Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Primal-Dual Approach

Qinbo Bai,[1] **Amrit Singh Bedi,**[2] **Mridul Agarwal,**[1] **Alec Koppel,**[3] **Vaneet Aggarwal,**[1]

[1] Purdue University
[2] US Army Research Laboratory
[3] Supply Chain Optimization Technologies, Amazon
bai113@purdue.edu, amrit0714@gmail.com, agarw180@purdue.edu, alec.e.koppel.civ@mail.mil, vaneet@purdue.edu

Abstract

Reinforcement learning is widely used in applications where one needs to perform sequential decisions while interacting with the environment. The problem becomes more challenging when the decision requirement includes satisfying some safety constraints. The problem is mathematically formulated as constrained Markov decision process (CMDP). In the literature, various algorithms are available to solve CMDP problems in a model-free manner to achieve ϵ-optimal cumulative reward with ϵ feasible policies. An ϵ-feasible policy implies that it suffers from constraint violation. An important question here is whether we can achieve ϵ-optimal cumulative reward with zero constraint violations or not. To achieve that, we advocate the use of randomized primal-dual approach to solve the CMDP problems and propose a conservative stochastic primal-dual algorithm (CSPDA) which is shown to exhibit $\tilde{\mathcal{O}}\left(1/\epsilon^2\right)$ sample complexity to achieve ϵ-optimal cumulative reward with zero constraint violations . In the prior works, the best available sample complexity for the ϵ-optimal policy with zero constraint violation is $\tilde{\mathcal{O}}\left(1/\epsilon^5\right)$. Hence, the proposed algorithm provides a significant improvement as compared to the state of the art.

Introduction

Reinforcement learning (RL) is a machine learning framework which learns to perform a task by repeatedly interacting with the environment. This framework is widely utilized in a wide range of applications such as robotics, communications, computer vision, autonomous driving, etc. (Arulkumaran et al. 2017; Kiran et al. 2021). The problem is mathematically formulated as a Markov Decision Process (MDP) which constitute of a state, action, and transition probabilities of going from one state to the other after taking a particular action. On taking an action, a reward is achieved and the overall objective is to maximize the sum of discounted rewards. However, in various realistic environments, the agent needs to decide action where certain constraints need to be satisfied (e.g., average power constraint in wireless sensor networks (Buratti et al. 2009), queue stability constraints (Xiang et al. 2015), safe exploration (Moldovan and Abbeel

2012), etc.). The standard MDP equipped with the cost function for the constraints is called constrained Markov Decision process (CMDP) framework (Altman 1999). It is well known by (Altman 1999) that the resulting CMDP problem can be equivalently written as a linear program (LP) and hence efficient algorithms are available in the literature. But to solve the LP, one needs access to the transition probabilities of the environment, which is not available in realistic environment, and thus efficient approaches to develop model-free algorithms for CMDP are required.

Various algorithms are proposed in the literature to solve the CMDP problem without apriori knowledge of the transition probability (See Table 1 for comparisons). The performance of these algorithms is measured by the number of samples (number of state-action-state transitions) required to achieve ϵ-optimal (objective sub-optimality) ϵ-feasible (constraint violations) policies. An ϵ-feasible policy means that the constraints are not completely satisfied by the obtained policy. However, in many applications, such as in power systems (Vu et al. 2020) or autonomous vehicle control (Wen et al. 2020), violations of constraint could be catastrophic in practice. Hence, achieving optimal objective guarantees without constraint violation is an important problem and is the focus of the paper. More precisely, we ask the question, "*Is it possible to achieve the optimal sublinear convergence rate for the objective while achieving zero constraint violations for CMDP problem without apriori knowledge of the transition probability?*"

We answer the above question in affirmative in this work. We remark that the sample complexity result in this work exhibit tight dependencies on the cardinality of state and action spaces (cf. Table 1 in Appendix). The key contributions can be summarized as follows:

- To best of our knowledge, this work is the first attempt to solve CMDPs to achieve optimal sample complexity with zero constraint violation. There exist one exception in the literature which achieves the zero constraint violation but at the cost of $\tilde{\mathcal{O}}\left(1/\epsilon^5\right)$ sample complexity to achieve ϵ optimal policy (Wei, Liu, and Ying 2021). In contrast, we are able to achieve zero constraint violation with $\tilde{\mathcal{O}}\left(1/\epsilon^2\right)$ sample complexity[1].

