerical Experiments

To empirically evaluate our algorithm, we performed extensive numerical experiments in two representative settings: (i) a tabular environment, Discharge of Sewage, introduced in Zhang et al. [2024]; and (ii) a real-world continuous control task in wireless communications following the setup in Zhang and Liang [2020], Liang et al. [2023]. Figure 1 illustrates the two environments. We measure the maximum constraint violation by $\max_{y\in Y}\{V_{c_y}^\pi(\mu) - u_y\}$, with positive values indicating a violated constraint. We compare SI-PDPO against four baselines:



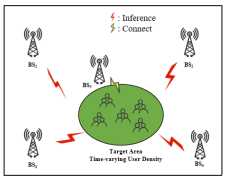


Figure 1: Illustration of the two experimental environments. Left: sewage outfall locations (image from NASA, used for illustrative purposes). Right: power control with time-varying user density in wireless networks.

Table 1: Results on discharge of sewage.

Method	Reward	Max violation
PO	6.6608 ± 0.0084	0.1683 ± 0.0169
PDPO	6.4851 ± 0.0676	0.0951 ± 0.0179
SI-CPO	6.3440 ± 0.0862	0.0246 ± 0.0017
SI-PDPO	6.5990 ± 0.0217	0.0443 ± 0.0068
EPO	6.6321 ± 0.0096	0.0991 ± 0.0123

Table 2: Results on power control.

Method	Reward	Max violation
PO	88.934 ± 0.344	0.1465 ± 0.0105
PDPO	85.212 ± 1.566	0.0721 ± 0.0089
SI-CPO	74.527 ± 1.129	0.0358 ± 0.0051
SI-PDPO	82.604 ± 0.926	0.0468 ± 0.0093
EPO	65.236 ± 1.900	0.0210 ± 0.0042

1) Policy Optimization (PO): we omitted the constraints and solve $\max_\pi V_r^\pi(\mu)$ with standard policy optimization algorithms.

2) Primal-Dual Policy Optimization (PDPO): we transform the SICMDP problem to a standard CMDP by discretizing Y, and then use primal-dual policy optimization problem as in Tessler et al. [2018], Ding et al. [2020] to solve the CMDP problem.

3) Semi-Infinite Constrained Policy Optimization (SI-CPO): we solve the SICMDP problem by primal-type SI-CPO in Zhang et al. [2024].

4) Exchange Policy Optimization (EPO): we solve the SICMDP problem using EPO in Zhang et al. [2025].

The experimental results show that SI-PDPO exhibits superior performance over the baselines. Specifically, it can effectively overcome the computational intractability and over-conservative bias of the SI-CPO algorithm.

4.1 Discharge of Sewage

The discharge of sewage setting is firstly introduced in Zhang et al. [2024]. Suppose that there are $|\mathcal{S}|$ sewage outfalls distributed across a two-dimensional region $[0, 2]^2$, with only one outfall active at any given time. The active outfall contributes to local pollution, with the effect decaying as a function of distance from the source. At each decision step, the agent may either retain the current active outfall or switch to one of the $|\mathcal{A}|$ predefined nearby alternatives. Each switch incurs a constant negative reward. The goal is to learn a switching policy that minimizes the negative reward while mitigating excessive pollution across the whole region. This setup can be formalized as a SICMDP with $Y = [0, 2]^2$ as the region of interest. The cost function is defined as $c_y(s, a) = c_y(s) = 1/(1 + \|y - s\|_2^2)$, where $y \in Y$ denotes a spatial location and s is the position of the active outfall. One may refer to Appendix B for a more detailed description of the environment.

We employ the Natural Policy Gradient (NPG) method as the policy optimization subroutine for SI-PDPO as well as for the four baseline algorithms. For policy evaluation, we use the standard Monte Carlo method. Implementation details are provided in Appendix B. Figure 2 shows the training curves of the reward-based value function $V_r^{\pi_{\theta^{(t)}}}(\mu)$ and the maximum constraint violation $\max_y V_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - u_y$ throughout training. Table 1 summarizes the final performance. Both PO and PDPO exhibit rapid increases in the reward-based value function during the initial phase of training.

However, both methods exhibit significant constraint violations, indicating that semi-infinite constraints cannot be ignored or naively addressed through discretization techniques. SI-CPO substantially reduces constraint violation but exhibits slow improvement in the reward-based value function. This behavior illustrates the over-conservative bias introduced by the inner-loop optimization proce-

Table 3: Total runtime (seconds).

Task	SI-CPO	SI-PDPO	EPO
Sewage	75.0	78.0	99.9
Power control	375.6	451.2	409.1

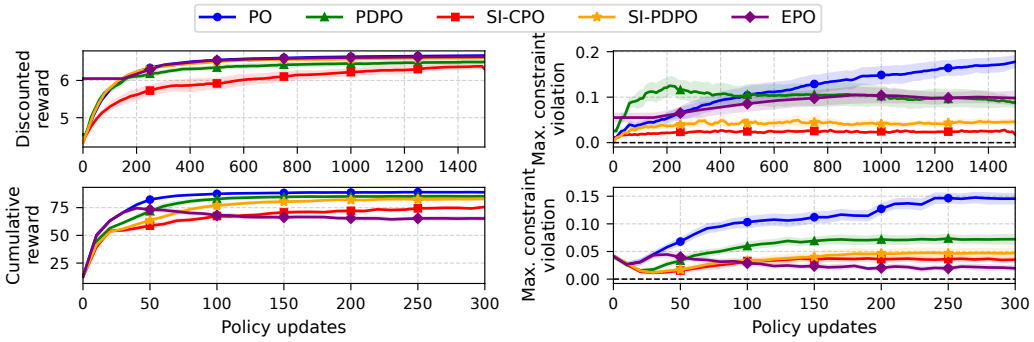


Figure 2: Training curves for the discharge of sewage (top row) and power control (bottom row) tasks. Left: reward. Right: maximum constraint violation. Shaded bands show 95% confidence intervals.

dure. In contrast, SI-PDPO maintains consistently low constraint violation throughout training while rapidly achieving high reward-based value, thereby effectively balancing performance and constraint satisfaction.

To further illustrate the over-conservative bias of SI-CPO and the effectiveness of our approach, we conducted a comparative study between SI-CPO and SI-PDPO under different levels of estimation errors $|\widehat{V}_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - V_{c_y}^{\pi_{\theta^{(t)}}}(\mu)|$, as shown in Figure 3. We find that the over-conservative bias increases when the estimation error $|\widehat{V}_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - V_{c_y}^{\pi_{\theta^{(t)}}}(\mu)|$ becomes larger. This behavior aligns with the theoretical explanations that larger noise in policy evaluation can exacerbate the bias. The results also confirm that SI-PDPO avoids the over-conservative bias observed in SI-CPO and rapidly achieves higher reward-based value.

We also compare the total wall-clock time of SI-CPO, SI-PDPO, and EPO. As shown in Table 3, the three methods have comparable runtime.

4.2 Power Control in Communication Systems

We apply our method to a more challenging real-world continuous control task-power control in communication systems. Specifically, we consider a wireless base station located within a two-dimensional rectangle region $[80, 130] \times [70, 100]$, where users are spatially distributed across the area. Within this region, a designated subarea, for example, a stadium, exhibits significant temporal fluctuations in user density. Our power control scheduling is designed to improve overall performance in this dynamic subregion, as illustrated in Figure 1. The state of our system captures both the distribution of signal quality and user density across the region.

At each decision step, the agent selects a transmission power level $a \in \mathcal{A} \subset \mathbb{R}_+$, where higher power enhances service coverage but incurs higher energy cost and may increase interference experienced by edge users in neighboring cells. And the state variable will evolve according to the previous state and the action we take. The goal is to learn a power allocation policy that maximizes energy efficiency while maintaining acceptable service quality at every location within the subregion. This problem can be formulated as a SICMDP with the particular subregion $Y = [80, 130] \times [70, 100]$, where the cost function $c_y(s, a)$ reflects the service quality delivered to location y, incorporating both signal attenuation and user density. We set $c_y(s, a) = 0$ if the service quality is acceptable at location y, and $c_y(s, a) > 0$ otherwise. And we define the constraint bound as $u_y = 0$ for all $y \in Y$.

We employ the Proximal Policy Optimization (PPO) method as the policy optimization subroutine for SI-PDPO as well as for the four baseline algorithms. For policy evaluation, we use the standard Monte Carlo method. Implementation details are provided in Appendix B. Figure 2 shows the training

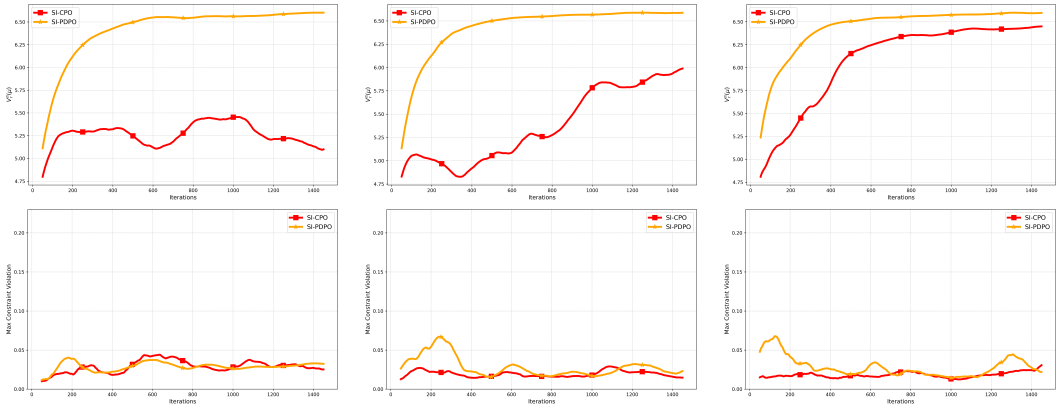


Figure 3: Comparison of SI-CPO and SI-PDPO under different levels of estimation errors $|\widehat{V}_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - V_{c_y}^{\pi_{\theta^{(t)}}}(\mu)|$ in the policy evaluation subroutine. Top row: reward-based value; bottom row: constraint violation. Left/middle/right: 300/1000/2000 Monte Carlo trajectories. Curves are smoothed with a moving average method.

curves of the cumulative reward and the maximum constraint violation throughout training. Table 2 summarizes the final performance. Like in the Discharge of Sewage task, PO and PDPO fail to enforce the semi-infinite constraints. SI-CPO exhibits slow improvement in the reward-based value function at the initial phase of training as a result of the over-conservative bias. In contrast, SI-PDPO rapidly achieves high reward-based value at the very beginning of the training process. In the later phase of the training process, SI-PDPO achieves a higher reward-based value than SI-CPO. Both methods maintain low constraint violation, with SI-CPO attaining a lower violation than SI-PDPO. We also compare the total wall-clock time in this task, as shown in Table 3.