- We utilized the idea of conservative constraints in the dual domain to derive the zero constraint violations. Conservative constrains were used recently for showing zero con-

[1]The detailed proof can be found in Appendix in (Bai et al. 2021)

straint violations in online constrained convex optimization in (Akhtar, Bedi, and Rajawat 2021), while the problem of CMDP is much more challenging than online constrained optimization. The dual constraint violations are then used to derive the primal domain results utilizing the novel analysis unique to this work (cf. Sec.). We remark that directly applying the conservative constraint idea in the primal domain does not result in the optimal dependence on the discount factor.

- The proposed algorithm utilizes adaptive state-action pair sampling (cf. Eq. (12)), due to which the stochastic gradient estimates exhibit unbounded second order moments. This makes the analysis challenging, and standard saddle point algorithms cannot be used. This difficulty is handled by using KL divergence as the performance metric for the dual update similar to (Zhang et al. 2021).

- We have performed proof of concept experiments to support the theoretical findings.

Related Work

Unconstrained RL. In the recent years, reinforcement learning has been well studied for unconstrained tabular settings. Different algorithms are compared based upon the sample complexity of the algorithm which describes the number of samples T required to achieve an ϵ optimal policy. For the infinite horizon discounted reward setting, (Lattimore and Hutter 2012) modified the famous model-based UCRL algorithm (Jaksch, Ortner, and Auer 2010) to achieve the PAC upper bound of $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ on the sample complexity. (Li et al. 2021) improved the model-free vanilla Q-learning algorithm to achieve the sample Complexity $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$. For the episodic setting with episode length of H, (Azar, Osband, and Munos 2017) proposed the model-based UCBVI algorithm and achieved a sample complexity of $\tilde{O}\left(\frac{H^3|\mathcal{S}||\mathcal{A}|}{\epsilon^2}\right)$ which is equivalent to the lower bound provided in the paper. Along the similar lines, (Jin et al. 2018) proposed a model-free UCB Q-learning and achieved the sample complexity of $\tilde{O}(\frac{H^5|\mathcal{S}||\mathcal{A}|}{\epsilon^2})$. Above all, there exists a number of near-optimal algorithms (either model-based or model-free) in the unconstrained tabular settings for RL.

Model-based Constrained RL. Once the estimated transition model is either given or estimated accurately enough, it makes intuitive sense to utilize a model-based algorithm to solve the constrained RL (CRL) problem because the problem boils down to solving only a linear program (Altman 1999). Under the model-based framework, the authors (Efroni, Mannor, and Pirotta 2020) proposed 4 algorithms namely OptCMDP & OptCMDP-bonus, OptDual, and Opt-PrimalDual which solve the problem in the primal, dual, and primal-dual domains, respectively. (Brantley et al. 2020) proposed a modular algorithm, CONRL, which utilizes the principle of optimism and can be applied to standard CRL setting and also extended to the concave-convex and knapsack setting. (Kalagarla, Jain, and Nuzzo 2021) proposed the UC-CFH algorithm which also works using the optimism

principle and provided a PAC analysis for their algorithm. (Ding et al. 2021) considered a linear MDP with constraints setting and proposed the OPDOP algorithm and extended it to the tabular setting as well.