4.3 Additional Experimental Results

Appendix C provides additional experiments on Monte-Carlo sample complexity, sensitivity to the regularization parameter κ, and the convergence behavior of the dual measure. The results show that SI-PDPO stabilizes with moderate sample sizes, is robust across a wide range of κ, and maintains stable dual updates without oscillation or degeneration. The PDPO grid-resolution experiment is provided in Appendix C.4.

5 Conclusion

We have proposed **SI-PDPO**, the first primal-dual policy optimization algorithm for semi-infinitely constrained RL. Unlike prior methods such as SI-CPO, which rely on inner-loop maximization and suffer from computational intractability and over-conservatism, SI-PDPO formulates the problem as a regularized saddle-point optimization using an infinite-dimensional dual variable and smoothly reweights different constraints. We establish strong duality under mild conditions. We also quantify the regularization bias and establish finite-time convergence guarantees when natural policy gradient is used to update the policy. Empirical results demonstrate that SI-PDPO avoids inner-loop maximization, effectively addresses the over-conservative bias, and achieves superior performance compared to existing baselines.

Limitations We outline the limitations of our work as follows. First, regularization introduces a bias, as quantified in Theorem 2. Second, our algorithm relies on the Monte Carlo technique to approximate integrals or dual density. Though convenient, it may be less efficient than advanced numerical integration methods or density estimators. Third, our method cannot guarantee zero constraint violation during training.

Acknowledgments and Disclosure of Funding

The work of Liangyu Zhang was supported in part by the National Natural Science Foundation of China under Grant No. 12501707. The work of Haishan Ye was supported in part by the National Natural Science Foundation of China under Grant No. 72471185.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

Nicole Bäuerle and Ulrich Rieder. *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011.

Dimitris Bertsimas and John Tsitsiklis. Simulated annealing. *Statistical science*, 8(1):10–15, 1993.

Bruno Betro. An accelerated central cutting plane algorithm for linear semi-infinite programming. *Mathematical Programming*, 101(3):479–495, 2004.

Marco C Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

Thiago Mureebe Carrijo, Wesley Bueno Cardoso, and Ardiley Torres Avelar. Linear semi-infinite programming approach for entanglement quantification. *Physical Review A*, 104(2):022413, 2021.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Sebastian Daum and Ralf Werner. A novel feasible discretization method for linear semi-infinite programming applied to basket option pricing. *Optimization*, 60(10-11):1379–1398, 2011.

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

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained MDPs. *Advances in Neural Information Processing Systems*, 36:66138–66200, 2023.

Dejan V Djonin and Vikram Krishnamurthy. Mimo transmission control in fading channels—a constrained markov decision process formulation with monotone randomized policies. *IEEE Transactions on Signal processing*, 55(10):5069–5083, 2007.

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

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Günther Gramlich, Rainer Hettich, and Ekkehard W Sachs. Local convergence of sqp methods in semi-infinite programming. *SIAM Journal on Optimization*, 5(3):641–658, 1995.

Rainer Hettich and Kenneth O Kortanek. Semi-infinite programming: theory, methods, and applications. *SIAM review*, 35(3):380–429, 1993.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained markov decision process. *Operations Research Letters*, 54:107107, 2024.

Renjie Liang, Haiyang Lyu, and Jiancun Fan. A deep reinforcement learning-based power control scheme for the 5g wireless systems. *China Communications*, 20(10):109–119, 2023.

Yongshuai Liu, Jiaxin Ding, and Xin Liu. IPO: Interior-point policy optimization under constraints. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4940–4947, 2020.

Wilhelmus AJ Luxemburg. *Riesz spaces*, volume 1. Elsevier, 2000.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

Nicholas Mastronarde and Mihaela van der Schaar. Fast reinforcement learning for energy-efficient wireless communication. *IEEE Transactions on Signal Processing*, 59(12):6262–6266, 2011.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained MDPs. *arXiv preprint arXiv:2402.15776*, 2024.

Pablo D Ochoa and Virginia N Vera de Serio. Stability of the primal-dual partition in linear semi-infinite programming. *Optimization*, 61(12):1449–1465, 2012.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Liqun Qi, Chen Ling, Xiaojiao Tong, and Guanglu Zhou. A smoothing projected newton-type algorithm for semi-infinite programming. *Computational Optimization and Applications*, 42(1): 1–30, 2009.

Rembert Reemtsen. Discretization methods for the solution of semi-infinite programming problems. *Journal of Optimization Theory and Applications*, 71:85–103, 1991.

Eugene Y Remez. Sur la détermination des polynômes d'approximation de degré donnée. *Comm. Soc. Math. Kharkov*, 10(196):41–63, 1934.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Optimal strong regret and violation in constrained mdps via policy optimization. *arXiv preprint arXiv:2410.02275*, 2024.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Bo Wei, William B Haskell, and Sixiang Zhao. An inexact primal-dual algorithm for semi-infinite programming. *Mathematical Methods of Operations Research*, 91(3):501–544, 2020.

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

Jiaming Zhang, Yujie Yang, Haoning Wang, Liping Zhang, and Shengbo Eben Li. Exchange policy optimization algorithm for semi-infinite safe reinforcement learning. *arXiv preprint arXiv:2511.04147*, 2025.

Liangyu Zhang, Yang Peng, Wenhao Yang, and Zhihua Zhang. Semi-infinitely constrained markov decision processes and provably efficient reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3722–3735, 2024.

Lin Zhang and Ying-Chang Liang. Deep reinforcement learning for multi-agent power control in heterogeneous networks. *IEEE Transactions on Wireless Communications*, 20(4):2551–2564, 2020.

Liping Zhang, Soon-Yi Wu, and Marco A López. A new exchange method for convex semi-infinite programming. *SIAM Journal on optimization*, 20(6):2959–2977, 2010.

A Omitted Proofs

Proof of Theorem 1. WLOG we can set $u_y = 0$. In the space $W := \mathbb{R} \times C(Y)$, define the sets

$$G := \left\{(r, g) \in W : r \le V_r^\pi,\ g \ge V_{c_y}^\pi \text{ for some } \pi\right\},$$
$$H := \left\{(r, g) \in W : r > V_r^{\pi^*},\ g < 0\right\},$$

where for two functions $f, h \in C(Y)$, $f < h$ means $f(y) < h(y)$ for all $y \in Y$. According to Lemma A.1, G is convex, and H is clearly open and convex.

We claim $G \cap H = \emptyset$. Suppose for contradiction that $(r, g) \in G \cap H$. Then there exists π with $r \le V_r^\pi$ and $g \ge V_{c_y}^\pi$. Since $(r, g) \in H$, we have $r > V_r^{\pi^*}$ and $g < 0$. This gives $V_r^\pi \ge r > V_r^{\pi^*}$ and $V_{c_y}^\pi \le g < 0$, meaning π is strictly feasible with $V_r^\pi > V_r^{\pi^*}$, contradicting the optimality of π^*.

Since G is convex, H is open and convex, and $G \cap H = \emptyset$, the separating hyperplane theorem yields a non-zero element $(\mu_0, \Lambda^*) \in W^* = \mathbb{R} \times \mathcal{M}(Y)$ with the orientation that G lies on the $\ge$ side of H, so that

$$\mu_0 r_1 + \langle \Lambda^*, g_1 \rangle \ge \mu_0 r_2 + \langle \Lambda^*, g_2 \rangle$$

for all $(r_1, g_1) \in G$ and $(r_2, g_2) \in H$.

From G: since r_1 can be arbitrarily negative, we need $\mu_0 \le 0$. From G: since g_1 can be arbitrarily large (positive), Λ^* must be a non-negative measure, i.e., $\Lambda^* \in \mathcal{M}_+(Y)$.

We now show $\mu_0 < 0$. If $\mu_0 = 0$, choose the uniform Slater policy π_0 with $V_{c_y}^{\pi_0} \le -\xi$ for all y, and set $g_2 = -\epsilon \mathbf{1}$ for $0 < \epsilon < \xi$. The separation inequality gives

$$-\epsilon \|\Lambda^*\|_{\mathrm{TV}} \le \langle \Lambda^*, V_{c_y}^{\pi_0} \rangle \le -\xi \|\Lambda^*\|_{\mathrm{TV}}.$$

Hence $\Lambda^* = 0$, contradicting the nonzero separator. Therefore $\mu_0 < 0$, and we normalize it to $\mu_0 = -1$.

The separation inequality becomes: for all π,

$$-V_r^\pi + \left\langle \Lambda^*, V_{c_y}^\pi \right\rangle \ge -r_2 + \langle \Lambda^*, g_2 \rangle$$

for all $(r_2, g_2) \in H$. Taking $r_2 \to (V_r^{\pi^*})^+$ and $g_2 \to 0^-$:

$$-V_r^\pi + \left\langle \Lambda^*, V_{c_y}^\pi \right\rangle \ge -V_r^{\pi^*}.$$

Rearranging: $V_r^\pi - \left\langle \Lambda^*, V_{c_y}^\pi \right\rangle \le V_r^{\pi^*}$ for all π. Since π^* is feasible ($V_{c_y}^{\pi^*} \le 0$) and $\Lambda^* \ge 0$, we have $\left\langle \Lambda^*, V_{c_y}^{\pi^*} \right\rangle \le 0$, hence

$$V_r^{\pi^*} \le V_r^{\pi^*} - \left\langle \Lambda^*, V_{c_y}^{\pi^*} \right\rangle \le V_r^{\pi^*},$$

which gives $\left\langle \Lambda^*, V_{c_y}^{\pi^*} \right\rangle = 0$ (complementary slackness) and $\sup_\pi \left\{ V_r^\pi - \left\langle \Lambda^*, V_{c_y}^\pi \right\rangle \right\} = V_r^{\pi^*}$. □

Lemma A.1. *In the space $W := \mathbb{R} \times C(Y)$, define the set*

$$G := \left\{(r, g) : r \le V_r^\pi,\ g \ge V_{c_y}^\pi \text{ for some } \pi\right\}.$$

Here for two functions $f, h \in C(Y)$, $f \le h$ means $f(y) \le h(y)$, $\forall y \in Y$. Then G is convex.

Proof of Lemma A.1. For any $(r_1, g_1), (r_2, g_2) \in G$, suppose for $i = 1, 2$,

$$r_i \le V_r^{\pi_i},\ g_i \ge V_{c_y}^{\pi_i}.$$

We use ν_1, ν_2 to denote the discounted occupancy measures of π_1, π_2 over $\mathcal{S} \times \mathcal{A}$. Specifically, for $i = 1, 2$,

$$\nu_i(s, a) := \sum_{t=0}^{\infty} \gamma^t P^{\pi_i}(S_t = s, A_t = a).$$

We have the following basic facts:

$$V_r^{\pi_i} = \sum_{s\in\mathcal{S},a\in\mathcal{A}} r(s,a)\nu_i(s,a),$$
$$V_{c_y}^{\pi_i} = \sum_{s\in\mathcal{S},a\in\mathcal{A}} c_y(s,a)\nu_i(s,a),$$
$$\pi_i(a \mid s) = \frac{\nu_i(s,a)}{\sum_{a'\in\mathcal{A}}\nu_i(s,a')},$$

where $i = 1, 2$. For any $\alpha \in (0,1)$, set $\nu_\alpha := (1-\alpha)\nu_1 + \alpha\nu_2$. Then ν_α is also an occupancy measure. Let π_α be the policy associated with ν_α. Concretely, $\pi_\alpha(a \mid s) = \nu_\alpha(s,a)/(\sum_{a'\in\mathcal{A}}\nu_\alpha(s,a'))$. Note that $V_r^{\pi_\alpha} = (1-\alpha)V_r^{\pi_1} + \alpha V_r^{\pi_2}$ and $V_{c_y}^{\pi_\alpha} = (1-\alpha)V_{c_y}^{\pi_1} + \alpha V_{c_y}^{\pi_2}$, we have

$$(1-\alpha)r_1 + \alpha r_2 \le (1-\alpha)V_r^{\pi_1} + \alpha V_r^{\pi_2} = V_r^{\pi_\alpha},$$
$$(1-\alpha)g_1 + \alpha g_2 \ge (1-\alpha)V_{c_y}^{\pi_1} + \alpha V_{c_y}^{\pi_2} = V_{c_y}^{\pi_\alpha}.$$

Therefore, $(1-\alpha)(r_1,g_1) + \alpha(r_2,g_2) \in G$ and G is convex. □

A.1 Proof of Theorem 2

We first establish the smoothing construction used in both Theorems 2 and 3.

Lemma A.2 (Smoothing a finite measure)**.** *Let Y be compact and ϕ_u a probability measure with full support on Y. For $h > 0$, define $b_h(y) = \phi_u(B(y,h)\cap Y)$ and $q_Y(h) = \inf_{y\in Y} b_h(y)$. Then $q_Y(h) > 0$. For any nonnegative finite measure Γ of mass m, the kernel*

$$K_h(y,A) = \frac{\phi_u(A\cap B(y,h))}{b_h(y)}$$

defines a measure $\Gamma_h(A) = \int_Y K_h(y,A)\,\Gamma(dy)$ of mass m, with density

$$\frac{d\Gamma_h}{d\phi_u}(z) = \int_Y \frac{\mathbf{1}\{\|z-y\| < h\}}{b_h(y)}\,\Gamma(dy) \le \frac{m}{q_Y(h)}. \tag{8}$$

For $\Lambda_u = \rho_0\phi_u$, where $\rho_0 > 0$,

$$B(\Gamma_h, \Lambda_u) \le m\log\frac{m}{\rho_0 q_Y(h)} - m + \rho_0. \tag{9}$$

If $f : Y \to \mathbb{R}$ is L-Lipschitz, then

$$|\langle\Gamma_h, f\rangle - \langle\Gamma, f\rangle| \le mLh. \tag{10}$$

The logarithmic term in Equation (9) *is interpreted as zero when $m = 0$.*

Proof. Compactness gives a finite cover $Y \subset \bigcup_{j=1}^J B(y_j, h/2)$. Full support implies $a_j := \phi_u(B(y_j,h/2)\cap Y) > 0$. For any $y \in Y$, select j with $\|y - y_j\| < h/2$. The triangle inequality gives $B(y_j,h/2)\cap Y \subset B(y,h)\cap Y$, so

$$b_h(y) \ge a_j \ge \min_{1\le j\le J} a_j > 0.$$

Taking the infimum over y proves the first assertion. The map $(y,z) \mapsto \mathbf{1}\{\|z-y\| < h\}$ is Borel measurable; integration shows that b_h and the kernel K_h are measurable. Also $K_h(y,Y) = 1$. Tonelli's theorem therefore gives

$$\Gamma_h(A) = \int_Y\int_Y \mathbf{1}_A(z)\frac{\mathbf{1}\{\|z-y\| < h\}}{b_h(y)}\,\phi_u(dz)\,\Gamma(dy)$$
$$= \int_A\left[\int_Y \frac{\mathbf{1}\{\|z-y\| < h\}}{b_h(y)}\,\Gamma(dy)\right]\phi_u(dz).$$

This proves Equation (8), absolute continuity, and $\Gamma_h(Y) = m$. If $m > 0$, write the density as λ_h. Since $0 \le \lambda_h \le m/q_Y(h)$ and $\int_Y \lambda_h \, d\phi_u = m$,

$$\begin{aligned} B(\Gamma_h, \Lambda_u) &= \int_Y \lambda_h \log \frac{\lambda_h}{\rho_0} \, d\phi_u - m + \rho_0 \\ &\le \int_Y \lambda_h \log \frac{m}{\rho_0 q_Y(h)} \, d\phi_u - m + \rho_0 \\ &= m \log \frac{m}{\rho_0 q_Y(h)} - m + \rho_0. \end{aligned}$$

Here $0 \log 0 = 0$. If $m = 0$, both Γ and Γ_h are zero, and $B(0, \Lambda_u) = \rho_0$, which proves the same bound. Finally, $K_h(y, \cdot)$ is supported within distance h of y. Hence

$$\begin{aligned} |\langle \Gamma_h, f \rangle - \langle \Gamma, f \rangle| &= \left| \int_Y \int_Y (f(z) - f(y)) \, K_h(y, dz) \, \Gamma(dy) \right| \\ &\le \int_Y \int_Y Lh \, K_h(y, dz) \, \Gamma(dy) = mLh. \end{aligned}$$

This proves Equation (10). □

Proof of Theorem 2. Write $p^* = V_r^{\pi^*}(\mu)$ and $D(\Lambda) = \sup_{\pi \in \Pi} \mathcal{L}(\pi, \Lambda)$. The function $x \log x - x + 1$ is nonnegative for $x \ge 0$, with its minimum at $x = 1$. Consequently $B(\Lambda, \Lambda_u) \ge 0$ for every nonnegative measure Λ, including the case where $B = +\infty$. Strong duality yields

$$p_\kappa = \inf_\Lambda \{ D(\Lambda) + \kappa B(\Lambda, \Lambda_u) \} \ge \inf_\Lambda D(\Lambda) = p^*.$$

For the upper bound, apply Lemma A.2 to $\Gamma = \Lambda^*$ and let Λ_h^* be the resulting smoothed measure. For every policy π, Equation (10) gives

$$\begin{aligned} \mathcal{L}(\pi, \Lambda_h^*) &= \mathcal{L}(\pi, \Lambda^*) + \langle \Lambda^* - \Lambda_h^*, g^\pi \rangle \\ &\le \mathcal{L}(\pi, \Lambda^*) + r_* L_g h. \end{aligned}$$

Taking the supremum over π and using $D(\Lambda^*) = p^*$ gives $D(\Lambda_h^*) \le p^* + r_* L_g h$. Testing the infimum defining p_κ at Λ_h^*, and then applying Equation (9), proves

$$\begin{aligned} p_\kappa - p^* &\le D(\Lambda_h^*) - p^* + \kappa B(\Lambda_h^*, \Lambda_u) \\ &\le r_* L_g h + \kappa \left[r_* \log \frac{r_*}{\rho_0 q_Y(h)} - r_* + \rho_0 \right]. \end{aligned}$$

This argument uses only the finite mass of Λ^* and thus also applies to atomic measures. To prove vanishing bias, fix $\epsilon > 0$ and choose $h > 0$ with $r_* L_g h < \epsilon/2$ (any h suffices if $r_* L_g = 0$). For this fixed h, the bracketed constant is finite and nonnegative, so the second term is below $\epsilon/2$ for all sufficiently small κ. Together with the lower bound, this proves $p_\kappa \to p^*$. If $q_Y(h) \ge q_0 h^m$ for sufficiently small h and $r_* > 0$, choosing $h = \kappa$ gives

$$p_\kappa - p^* \le \kappa \left[r_* L_g + r_* \log \frac{r_*}{\rho_0 q_0} - r_* + \rho_0 + m r_* \log \frac{1}{\kappa} \right].$$

This is $O(\kappa \log(1/\kappa))$ as $\kappa \to 0$. When $r_* = 0$, the bound is $\kappa \rho_0$, which also satisfies the stated rate. □

A.2 Proof of Theorem 3

Throughout this proof, Assumptions 2–4 hold. We write $\pi_t = \pi_{\theta^{(t)}}$, $\Lambda_t = \Lambda^{(t)}$, and $g_t = g^{\pi_t}$. The normalized discounted occupancies are

$$d_\mu^\pi(s) = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k \Pr_\mu^\pi(s_k = s), \qquad \nu_\mu^\pi(s, a) = d_\mu^\pi(s) \pi(a \mid s).$$

Let $\nu_t = \nu_\mu^{\pi_t}$, $\nu^* = \nu_\mu^{\pi^*}$, and $\psi_t(s, a) = \nabla_\theta \log \pi_{\theta^{(t)}}(a \mid s)$. For a bounded signal f, define $Q_f^\pi(s, a) = f(s, a) + \gamma \mathbb{E}[V_f^\pi(s') \mid s, a]$ and $A_f^\pi(s, a) = Q_f^\pi(s, a) - V_f^\pi(s)$.