Model-free CRL. As compared to the model-based algorithms, existing results for the model-free algorithms are fewer. The authors of (Achiam et al. 2017) proposed a constrained policy optimization (CPO) algorithm and authors of (Tessler, Mankowitz, and Mannor 2018) proposed a reward constrained policy optimization (RCPO) algorithm. The authors of (Gattami, Bai, and Aggarwal 2021) related CMDP to zero-sum Markov-Bandit games, and provided efficient solutions for CMDP. However, these works did not provide any convergence rates for their algorithms. Furthermore, the authors in (Ding et al. 2020) proposed a primal-dual natural policy gradient algorithm both in tabular and general settings and have provided a regret and constraint violation analysis. A primal only constraint rectified policy optimization (CRPO) algorithm is proposed in (Xu, Liang, and Lan 2021) to achieve sublinear convergence rate to the global optimal policy and sublinear convergence rate for the constraint violations as well. Most of the existing approaches with specific sample complexity and constraint violation error bound are summarized in Appendix. Recently, (Chen, Dong, and Wang 2021) translated the constrained RL problem into a saddle point problem and proposed a primal-dual algorithm which achieved $\tilde{O}(1/\epsilon^2)$ sample complexity to obtain ϵ-optimal ϵ-feasible solution. However, the policy is considered as the primal variable in the algorithm and an estimation of Q-table is required in the primal update, which introduces extra sample complexity and computation complexity.

Online Constrained Convex Optimization. In the field of standard online convex optimization with constraints, the problem of reducing the regret and constraint violation is well investigated in the recent years (Mahdavi, Jin, and Yang 2012; Akhtar, Bedi, and Rajawat 2021). Recently, the authors of (Akhtar, Bedi, and Rajawat 2021) utilized the idea of conservative constraints to achieve ϵ-optimal solution with $\tilde{O}(1/\epsilon^2)$ sample complexity and zero constraint violations. We utilize the conservative idea in this work to more complex setting of constrained RL problems to achieve zero constraint violations.

Problem Formulation

Consider an infinite horizon discounted reward constrained Markov Decision Process (CMDP) which is defined by tuple $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r}, \mathbf{g}^i, I, \gamma, \boldsymbol{\rho})$. In this model, $\mathcal{S}$ denotes the finite state space (with $|\mathcal{S}|$ number of states), $\mathcal{A}$ is the finite action space (with $|\mathcal{A}|$ number of actions), and $\mathbf{P} : \mathcal{S} \times \mathcal{A} \to \Delta^{|\mathcal{S}|}$ gives the transition dynamics of the CMDP (where Δ^d denotes the probability simplex in d dimension). More specifically, $\mathbf{P}(\cdot|s, a)$ describes the probability distribution of next state conditioned on the current state s and action a. We denote $\mathbf{P}(s'|s, a)$ as $\mathbf{P}_a(s, s')$ for simplicity. In the CMDP tuple, $\mathbf{r} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\mathbf{g}^i : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the ith constraint cost function, and I denotes the number of constraints. Further, γ is the discounted factor and $\boldsymbol{\rho}$ is the initial distribution of the states.

It is well known that there always exists a deterministic optimal policy for unconstrained MDP problem. However, the optimal policy for CMDP could be stochastic. Furthermore, (Altman 1999, Theorem 3.1) shows that it is enough to consider stationary stochastic policies. Thus, let us define the stationary stochastic policy as $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$, which maps a state to a distribution in the action space. The value functions for both reward and constraint's cost following such policy π are given by (Chen, Dong, and Wang 2021)

$$V_{\mathbf{r}}^{\pi}(s) = (1 - \gamma)\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t)\right],$$

$$V_{\mathbf{g}^i}^{\pi}(s) = (1 - \gamma)\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}^i(s_t, a_t)\right], \quad (1)$$

for all $s \in \mathcal{S}$. At each instant t, for given state s_t and action $a_t \sim \pi(\cdot|s_t)$, the next state s_{t+1} is distributed as $s_{t+1} \sim \mathbf{P}(\cdot|s_t, a_t)$. The expectation in (1) is with respect to the transition dynamics of the environment and the stochastic policy π. Let us denote $J_{\mathbf{r}}^{\pi}$ and $J_{\mathbf{g}^i}^{\pi}$ as the expected value function w.r.t. the initial distribution such as

$$J_{\mathbf{r}, \boldsymbol{\rho}}(\pi) = \mathbb{E}_{s_0 \sim \boldsymbol{\rho}}[V_{\mathbf{r}}^{\pi}(s_0)],$$
$$J_{\mathbf{g}^i, \boldsymbol{\rho}}(\pi) = \mathbb{E}_{s_0 \sim \boldsymbol{\rho}}[V_{\mathbf{g}^i}^{\pi}(s_0)], \quad \forall i. \quad (2)$$