For $T \geq 1$, the population iteration analyzed here is

$$\theta^{(t+1)} = \theta^{(t)} + \eta w_t, \qquad w_t = w_{r,t} - \int_Y w_{c_y,t}\, \Lambda_t(dy), \tag{11}$$

$$\Lambda_{t+1} \in \operatorname*{argmin}_{\Lambda \in \mathcal{V}} \{\alpha \langle \Lambda_t - \Lambda, g_t \rangle + \alpha\kappa B(\Lambda, \Lambda_u) + B(\Lambda, \Lambda_t)\}. \tag{12}$$

Here $\mathcal{V} = \{\Lambda \in \mathcal{M}_+(Y) : \|\Lambda\|_{\mathrm{TV}} \leq \bar{\rho}\}$ and $\Lambda_0 = \Lambda_u = \rho_0 \phi_u$. Put $R = R_{\max}/(1-\gamma)$, $W = (1+\bar{\rho})W_0$, and $\delta = (1+\bar{\rho})\sqrt{\varepsilon_{\mathrm{bias}}}/(1-\gamma)$.

Lemma A.3 (Policy performance and gradient identities)**.** *For any two policies π, π' and any bounded signal f,*

$$V_f^{\pi'}(\mu) - V_f^{\pi}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{\nu_\mu^{\pi'}}[A_f^{\pi}]. \tag{13}$$

For a differentiable positive policy π_θ, writing $\psi_\theta = \nabla_\theta \log \pi_\theta$,

$$\nabla_\theta V_f^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{\nu_\mu^{\pi_\theta}}[\psi_\theta A_f^{\pi_\theta}]. \tag{14}$$

Proof. The Bellman equation gives $A_f^{\pi}(s,a) = f(s,a) + \gamma \mathbb{E}[V_f^{\pi}(s') \mid s,a] - V_f^{\pi}(s)$. Under π', for any integer $H \geq 1$,

$$\mathbb{E}_\mu^{\pi'} \sum_{k=0}^{H-1} \gamma^k A_f^{\pi}(s_k, a_k) = \mathbb{E}_\mu^{\pi'} \sum_{k=0}^{H-1} \gamma^k f(s_k, a_k) - V_f^{\pi}(\mu) + \gamma^H \mathbb{E}_\mu^{\pi'} V_f^{\pi}(s_H).$$

Since $|V_f^{\pi}| \leq \|f\|_\infty/(1-\gamma)$, the terminal term vanishes as $H \to \infty$. The left side converges to $(1-\gamma)^{-1}\mathbb{E}_{\nu_\mu^{\pi'}} A_f^{\pi}$, and the reward sum converges to $V_f^{\pi'}(\mu)$. This proves Equation (13).

For Equation (14), differentiate the finite-dimensional Bellman equation. Let P_θ be the policy-induced transition matrix and

$$b_\theta(s) = \sum_a \pi_\theta(a \mid s) \psi_\theta(s,a) Q_f^{\pi_\theta}(s,a).$$

The derivative satisfies $\nabla_\theta V_f^{\pi_\theta} = b_\theta + \gamma P_\theta \nabla_\theta V_f^{\pi_\theta}$. The inverse $(I - \gamma P_\theta)^{-1} = \sum_{k \geq 0} \gamma^k P_\theta^k$ exists, so

$$\nabla_\theta V_f^{\pi_\theta}(\mu) = \sum_{k \geq 0} \gamma^k \mathbb{E}_\mu^{\pi_\theta} b_\theta(s_k) = \frac{1}{1-\gamma} \mathbb{E}_{\nu_\mu^{\pi_\theta}}[\psi_\theta Q_f^{\pi_\theta}].$$

Finally, $\sum_a \pi_\theta(a \mid s)\psi_\theta(s,a) = \nabla_\theta \sum_a \pi_\theta(a \mid s) = 0$. Subtracting $V_f^{\pi_\theta}(s)$ inside the last expectation gives Equation (14). □

Lemma A.4 (Bound on the optimal dual mass)**.** *There is a strong-duality multiplier Λ^* satisfying*

$$\|\Lambda^*\|_{\mathrm{TV}} \leq \rho_* = \frac{2R}{\xi} < \bar{\rho}. \tag{15}$$

Proof. Assumption 2 and Theorem 1 give a nonnegative multiplier with $D(\Lambda^*) = V_r^{\pi^*}(\mu)$, where $D(\Lambda) = \sup_\pi \mathcal{L}(\pi, \Lambda)$. The reward bound implies $|V_r^{\pi}(\mu)| \leq R$ for every policy. Evaluating the dual function at π_{sl} therefore gives

$$\begin{aligned} R &\geq D(\Lambda^*) \\ &\geq V_r^{\pi_{\mathrm{sl}}}(\mu) - \langle \Lambda^*, g^{\pi_{\mathrm{sl}}} \rangle \\ &\geq -R + \xi \|\Lambda^*\|_{\mathrm{TV}}. \end{aligned}$$

Rearranging proves $\|\Lambda^*\|_{\mathrm{TV}} \leq 2R/\xi$. The strict final inequality follows from the choice of $\bar{\rho}$ in Theorem 3. □

Lemma A.5 (Compatible approximation of the Lagrangian advantage)**.** *For $\Lambda_t \in \mathcal{V}$, define*

$$\ell_t(s,a) = r(s,a) - \int_Y [c_y(s,a) - (1-\gamma)u_y]\, \Lambda_t(dy).$$

Then $V_{\ell_t}^{\pi}(\mu) = \mathcal{L}(\pi, \Lambda_t)$. The vector w_t in Equation (11) is the unique minimizer of $\mathbb{E}_{\nu_t}[(A_{\ell_t}^{\pi_t} - w^\top \psi_t)^2]$, and

$$\|w_t\|_2 \leq W, \qquad \|A_{\ell_t}^{\pi_t} - w_t^\top \psi_t\|_{L_2(\nu^*)} \leq (1+\bar{\rho})\sqrt{\varepsilon_{\mathrm{bias}}}. \tag{16}$$

Proof. Bounded signals and finite mass permit interchange of the discounted sum and the integral over Y. Since $\sum_{k\geq 0}\gamma^k(1-\gamma)u_y = u_y$, this gives $V^\pi_{\ell_t}(\mu) = V^\pi_r(\mu) - \langle \Lambda_t, g^\pi\rangle$. A signal constant in (s,a) has equal value and action-value functions and hence zero advantage. Thus

$$A^{\pi_t}_{\ell_t} = A^{\pi_t}_r - \int_Y A^{\pi_t}_{c_y}\,\Lambda_t(dy). \tag{17}$$

The normal equations for the regressions in Assumption 4 are

$$F_t w_{r,t} = \mathbb{E}_{\nu_t}[\psi_t A^{\pi_t}_r], \qquad F_t w_{c_y,t} = \mathbb{E}_{\nu_t}[\psi_t A^{\pi_t}_{c_y}].$$

The uniform coefficient bound ensures that the integral defining w_t is finite. Using Equation (17),

$$\mathbb{E}_{\nu_t}[\psi_t A^{\pi_t}_{\ell_t}] = F_t w_{r,t} - \int_Y F_t w_{c_y,t}\,\Lambda_t(dy) = F_t w_t.$$

Positive definiteness of F_t proves uniqueness of the minimizer. The triangle inequality gives $\|w_t\|_2 \leq W_0 + \|\Lambda_t\|_{\mathrm{TV}} W_0 \leq W$.