The goal here is to maximize the expected value function for reward $J_{\mathbf{r}, \boldsymbol{\rho}}(\pi)$ with respect to policy π subject to satisfying the constraints value function, formulated as

$$\max_{\pi} \quad J_{\mathbf{r}, \boldsymbol{\rho}}(\pi)$$
$$s.t. \quad J_{\mathbf{g}^i, \boldsymbol{\rho}}(\pi) \geq 0 \quad \forall i \in [I], \quad (3)$$

where $[I]$ denoted the index of constraints. We note that the problem in Eq. (3) optimizes in the policy space. However, the value function is a non-linear function with respect to policy. It is well known that the problem in (3) can be equivalently written in terms of a linear program in the occupancy measure space (Altman 1999). Thus, we introduce the concept of occupancy measure as follows. For a given policy π, the occupancy measure is defined as

$$\boldsymbol{\lambda}(s, a) = (1 - \gamma)\left(\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a)\right), \quad (4)$$

where $s_0 \sim \boldsymbol{\rho}$, $a_t \sim \pi(\cdot|s_T)$, $\mathbb{P}(s_t = s, a_t = a)$ is the probability of visiting state s and taking action a in step t. By the definition in (4), the value functions in Eq. (3) can be expressed as

$$\mathbb{E}_{s \sim \boldsymbol{\rho}}[V_{\mathbf{r}}^{\pi}(s)] = \langle \boldsymbol{\lambda}, \mathbf{r} \rangle, \quad \mathbb{E}_{s \sim \boldsymbol{\rho}}[V_{\mathbf{g}^i}^{\pi}(s)] = \langle \boldsymbol{\lambda}, \mathbf{g}^i \rangle. \quad (5)$$

Following the equivalence in (5), the original problem of (3) in policy space is equivalent to the following Linear program (LP) in the occupancy measure space given by (Altman 1999, Theorem 3.3(a)(b))

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \quad \langle \boldsymbol{\lambda}, \mathbf{r} \rangle$$
$$s.t. \quad \langle \boldsymbol{\lambda}, \mathbf{g}^i \rangle \geq 0 \quad \forall i \in [I], \quad (6)$$
$$\sum_{a \in \mathcal{A}} (\mathbf{I} - \gamma \mathbf{P}_a^T) \boldsymbol{\lambda}_a = (1 - \gamma)\boldsymbol{\rho},$$

where $\boldsymbol{\lambda}_a = [\boldsymbol{\lambda}(1, a), \cdots, \boldsymbol{\lambda}(|\mathcal{S}|, a)] \in \mathbb{R}^{|\mathcal{S}|}$ is the ath column of $\boldsymbol{\lambda}$. Notice that the equality constant in Eq. (6) sums

up to 1, which means $\boldsymbol{\lambda}$ is a valid probability measure and we define $\Lambda := \{\boldsymbol{\lambda}| \sum_{s,a} \boldsymbol{\lambda}(s, a) = 1\}$ as a probability simplex. For a given occupancy measure $\boldsymbol{\lambda}$, we can recover the policy $\pi_{\boldsymbol{\lambda}}$ as

$$\pi_{\boldsymbol{\lambda}}(a|s) = \frac{\boldsymbol{\lambda}(s, a)}{\sum_{a'} \boldsymbol{\lambda}(s, a')}. \quad (7)$$

By (Altman 1999, Theorem 3.3(c)), it is known that if $\boldsymbol{\lambda}^*$ is the optimal solution for problem in Eq. (6), then $\pi_{\boldsymbol{\lambda}^*}$ will be an optimal policy for problem in Eq. (3).

Algorithm Development

The problem in (3) is well studied in the literature and various model-based algorithms are proposed (Ding et al. 2020; Xu, Liang, and Lan 2021). All of the existing approaches are able to achieve an objective optimality gap of $\tilde{\mathcal{O}}(1/\epsilon^2)$ with constraint violations of $\tilde{\mathcal{O}}(\epsilon)$ where ϵ is the accuracy parameter. Recently, the authors in (Wei, Liu, and Ying 2021) proposed a triple-Q algorithm to achieve zero constraint violations at the cost of achieving objective optimality gap of $\tilde{\mathcal{O}}(1/\epsilon^5)$. The goal here is to develop an algorithm to achieve zero constraint violation without suffering for the objective optimality gap. To do so, we consider the conservative stochastic optimization framework presented in (Akhtar, Bedi, and Rajawat 2021) and utilize it to propose a conservative version of the constrained MDPs problem in (6) as

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \quad \langle \boldsymbol{\lambda}, \mathbf{r} \rangle \quad (8a)$$

$$s.t. \quad \langle \boldsymbol{\lambda}, \mathbf{g}^i \rangle \geq \kappa \quad \forall i \in [I], \quad (8b)$$

$$\sum_{a \in \mathcal{A}} (\mathbf{I} - \gamma \mathbf{P}_a^T) \boldsymbol{\lambda}_a = (1 - \gamma)\boldsymbol{\rho}, \quad (8c)$$

where $\kappa > 0$ is tuning parameter which controls the conservative nature for the constraints. The idea is to consider a tighter version (controlled by κ) of the original inequality constraint in (6) which allows us to achieve zero constraint violation for CMDPs which does not hold for any existing algorithm. We will specify the specific value of the parameter κ later in the convergence analysis section (cf. Sec.). Note that the conservative version of the problem in Eq. (8) is still a LP and hence the strong duality holds, which motivates us to develop the primal-dual based algorithms to solve the problem in (8). By the KKT theorem, the problem in Eq. (8) is equivalent to the following a saddle point problem which we obtain by writing the Lagrangian of (8) as

$$\mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \langle \boldsymbol{\lambda}, \mathbf{r} \rangle + \sum_{i \in [I]} u^i \left(\langle \mathbf{g}^i, \boldsymbol{\lambda} \rangle - \kappa\right)$$
$$+ (1 - \gamma)\langle \boldsymbol{\rho}, \mathbf{v} \rangle + \sum_{a \in \mathcal{A}} \boldsymbol{\lambda}_a^T (\gamma \mathbf{P}_a - \mathbf{I})\mathbf{v} \quad (9)$$

$$= \langle \boldsymbol{\lambda}, \mathbf{r} \rangle + \langle \mathbf{u}, \mathbf{G}^T \boldsymbol{\lambda} - \kappa \mathbf{1} \rangle + (1 - \gamma)\langle \boldsymbol{\rho}, \mathbf{v} \rangle$$
$$+ \sum_{a \in \mathcal{A}} \langle \boldsymbol{\lambda}_a, (\gamma \mathbf{P}_a - \mathbf{I})\mathbf{v} \rangle, \quad (10)$$

where $\mathbf{u} := [u_1, u_2, \cdots, u^i]^T$ is a column vector of the dual variable corresponding to constraints in (8b), $\mathbf{v}$ is the dual variable corresponding to equality constraint in (8c),

$\mathbf{G} := [\mathbf{g}^1, \cdots, \mathbf{g}^I] \in \mathbb{R}^{\mathcal{S} \times \mathcal{A} \times I}$ collects all the $\mathbf{g}^i$'s corresponding to I constraints in (8b), and $\mathbf{1}$ is the all one column vector. From the Lagrangian in (10), the equivalent saddle point problem is given by

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\mathbf{u} \geq 0, \mathbf{v}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}). \tag{11}$$