Let $e_{r,t} = A^{\pi_t}_r - w^\top_{r,t}\psi_t$ and $e_{c_y,t} = A^{\pi_t}_{c_y} - w^\top_{c_y,t}\psi_t$. The $L_2(\nu^*)$ triangle inequality and Assumption 4 give

$$\begin{aligned}
\|A^{\pi_t}_{\ell_t} - w_t^\top\psi_t\|_{L_2(\nu^*)} &= \left\| e_{r,t} - \int_Y e_{c_y,t}\,\Lambda_t(dy)\right\|_{L_2(\nu^*)} \\
&\leq \|e_{r,t}\|_{L_2(\nu^*)} + \int_Y \|e_{c_y,t}\|_{L_2(\nu^*)}\,\Lambda_t(dy) \\
&\leq (1+\|\Lambda_t\|_{\mathrm{TV}})\sqrt{\varepsilon_{\mathrm{bias}}} \leq (1+\bar\rho)\sqrt{\varepsilon_{\mathrm{bias}}}.
\end{aligned}$$

This proves Equation (16). Equation (14) also shows that $w_t = (1-\gamma)F_t^{-1}\nabla_\theta\mathcal{L}(\pi_t,\Lambda_t)$, so the primal update is the natural-gradient step with the constant factor $1-\gamma$ absorbed into its step size. □

Lemma A.6 (Pinsker inequality for finite measures)**.** *For nonnegative measures Γ,Λ of mass at most $\bar\rho$,*

$$B(\Gamma,\Lambda) \geq \frac{\|\Gamma-\Lambda\|^2_{\mathrm{TV}}}{2\bar\rho}. \tag{18}$$

Proof. If $B(\Gamma,\Lambda) = +\infty$, the result is immediate. Otherwise $\Gamma \ll \Lambda$. Take the dominating measure $\Phi = \Gamma + \Lambda$ and write $p = d\Gamma/d\Phi$, $q = d\Lambda/d\Phi$. On $\{q=0\}$, absolute continuity gives $p=0$ as well. With $H(x) = x\log x - x + 1$ and $0\log 0 = 0$,

$$B(\Gamma,\Lambda) = \int qH(p/q)\,d\Phi.$$

We first prove the scalar inequality

$$\left(\frac{4}{3} + \frac{2x}{3}\right)H(x) \geq (x-1)^2, \qquad x \geq 0. \tag{19}$$

Let $F(x) = (4/3+2x/3)H(x) - (x-1)^2$. For $x>0$, direct differentiation gives $F(1) = F'(1) = 0$ and $F''(x) = 4H(x)/(3x) \geq 0$. Thus F is convex with its minimum at 1, and $F(x)\geq 0$. Continuity at 0 extends the inequality to $x=0$.

Equation (19) and Cauchy–Schwarz now give

$$\begin{aligned}
\|\Gamma-\Lambda\|^2_{\mathrm{TV}} &= \left(\int |p-q|\,d\Phi\right)^2 \\
&\leq \left(\int \left(\tfrac{4}{3}q + \tfrac{2}{3}p\right)d\Phi\right)\left(\int qH(p/q)\,d\Phi\right) \\
&= \left(\tfrac{4}{3}\Lambda(Y) + \tfrac{2}{3}\Gamma(Y)\right)B(\Gamma,\Lambda) \\
&\leq 2\bar\rho B(\Gamma,\Lambda).
\end{aligned}$$

The integrands on $\{q=0\}$ are taken as zero. Rearranging proves Equation (18). □

Lemma A.7 (Dual prox and three-point inequality)**.** *The update in Equation* (12) *has a positive density* $\lambda_{t+1} = d\Lambda_{t+1}/d\phi_u$*. Writing* $\lambda_t = d\Lambda_t/d\phi_u$*, it is given by*

$$z_t(y) = \lambda_t(y)^{\frac{1}{1+\alpha\kappa}} \rho_0^{\frac{\alpha\kappa}{1+\alpha\kappa}} \exp\left(\frac{\alpha g_t(y)}{1+\alpha\kappa}\right), \qquad \lambda_{t+1} = c_t z_t, \quad c_t = \min\left\{1, \frac{\bar{\rho}}{\int_Y z_t \, d\phi_u}\right\}. \tag{20}$$

For every $\Lambda \in \mathcal{V}$ *with* $B(\Lambda, \Lambda_u) < \infty$*,*

$$\alpha\langle \Lambda - \Lambda_{t+1}, g_t\rangle + \alpha\kappa[B(\Lambda_{t+1}, \Lambda_u) - B(\Lambda, \Lambda_u)] \le B(\Lambda, \Lambda_t) - B(\Lambda, \Lambda_{t+1}) - B(\Lambda_{t+1}, \Lambda_t). \tag{21}$$

Proof. Initially $\lambda_0 = \rho_0 > 0$. Suppose λ_t is bounded above and bounded away from zero. Since $|g_t| \le G$, the same is true of z_t in Equation (20). Its integral is finite and positive; hence $0 < c_t \le 1$, and λ_{t+1} is also bounded above and away from zero. Induction proves this for every finite t.

Any measure with a finite objective in Equation (12) is absolutely continuous with respect to ϕ_u. For its density λ, the terms depending on λ are

$$\begin{aligned} J_t(\lambda) = -\alpha \int_Y \lambda g_t \, d\phi_u + \alpha\kappa \int_Y \left[\lambda \log \frac{\lambda}{\rho_0} - \lambda + \rho_0\right] d\phi_u \\ + \int_Y \left[\lambda \log \frac{\lambda}{\lambda_t} - \lambda + \lambda_t\right] d\phi_u, \end{aligned}$$

subject to $\lambda \ge 0$ and $\int_Y \lambda \, d\phi_u \le \bar{\rho}$. Define $\zeta_t = -(1+\alpha\kappa)\log c_t \ge 0$. At $\lambda_{t+1} = c_t z_t$,

$$-\alpha g_t + \alpha\kappa \log \frac{\lambda_{t+1}}{\rho_0} + \log \frac{\lambda_{t+1}}{\lambda_t} = -\zeta_t, \qquad \zeta_t \left(\int_Y \lambda_{t+1} \, d\phi_u - \bar{\rho}\right) = 0. \tag{22}$$

The second identity follows because $c_t < 1$ forces the mass to equal $\bar{\rho}$, whereas $c_t = 1$ gives $\zeta_t = 0$. For any feasible λ,

$$\int_Y J_t'(\lambda_{t+1})(\lambda - \lambda_{t+1}) \, d\phi_u = -\zeta_t \left(\int_Y \lambda \, d\phi_u - \int_Y \lambda_{t+1} \, d\phi_u\right) \ge 0.$$

Convexity of the entropy integrands implies $J_t(\lambda) \ge J_t(\lambda_{t+1})$, proving the minimizer formula. Strict convexity gives uniqueness up to ϕ_u-null sets.

Now take a comparator Λ with density λ and finite $B(\Lambda, \Lambda_u)$. The bounds on λ_t imply that $B(\Lambda, \Lambda_t)$ and $B(\Lambda, \Lambda_{t+1})$ are finite as well. Rearranging the preceding variational inequality gives

$$\begin{aligned} \alpha\langle \Lambda - \Lambda_{t+1}, g_t\rangle \le \alpha\kappa \int_Y \log \frac{\lambda_{t+1}}{\rho_0}(\lambda - \lambda_{t+1}) \, d\phi_u \\ + \int_Y \log \frac{\lambda_{t+1}}{\lambda_t}(\lambda - \lambda_{t+1}) \, d\phi_u. \end{aligned}$$

Convexity of $B(\cdot, \Lambda_u)$ bounds the first integral by $B(\Lambda, \Lambda_u) - B(\Lambda_{t+1}, \Lambda_u)$. Expanding the three divergences gives the exact identity

$$\int_Y \log \frac{\lambda_{t+1}}{\lambda_t}(\lambda - \lambda_{t+1}) \, d\phi_u = B(\Lambda, \Lambda_t) - B(\Lambda, \Lambda_{t+1}) - B(\Lambda_{t+1}, \Lambda_t).$$

Substituting these two relations proves Equation (21). □

Proof of Theorem 3. We first control the errors from the primal and dual updates, then combine them to bound the reward gap and constraint violation.

Primal updates. Define $\Psi_t = \mathbb{E}_{s\sim d_\mu^{\pi^*}} D_{\mathrm{KL}}(\pi^*(\cdot \mid s) \| \pi_t(\cdot \mid s))$. Positive policies on finite state and action spaces make each Ψ_t finite and nonnegative. Integrating the policy score along the update segment gives

$$\begin{aligned} \log \pi_{t+1}(a \mid s) - \log \pi_t(a \mid s) &= \eta \int_0^1 \nabla_\theta \log \pi_{\theta^{(t)} + v\eta w_t}(a \mid s)^\top w_t \, dv \\ &\ge \eta \psi_t(s,a)^\top w_t - \frac{\beta\eta^2}{2}\|w_t\|_2^2 \\ &\ge \eta \psi_t(s,a)^\top w_t - \frac{\beta\eta^2 W^2}{2}. \end{aligned}$$

The middle inequality uses the β-Lipschitz score in Assumption 3; the last uses Lemma A.5. Taking expectation under ν^* gives

$$\mathbb{E}_{\nu^*}[\psi_t^\top w_t] \le \frac{\Psi_t - \Psi_{t+1}}{\eta} + \frac{\beta\eta W^2}{2}, \tag{23}$$

because $\Psi_t - \Psi_{t+1} = \mathbb{E}_{\nu^*} \log(\pi_{t+1}/\pi_t)$.

Let $e_t = A_{\ell_t}^{\pi_t} - w_t^\top \psi_t$. Apply Lemma A.3 to ℓ_t. The regularizer cancels when the multiplier is fixed, so

$$\begin{aligned} L_\kappa(\pi^*, \Lambda_t) - L_\kappa(\pi_t, \Lambda_t) &= V_{\ell_t}^{\pi^*}(\mu) - V_{\ell_t}^{\pi_t}(\mu) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{\nu^*}[w_t^\top \psi_t + e_t] \\ &\le \frac{\Psi_t - \Psi_{t+1}}{(1-\gamma)\eta} + \frac{\beta\eta W^2}{2(1-\gamma)} + \delta. \end{aligned}$$

Here Cauchy–Schwarz and Lemma A.5 imply $\mathbb{E}_{\nu^*} e_t \le \|e_t\|_{L_2(\nu^*)} \le (1+\bar\rho)\sqrt{\varepsilon_{\mathrm{bias}}}$. Summing the last inequality yields

$$\sum_{t=0}^{T-1}[L_\kappa(\pi^*, \Lambda_t) - L_\kappa(\pi_t, \Lambda_t)] \le \frac{\Psi_0 - \Psi_T}{(1-\gamma)\eta} + \frac{T\beta\eta W^2}{2(1-\gamma)} + T\delta.$$

Dropping $\Psi_T \ge 0$ and dividing by T, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}[L_\kappa(\pi^*, \Lambda_t) - L_\kappa(\pi_t, \Lambda_t)] \le \frac{\Psi_0}{(1-\gamma)\eta T} + \frac{\beta\eta W^2}{2(1-\gamma)} + \delta =: P_T. \tag{24}$$