Since the Lagrange function is linear w.r.t. both primal and dual variable, it is known that the saddle point can be solved by the primal-dual gradient descent (Nedić and Ozdaglar 2009). However, since we assume that the transition dynamics P_a is unknown, then directly evaluating gradients of Lagrangian in (11) with respect to primal and dual variables is not possible. To circumvent this issue, we resort to a randomized primal dual approach proposed in (Wang 2020) to solve the problem in a model-free stochastic manner. We assume the presence of a generative model which is a common assumption in control/RL applications. The generative model results the next state s' for a given state s and action a in the model and provides a reward $\mathbf{r}(s, a)$ to train the policy. To this end, we consider a distribution $\boldsymbol{\zeta}$ over $\mathcal{S} \times \mathcal{A}$ to write a stochastic approximation for the Lagrangian $\mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v})$ in (11) as

$$\mathcal{L}^{\boldsymbol{\zeta}}_{(s,a,s'),s_0}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = (1 - \gamma)\mathbf{v}(s_0) + \tag{12}$$

$$\mathbf{1}_{\boldsymbol{\zeta}(s,a)>0} \cdot \frac{\boldsymbol{\lambda}(s,a)(Z_{sa} - M)}{\boldsymbol{\zeta}(s,a)} - \sum_{i \in [I]} \kappa u_i,$$

where

$$Z_{sa} := \mathbf{r}(s, a) + \gamma \mathbf{v}(s') - \mathbf{v}(s) + \sum_{i \in [I]} u_i \mathbf{g}^i(s, a), \tag{13}$$

and $s_0 \sim \boldsymbol{\rho}$, the current state action pair $(s, a) \sim \boldsymbol{\zeta}$, and the next state $s' \sim \mathbf{P}(\cdot|s, a)$. We remark that M in (14) is a shift parameter which is used in the convergence analysis. We notice that the stochastic approximation $\mathcal{L}^{\boldsymbol{\zeta}}_{(s,a,s'),s_0}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v})$ in (12) is an unbiased estimator for the Lagrangian function in Eq. (10) which implies that $\mathbb{E}_{\boldsymbol{\zeta} \times \mathbf{P}(\cdot|s,a),\boldsymbol{\rho}}[\mathcal{L}^{\boldsymbol{\zeta}}_{(s,a,s'),s_0}] = \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v})$ with $\text{supp}(\boldsymbol{\zeta}) \subset \text{supp}(\boldsymbol{\lambda})$. We could see $\boldsymbol{\zeta}$ as a adaptive state-action pair distribution which helps to control the variance of the stochastic gradient estimator. The stochastic gradients of the Lagrangian with respect to primal and dual variables are given by

$$\hat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \mathbf{1}_{\boldsymbol{\zeta}(s,a)>0} \cdot \frac{Z_{sa} - M}{\boldsymbol{\zeta}(s,a)} \cdot \mathbf{E}_{sa}, \tag{14}$$

$$\hat{\nabla}_{\mathbf{u}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \mathbf{1}_{\boldsymbol{\zeta}(s,a)>0} \cdot \frac{\boldsymbol{\lambda}(s,a)\mathbf{g}(s,a)}{\boldsymbol{\zeta}(s,a)} - \kappa \mathbf{1}, \tag{15}$$

$$\hat{\nabla}_{\mathbf{v}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \mathbf{e}(s_0') + \mathbf{1}_{\boldsymbol{\zeta}(s,a)>0} \cdot \frac{\boldsymbol{\lambda}(s,a)(\gamma \mathbf{e}(s') - \mathbf{e}(s))}{\boldsymbol{\zeta}(s,a)}, \tag{16}$$

where we define $\mathbf{e}(s_0') = (1 - \gamma)\mathbf{e}(s_0)$ with $\mathbf{e}(s_0) \in \mathbb{R}^{|\mathcal{S}|}$ being a column vector with all entries equal to 0 except only the sth entry equal to 1, $\mathbf{E}_{sa} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ is a matrix with only the (s, a) entry equaling to 1 and all other entries being 0, and $\mathbf{g}(s, a) = [\mathbf{g}^1(s, a), \cdots, \mathbf{g}^i(s, a)]^T$.