Dual updates. For any $\Lambda \in \mathcal{V}$ with $B(\Lambda, \Lambda_u) < \infty$, the one-step difference is

$$s_t(\Lambda) := L_\kappa(\pi_t, \Lambda_t) - L_\kappa(\pi_t, \Lambda) = \langle \Lambda - \Lambda_t, g_t\rangle + \kappa[B(\Lambda_t, \Lambda_u) - B(\Lambda, \Lambda_u)].$$

Add $\alpha\langle \Lambda_{t+1} - \Lambda_t, g_t\rangle + \alpha\kappa[B(\Lambda_t, \Lambda_u) - B(\Lambda_{t+1}, \Lambda_u)]$ to both sides of Equation (21). This yields

$$\begin{aligned} \alpha s_t(\Lambda) \le{}& B(\Lambda, \Lambda_t) - B(\Lambda, \Lambda_{t+1}) + \alpha\kappa[B(\Lambda_t, \Lambda_u) - B(\Lambda_{t+1}, \Lambda_u)] \\ &+ \alpha\langle \Lambda_{t+1} - \Lambda_t, g_t\rangle - B(\Lambda_{t+1}, \Lambda_t). \end{aligned}$$

Let $a_t = \|\Lambda_{t+1} - \Lambda_t\|_{\mathrm{TV}}$. Lemma A.6 and $|g_t| \le G$ imply

$$\begin{aligned} \alpha\langle \Lambda_{t+1} - \Lambda_t, g_t\rangle - B(\Lambda_{t+1}, \Lambda_t) &\le \alpha G a_t - \frac{a_t^2}{2\bar\rho} \\ &= \frac{\alpha^2\bar\rho G^2}{2} - \frac{(a_t - \alpha\bar\rho G)^2}{2\bar\rho} \le \frac{\alpha^2\bar\rho G^2}{2}. \end{aligned}$$

Summing the one-step inequalities gives both telescoping terms explicitly:

$$\begin{aligned} \alpha\sum_{t=0}^{T-1} s_t(\Lambda) \le{}& B(\Lambda, \Lambda_0) - B(\Lambda, \Lambda_T) \\ &+ \alpha\kappa[B(\Lambda_0, \Lambda_u) - B(\Lambda_T, \Lambda_u)] + \frac{T\alpha^2\bar\rho G^2}{2}. \end{aligned}$$

Since $\Lambda_0 = \Lambda_u$, the initial regularization potential is zero. Both terminal divergences are nonnegative. Dropping them and dividing by αT gives

$$\frac{1}{T}\sum_{t=0}^{T-1}[L_\kappa(\pi_t, \Lambda_t) - L_\kappa(\pi_t, \Lambda)] \le \frac{B(\Lambda, \Lambda_u)}{\alpha T} + \frac{\alpha\bar\rho G^2}{2} =: D_T(\Lambda). \tag{25}$$

Reward gap. Let $\bar g = g^{\bar\pi_T}$. Conditioning on the uniformly selected episode index gives, for every bounded signal f,

$$V_f^{\bar\pi_T}(\mu) = \frac{1}{T}\sum_{t=0}^{T-1} V_f^{\pi_t}(\mu).$$

Thus $\bar{g} = T^{-1}\sum_t g_t$ as well. Adding Equations (24) and (25) cancels the intermediate Lagrangian terms and gives

$$\frac{1}{T}\sum_{t=0}^{T-1}[L_\kappa(\pi^*, \Lambda_t) - L_\kappa(\pi_t, \Lambda)] \le P_T + D_T(\Lambda).$$

Expanding its left side yields

$$V_r^{\pi^*}(\mu) - V_r^{\bar{\pi}_T}(\mu) - \frac{1}{T}\sum_{t=0}^{T-1}\langle \Lambda_t, g^{\pi^*}\rangle + \langle \Lambda, \bar{g}\rangle + \frac{\kappa}{T}\sum_{t=0}^{T-1} B(\Lambda_t, \Lambda_u) - \kappa B(\Lambda, \Lambda_u).$$

Feasibility of π^* and nonnegativity of Λ_t make $-\langle \Lambda_t, g^{\pi^*}\rangle \ge 0$. Each divergence is also nonnegative. Dropping these two nonnegative contributions and moving the comparator's regularization term to the right yields

$$V_r^{\pi^*}(\mu) - V_r^{\bar{\pi}_T}(\mu) + \langle \Lambda, \bar{g}\rangle \le P_T + D_T(\Lambda) + \kappa B(\Lambda, \Lambda_u). \tag{26}$$

In Equation (26), take $\Lambda = 0$, which belongs to $\mathcal{V}$ and has $B(0, \Lambda_u) = \rho_0$. Hence

$$V_r^{\pi^*}(\mu) - V_r^{\bar{\pi}_T}(\mu) \le P_T + \frac{\rho_0}{\alpha T} + \frac{\alpha\bar{\rho}G^2}{2} + \kappa\rho_0.$$

For $\eta = \alpha = T^{-1/2}$, Equation (24) gives $P_T = C_\theta/\sqrt{T} + \delta$. Substitution proves Equation (6).

Constraint violation. Let $v_T = \sup_{y\in Y}[\bar{g}(y)]_+$. The mixture preserves the bound $\|\bar{g}\|_\infty \le G$ and the Lipschitz constant L_g. Compactness of Y therefore gives a maximizer y_T of $\bar{g}$. If $v_T = 0$, Equation (7) is immediate. Otherwise $\bar{g}(y_T) = v_T > 0$. Apply Lemma A.2 to the point mass $\bar{\rho}\delta_{y_T}$, obtaining

$$\Lambda_{y_T,h}(A) = \bar{\rho}\,\frac{\phi_u(A \cap B(y_T, h))}{\phi_u(B(y_T, h) \cap Y)}.$$

This measure belongs to $\mathcal{V}$. Equations (9) and (10) give

$$B(\Lambda_{y_T,h}, \Lambda_u) \le B_h, \qquad \langle \Lambda_{y_T,h}, \bar{g}\rangle \ge \bar{\rho}(v_T - L_g h). \tag{27}$$

Strong duality yields $V_r^{\pi^*}(\mu) \ge V_r^{\pi_t}(\mu) - \langle \Lambda^*, g_t\rangle$ for every t. Averaging and using Lemma A.4, together with $\bar{g}(y) \le v_T$ and $\Lambda^* \ge 0$, gives

$$V_r^{\pi^*}(\mu) - V_r^{\bar{\pi}_T}(\mu) \ge -\langle \Lambda^*, \bar{g}\rangle \ge -\rho_* v_T. \tag{28}$$

Use $\Lambda = \Lambda_{y_T,h}$ in Equation (26). Substituting Equations (27) and (28), and the dual regret bound, gives

$$(\bar{\rho} - \rho_*)v_T \le P_T + \frac{B_h}{\alpha T} + \frac{\alpha\bar{\rho}G^2}{2} + \kappa B_h + \bar{\rho}L_g h. \tag{29}$$

Let A denote the right side divided by $\bar{\rho} - \rho_* > 0$. All its terms are nonnegative, and $0 \le v_T \le \min\{G, A\}$. If $A \le G$, then $A^2 \le GA$; if $A \ge G$, then $G^2 \le GA$. In either case $\min\{G, A\} \le \sqrt{GA}$. Finally, substitute $P_T = C_\theta/\sqrt{T} + \delta$ and $\alpha = T^{-1/2}$ into Equation (29). This gives Equation (7) and completes the proof. □

B Experimental Setup And Implementation Details

This section outlines the implementation details and computational settings used in our experiments. We evaluate our algorithm on two representative tasks: discharge of sewage and power control in communication systems. The first experiment (Discharge of Sewage) is implemented using PyTorch, RLlib, and Gurobi. The second experiment (Power Control in Communication Systems) is also implemented in PyTorch and uses a simulation environment based on wireless propagation and interference models. Wall-clock time is measured on one NVIDIA RTX 4090 in serial execution.

All methods are evaluated over 20 random seeds with a fixed environment. They use 1500 policy updates in sewage and 300 in power control. PDPO trains on a fixed 22×22 grid; SI-PDPO resamples 300/200 indices per update in sewage/power control. Evaluation uses independent 1001×1001 and 100×60 grids, respectively. Table entries are time-weighted trapezoidal means over the final

20% of updates, averaged across seeds; $\pm$ denotes a normal-approximation 95% confidence interval, 1.96 times the standard error of the seed-level means. Shaded curve bands denote pointwise 95% confidence intervals.

Power control uses Euclidean distance after normalizing the spatial coordinates by $(50, 30)$, with uniform averaging of the $k = 2$ nearest stored values. Sewage uses Euclidean distance on $[0, 2]^2$: average the two nearest values if the second-neighbor distance is below 0.05, otherwise use the nearest value. Search remains within Y, including its boundary. At each update, sample a fresh uniform support, transfer the old dual values (use the constant density at $t = 0$), perform the primal update, then apply the dual mirror update and mass projection on that same support.

B.1 Details of the Construction of Discharge of Sewage

The environment is randomly generated with $|\mathcal{S}| = 8$, $|\mathcal{A}| = 4$, and $\gamma = 0.9$.

Positions of sewage outfalls, transition dynamics, and rewards are sampled uniformly over Y, the probability simplex, and $[0, 1]$, respectively. The state-occupancy measure d is generated using the uniform policy, and Δ is set to 10^{-6}.

The optimal policy π^* is computed by solving a corresponding linear program using the true transition dynamics P and a fine discretization of Y with 10^6 grid points.

B.2 Details of the Construction of Power Control in Communication Systems

We consider a wireless communication scenario within a $200\,\mathrm{m} \times 200\,\mathrm{m}$ indoor region. Within this space, a specific subregion—representing a large indoor venue such as a school auditorium—is randomly located in the area defined by $[80, 130] \times [70, 100]$. This subregion is characterized by significant temporal variation in user density, influenced by scheduled events or gatherings.