Algorithm 1: Conservative Stochastic Primal-Dual Algorithm (CSPDA) for constrained RL

Input: Sample size T. Initial distribution $\boldsymbol{\rho}$. Discounted factor γ.
Parameter: Step-size α, β. Slater variable φ, Shift-parameter M, Conservative variable κ and Constant $\delta \in (0, \frac{1}{2})$
Output: $\bar{\lambda} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\lambda}^t$, $\bar{u} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{u}^t$ and $\bar{v} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{v}^t$

1: Initialize $\mathbf{u}^1 \in \mathcal{U}$, $\mathbf{v}^1 \in \mathcal{V}$ and $\boldsymbol{\lambda}^1 = \frac{1}{|\mathcal{S}||\mathcal{A}|} \cdot \mathbf{1}$
2: **for** $t = 1, 2, ..., T$ **do**
3:      $\boldsymbol{\zeta}^t := (1 - \delta)\boldsymbol{\lambda}^t + \frac{\delta}{|\mathcal{S}||\mathcal{A}|}\mathbf{1}$
4:      Sample $(s_t, a_t) \sim \boldsymbol{\zeta}^t$ and $s_0 \sim \boldsymbol{\rho}$
5:      Sample $s_t' \sim \mathcal{P}(\cdot|a_t, s_t)$ from the generative model and observe reward r_{sa}
6:      Update value functions as $\mathbf{u}$ and $\mathbf{v}$ as

$$\mathbf{u}^{t+1} = \Pi_{\mathcal{U}}(\mathbf{u}^t - \alpha \hat{\nabla}_{\mathbf{u}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)) \tag{17}$$

$$\mathbf{v}^{t+1} = \Pi_{\mathcal{V}}(\mathbf{v}^t - \alpha \hat{\nabla}_{\mathbf{v}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)) \tag{18}$$

7:      Update occupancy measure as

$$\boldsymbol{\lambda}^{t+\frac{1}{2}} = \arg\max_{\boldsymbol{\lambda}} \left\langle \hat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t \right\rangle$$
$$- \frac{1}{\beta} KL(\boldsymbol{\lambda} \| \boldsymbol{\lambda}^t) \tag{19}$$

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^{t+\frac{1}{2}} / \|\boldsymbol{\lambda}^{t+\frac{1}{2}}\|_1 \tag{20}$$

8: **end for**

With all the stochastic gradient definitions in place, we are now ready to present the proposed novel algorithm called Conservative Stochastic Primal-Dual Algorithm (CSPDA) summarized in Algorithm 1. First, we initialize the primal and dual variables in step 1. In step 4 and 5, we sample (s_t, a_t, s_0) and then obtain s_t' from the generative model. In step 6, we update the dual variables by the gradient descent step and a projection opration (See Lemma 1 for the definition of $\mathcal{U}$ and $\mathcal{V}$). In step 7, we utilize the mirror ascent update and utilize the KL divergence as the Bregman divergence to obtain tight dependencies on the convergence rate analysis similar to (Wang 2020). Then, the occupancy measure is normalized so that it remains a valid distribution.

Convergence Analysis

In this section, we study the convergence rate of the proposed Algorithm 1 in detail. We start by analyzing the duality gap for the saddle point problem in (11). Then we show that the output of Algorithm 1 given by $\bar{\lambda}$ is ϵ-optimal for the conservative version of the dual domain optimization problem in (8) of CMDPs. Finally, we perform the analysis in the policy space and present the main results of this work. We prove that the induced policy $\bar{\pi}$ by the optimal occupancy measure $\bar{\lambda}$ is also ϵ-optimal and achieves zero constraint violation at the same time. Before discussing the convergence analysis, we provide a detailed description of the assump-

tions required for the work in this paper.

Assumption 1. *(Strict feasibility) There exists a strictly feasible occupancy measure $\hat{\boldsymbol{\lambda}} \geq 0$ to problem in (8) such that*

$$\left\langle \hat{\boldsymbol{\lambda}}, \mathbf{g}^i \right\rangle - \varphi \geq 0 \quad \forall i \in [I]$$
$$\text{and} \quad \sum_a (\mathbf{I} - \gamma \mathbf{P}_a^T)\hat{\boldsymbol{\lambda}}_a = (1-\gamma)\boldsymbol{\rho} \tag{21}$$

for some $\varphi > 0$.