While users are distributed throughout the entire environment, our power control scheduling specifically targets this dynamic subregion to improve communication performance under fluctuating demand.

To evaluate signal quality at each spatial location, we compute the signal-to-interference-plus-noise ratio (SINR) in linear scale, incorporating both large-scale and small-scale fading. Large-scale fading is modeled using the path loss function $\ell(d) = \left(\frac{d}{d_0}\right)^{-\alpha}$, where d is the distance between the transmitter and the location, d_0 is a reference distance, and $\alpha > 2$ is the path loss exponent. Small-scale fading is modeled using Rician fading, which accounts for a dominant line-of-sight (LOS) path in addition to multipath components. Specifically, the channel gain h_y follows a Rician distribution with Rician K-factor $K \geq 0$, representing the ratio of LOS to scattered power. The received power at location y is then given by:

$$P_{\mathrm{rx},y} = P \cdot h_y \cdot \ell(d_y),$$

where P is the transmit power of the serving base station. The SINR at location y is expressed as:

$$\mathrm{SINR}_y = \frac{P \cdot h_y \cdot \ell(d_y)}{\sum_{i \neq s} P_i \cdot h_{i,y} \cdot \ell(d_{i,y}) + \sigma^2},$$

where s denotes the index of the serving base station for location y, P_i and $h_{i,y}$ denote the transmit power and fading coefficient of interfering base station i, $d_{i,y}$ is the distance from interferer i to location y, and σ^2 is the noise power.

In addition, we use a time-varying two-dimensional Gaussian distribution to model the real-time spatial density of users, which reflects the fluctuating traffic demand within the target subregion.

Real-world user distributions in wireless networks often exhibit clustering behavior and spatiotemporal variability due to dynamic traffic patterns. To realistically simulate such dynamics, we model user density using a weighted sum of two-dimensional Gaussian functions, which effectively captures localized traffic hotspots that may vary over time.

We denote the user density at time t and spatial location (x, y) as:

$$\rho_t(x, y) = \sum_{k=1}^{N} w_k \cdot \mathcal{N}\left((x, y) \mid \mu_k, \Sigma_k\right),$$

where each term in the summation corresponds to a traffic hotspot. Here, $\mu_k \in \mathbb{R}^2$ denotes the center of the k-th hotspot, and $\Sigma_k \in \mathbb{R}^{2\times 2}$ specifies its spatial spread. The weight $w_k \in [0, 1]$ reflects the intensity of hotspot k, with the normalization condition $\sum_{k=1}^{N} w_k = 1$. Each $\mathcal{N}((x, y) \mid \mu_k, \Sigma_k)$ represents a 2D Gaussian probability density function evaluated at position (x, y).

The system is modeled with a state space of dimension $|\mathcal{S}| = 8$ and an action space of dimension $|\mathcal{A}| = 7$.

B.3 Implementation Details of SI-PDPO: Discharge of Sewage

For the first experiment, **Discharge of Sewage**, the parameterized policy class is chosen as the softmax policy: $\pi_\theta(a|s) = \frac{e^{\theta_{s,a}}}{\sum_{a'\in\mathcal{A}} e^{\theta_{s,a'}}}$, where $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is initialized as zero.

In this experiment, we define the sampling resolution over Y to be $N = 300$. The total number of training iterations is set to $T = 1500$. Each training iteration uses 1000 Monte Carlo trajectories, each with a fixed horizon length of 100. The discount factor is set to $\gamma = 0.9$, and the learning rate is fixed at $\alpha = 1$ throughout training.

The following hyper-parameters are used for the saddle-point optimization and dual variable regularization: the initial dual penalty coefficient $\rho_0 = 2$, the policy penalty parameter $\bar{\rho} = 500$, the step-size parameter $\kappa = 0.02$.

B.4 Implementation Details of SI-PDPO: Power Control in Communication Systems

In this experiment, the policy is parameterized by a fully connected neural network (FCN). The network takes the state representation as input and outputs a probability distribution over actions through a softmax layer. Specifically, the policy network consists of two hidden layers with 64 units each and ReLU activations. The output layer applies a softmax activation to produce a valid probability distribution over the discrete action space of size $|\mathcal{A}| = 7$.

The policy network is trained using the Adam optimizer with a learning rate of 5×10^{-4}. Each training iteration uses 50 Monte Carlo trajectories of horizon length 24, and training is run for 300 iterations. The discount factor is set to $\gamma = 0.8$.

We sample 200 constraint indices from Y per update. The other hyper-parameters are consistent with the power control experiment: $\rho_0 = 10$, $\bar{\rho} = 500$, and $\kappa = 0.15$. For some of the environmental parameters described in Appendix B.2, we set the path loss exponent $\alpha = 2$, the Rician K-factor $K = 8$, and the reference distance $d_0 = 1$.

In each training iteration, we use 6000 sampled points from Y to evaluate the cumulative reward over a 241-step deployment and the maximum constraint violation $\max_y V_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - u_y$ over full trajectories under the current policy.

C Additional Experimental Results

C.1 Monte Carlo Sampling Complexity

Figure 4 illustrates the training performance of SI-PDPO in the *Discharge of Sewage* task under varying numbers of Monte Carlo sample points $N \in \{10, 50, 300, 500, 1000\}$ drawn from the constraint index space Y, while keeping all other hyperparameters fixed.

As shown in the first plot of Figure 4, the reward value $V_r^\pi(\mu)$ steadily increases and converges to similar levels across all values of N, suggesting that SI-PDPO is relatively robust to the number of Monte Carlo samples used for reward estimation. Even with as few as $N = 10$, the algorithm achieves competitive long-term reward performance.

As shown in the second plot of Figure 4, we examine the effect of varying the number of Monte Carlo samples used to approximate the constraint value function, with sample sizes ranging from $N = 10$ to $N = 1000$. We observe that very small sample sizes (e.g., $N = 10$ or $N = 50$) lead to noticeably larger constraint violations and less stable convergence behavior, indicating that excessive sampling noise can negatively impact constraint satisfaction.

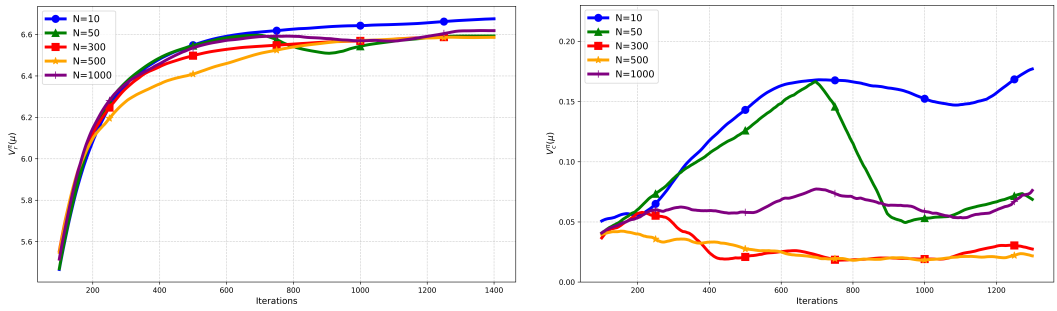


Figure 4: Effect of the number of Monte Carlo samples $y \in Y$ on SI-PDPO with fixed hyperparameters. Left: $V_r^{\pi_{\theta(t)}}(\mu)$ versus time. Right: $\max_{y \in Y} V_{c_y}^{\pi_{\theta(t)}}(\mu) - u_y$ versus time.

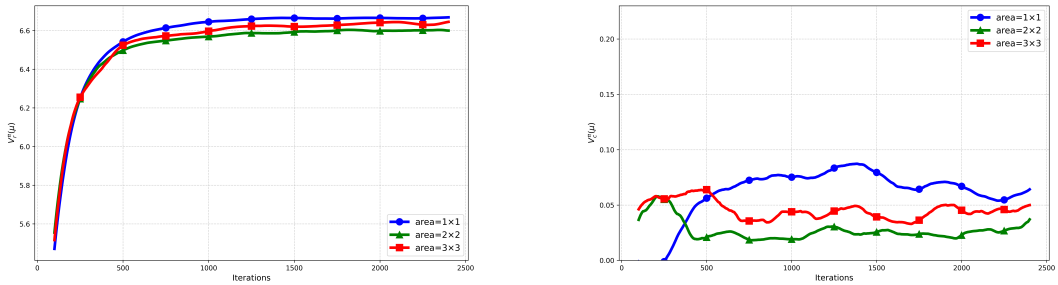


Figure 5: Effect of Monte Carlo sample size under different constraint region sizes. For regions of size 1×1, 2×2, and 3×3, the number of samples is scaled proportionally to the region area. The results show that such scaling yields stable and comparable convergence behavior, supporting the empirical sampling complexity scaling law.

In contrast, once the number of Monte Carlo samples exceeds a moderate threshold (e.g., $N \approx 300$), the constraint violation curves become significantly more stable and closely aligned. Increasing the sample size beyond this point (e.g., $N = 500$ or $N = 1000$) yields only marginal improvements in smoothness, suggesting diminishing returns in practice. These results indicate that SI-PDPO does not require highly accurate Monte Carlo estimates at each iteration; instead, a moderate number of samples suffices to achieve stable and effective constraint satisfaction.

To further investigate how sampling complexity scales with the size of the constraint domain, we conduct additional experiments on regions of different sizes. Specifically, starting from the 2×2 baseline setting, we consider smaller (1×1) and larger (3×3) continuous constraint regions. For these experiments, we scale the number of Monte Carlo samples proportionally to the area of the constraint region. Concretely, we use 75 samples for the 1×1 region and 675 samples for the 3×3 region.

The corresponding convergence and cost results are shown in Figure 5. Across different region sizes, we observe that scaling the sample size proportionally to the area of the constraint region yields comparable convergence behavior and stable optimization performance. This provides empirical evidence that the required number of Monte Carlo samples scales with the size of the constraint region. These results provide a practical guideline for selecting Monte Carlo sample sizes across problem instances with different constraint region sizes.

C.2 Sensitivity to the Regularization Parameter κ

Figure 6 examines the sensitivity of SI-PDPO to the regularization coefficient $\kappa \in \{0.01, 0.05, 0.15, 0.5\}$.

As shown in the first plot of Figure 6, reward performance $V_r^{\pi}(\mu)$ consistently converges to similar levels across all tested values of κ, again demonstrating robustness to this hyperparameter. Although $\kappa = 0.15$ shows slightly faster early convergence, final reward levels remain comparable.

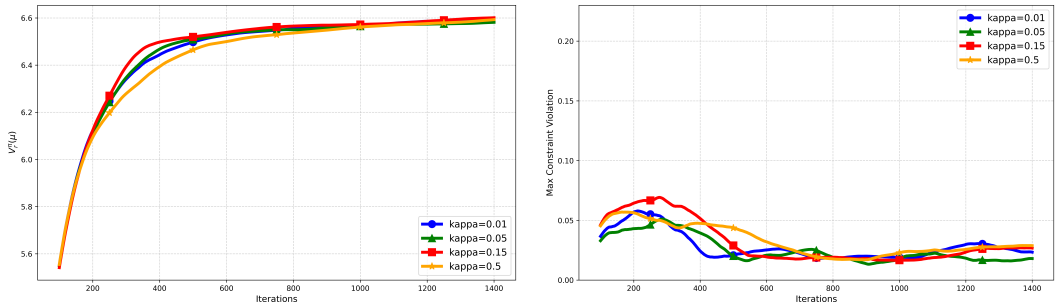


Figure 6: Effect of the regularization parameter κ on SI-PDPO with fixed hyperparameters. Left: $V_r^{\pi_{\theta^{(t)}}}(\mu)$ versus time. Right: $\max_{y \in Y} V_{c_y}^{\pi_{\theta^{(t)}}}(\mu) - u_y$ versus time.

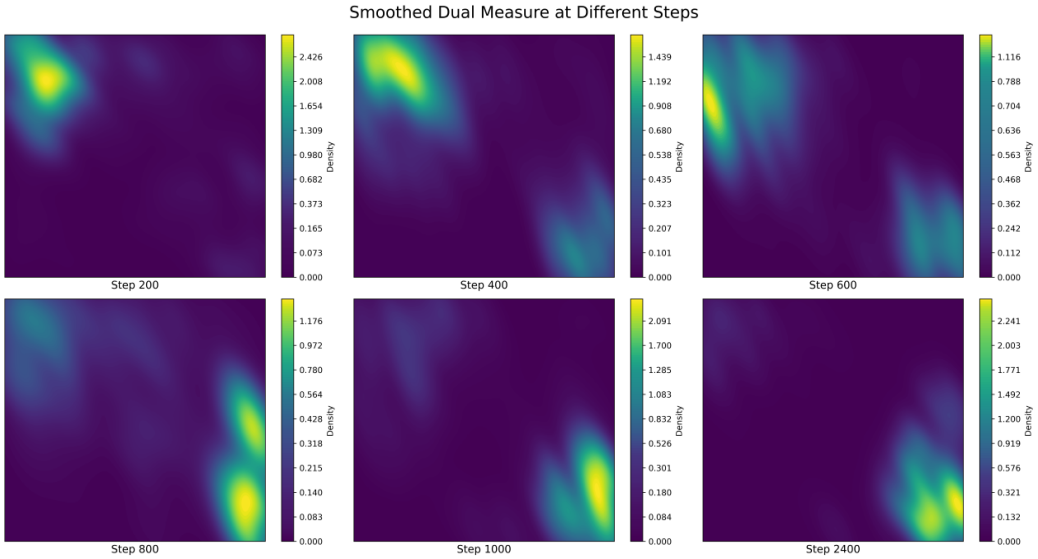


Figure 7: Evolution of the smoothed dual measure at different optimization steps in the discharge-of-sewage experiment.

The second plot in Figure 6 shows that constraint satisfaction is also robust to the choice of κ, with all tested values resulting in low and stable constraint violations throughout training. These results indicate that SI-PDPO remains reliable across a broad range of regularization strengths.

C.3 Empirical Convergence Analysis of the Dual Measure

In this appendix, we provide additional empirical analysis on the convergence behavior of the dual variable in our primal-dual framework. Specifically, we focus on the evolution of the dual measure Λ over optimization iterations and examine whether its spatial structure stabilizes as the algorithm progresses.

We conduct this analysis on the discharge-of-sewage task, where the dual variable corresponds to a non-negative measure over a continuous constraint domain. At selected optimization steps, we visualize the smoothed density of the dual measure to track how mass is distributed across the domain.

Figure 7 shows the evolution of the smoothed dual measure at different optimization steps. In early iterations, the dual measure exhibits relatively diffuse and transient high-density regions.

As optimization proceeds, the mass of the measure gradually concentrates on a smaller subset of the domain, and the locations of the dominant high-density regions become increasingly consistent across

iterations. This progressive spatial stabilization suggests that the proposed primal-dual updates lead to a stable configuration of the dual variable in practice. While our theoretical results establish strong duality and saddle-point optimality, these empirical observations provide complementary evidence that the dual measure does not exhibit oscillatory or degenerate behavior during optimization.

C.4 PDPO Grid Resolution

We compare PDPO with fixed 15×15, 22×22, and 32×32 training grids, keeping the policy-update budget and the aggregate dual scale fixed. This is a single-seed experiment. Evaluation uses the same independent grids as the main experiments, and entries are time-weighted means over the final 20% of updates.

Training grid	Reward	Maximum violation
15×15	6.4789	0.1185
22×22	6.3717	0.1092
32×32	6.2885	0.0858

Table 4: PDPO grid resolution in the discharge of sewage task.

Training grid	Reward	Maximum violation
15×15	89.128	0.1177
22×22	87.834	0.0949
32×32	85.962	0.0548

Table 5: PDPO grid resolution in the power control task.

D Omitted Algorithms

Algorithm 1 SI-PDPO

Input: state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r, a continuum of cost function c, index set Y, value for constraints u, discount factor γ, initial distribution μ.

Parameter: regularization parameter κ, parameter η_t, parameter α_t, number of samples N, maximum iteration number T, number of nearest neighbors k, TV-norm bound $\bar{\rho}$.

Initialize the density of dual variable $\lambda^{(0)}(y_i^{(0)}) = \rho_0$, for all $i \in \{1, ..., N\}$, $\theta^{(0)} = \theta_0 \in \Theta$.

for $t = 0, ..., T-1$ **do**

Sample N points $\{y_i^{(t)}\}_{i=1}^N$ i.i.d. from the uniform distribution over Y. At $t = 0$, use the constant initial density. For $t \geq 1$, use the task-specific nearest-neighbor transfer described in Appendix B; for $i \in \{1, ..., N\}$, update $\lambda^{(t)}(y_i^{(t)})$ using the average of $y_i^{(t)}$'s selected nearest neighbors in $\{y_i^{(t-1)}, i = 1, ..., N\}$:

$$\lambda^{(t)}(y_i^{(t)}) = \frac{1}{k} \sum_{j \in \mathcal{N}_k(y_i)} \lambda^{(t)}(y_j^{(t-1)}).$$

Obtain $\widehat{V}_r^{\pi_{\theta^{(t)}}}(\mu)$ and $\widehat{V}_{c_{y_i}}^{\pi_{\theta^{(t)}}}(\mu)$ for each $y_i^{(t)}$ via a policy evaluation subroutine.

Obtain $\widehat{\nabla}_{\theta^{(t)}} V_r^{\pi_{\theta^{(t)}}}(\mu)$ and $\widehat{\nabla}_{\theta^{(t)}} V_{c_{y_i}}^{\pi_{\theta^{(t)}}}(\mu)$ for each $y_i^{(t)}$ via a policy optimization subroutine.

Primal update:

$$\theta^{(t+1)} = \theta^{(t)} + \eta_t \left(\widehat{\nabla}_{\theta^{(t)}} V_r^{\pi_\theta}(\mu) - \frac{1}{N} \sum_{i=1}^N \lambda^{(t)}(y_i^{(t)}) \widehat{\nabla}_{\theta^{(t)}} V_{c_{y_i^{(t)}}}^{\pi_{\theta^{(t)}}}(\mu) \right).$$

Dual update:

$$h^{(t)}(y_i^{(t)}) = \exp\left(\frac{\alpha_t \left(\widehat{V}_{c_{y_i^{(t)}}}^{\pi_{\theta^{(t)}}}(\mu) - u_{y_i^{(t)}} \right)}{1 + \alpha_t \kappa} \right)$$

$$\lambda^{(t+1)}(y_i^{(t)}) = \min\left\{ \frac{\bar{\rho} N}{\sum_{j=1}^N z_j^{(t)}}, 1 \right\} z_i^{(t)}, \quad z_i^{(t)} = h^{(t)}(y_i^{(t)}) \rho_0^{\frac{\alpha_t \kappa}{1+\alpha_t \kappa}} \lambda^{(t)}(y_i^{(t)})^{\frac{1}{1+\alpha_t \kappa}}$$

end for **RETURN** $(\theta^{(T)}, \lambda^{(T)})$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions, assumptions, and scope of the results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion includes a dedicated limitations paragraph discussing the effect of regularization bias, Monte Carlo approximation, and possible constraint violation during training.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions and proofs are stated in Section 3.1, Section 3.2 (Theorem 2), Section 3.4 (Theorem 3), and Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental details are provided in Appendix B and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be made available at a later stage.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings, implementation details, hyperparameters, and additional experimental analyses are provided in Appendix B and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The tables and curves report 95% confidence intervals; Appendix B defines their calculation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B reports the computing hardware and implementation libraries used for the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational algorithmic work on constrained reinforcement learning and does not introduce a deployed system, dataset, or model with a direct path to societal harm.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Figure 1 credits NASA for the illustrative image used in the sewage example. The paper does not release third-party datasets, models, or code assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.