Assumption 1 is the stronger version of the popular Slater's condition which is often required in the analysis of convex optimization problems. A similar assumption is considered in the literature as well (Mahdavi, Jin, and Yang 2012; Akhtar, Bedi, and Rajawat 2021) and also helps to ensure the boundedness of dual variables (see Lemma 1). We remark that Assumption 1 is unique to utilize the idea of conservative constraints to obtain zero constraint violations in the long term. To be specific, Assumption 1 plays a crucial rule in proving that the optimal objective values of the original problem in (6) and it's conservative version in (8) are $\mathcal{O}(\kappa)$ apart as mentioned in Lemma 3.

Convergence Analysis for Duality Gap

In order to bound the duality gap, we note that the standard analysis of saddle point algorithms (Nedić and Ozdaglar 2009; Akhtar, Bedi, and Rajawat 2021) is not applicable because of the unbounded noise introduced into the updates due to the use of adaptive sampling of the state-action pairs (Wang 2020; Zhang et al. 2021). Therefore, it becomes necessary to obtain explicit bounds on the gradient as well as the variance of the stochastic estimates of the gradients. We start the analysis by consider the form of Slater's condition in Assumption 1, and show that the dual variables $\mathbf{u}$ and $\mathbf{v}$ are bounded (Note that the optimal dual variables now will be function of conservative variable κ as well).

Lemma 1 (Bounded dual variable $\mathbf{u}$ and $\mathbf{v}$). *Under the Assumption 1, the optimal dual variables $\mathbf{u}_\kappa^*$ and $\mathbf{v}_\kappa^*$ are bounded. Formally, it holds that $\|\mathbf{u}_\kappa^*\|_1 \leq \frac{2}{\varphi}$ and $\|\mathbf{v}_\kappa^*\|_\infty \leq \frac{1}{1-\gamma} + \frac{2}{(1-\gamma)\varphi}$.*

The proof of Lemma 1 is provided in Appendix. As a result, we define $\mathcal{U} := \left\{ \mathbf{u} \mid \|\mathbf{u}\|_1 \leq \frac{4}{\varphi} \right\}$ and $\mathcal{V} := \left\{ \mathbf{v} \mid \|\mathbf{v}\|_\infty \leq 2[\frac{1}{1-\gamma} + \frac{2}{(1-\gamma)\varphi}] \right\}$. Since we have mathematically defined the set $\mathcal{U}$ and $\mathcal{V}$, now we rewrite the saddle point formulation in (11) as

$$\max_{\boldsymbol{\lambda} \in \Lambda} \min_{(\mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V})} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}). \tag{22}$$

In the analysis presented next, we will work with the problem in (22). First, we decompose the duality gap in Lemma 2 as follows.

Lemma 2 (Duality gap). *For any dual variables $\mathbf{u}, \mathbf{v}$, let us define $\mathbf{w} = [\mathbf{u}^T, \mathbf{v}^T]^T$, and consider $\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{\boldsymbol{\lambda}}$ as defined in*

Algorithm 1, the duality gap can be bounded as

$$\mathcal{L}(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \boldsymbol{\lambda}_\kappa^*) - \mathcal{L}(\mathbf{u}, \mathbf{v}, \bar{\boldsymbol{\lambda}})$$
$$\leq \frac{1}{T} \sum_{t=1}^{T} \Big[\underbrace{\left\langle \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{w}^t, \boldsymbol{\lambda}^t), \boldsymbol{\lambda}_\kappa^* - \boldsymbol{\lambda}^t \right\rangle}_{(I)}$$
$$+ \underbrace{\left\langle \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^t, \boldsymbol{\lambda}^t), \mathbf{w}^t - \mathbf{w} \right\rangle}_{(II)} \Big]. \tag{23}$$

The bound on terms I and II in the statement of Lemma 2 are provided in the Appendix. This helps to prove the main result in Theorem 1, which establishes the final bound on the duality gap as follows.

Theorem 1. *Define $(\mathbf{u}^\dagger, \mathbf{v}^\dagger) := \arg\min_{\mathbf{u}, \mathbf{v}} \mathcal{L}(\mathbf{u}, \mathbf{v}, \bar{\boldsymbol{\lambda}})$. Recall $\boldsymbol{\lambda}_\kappa^*$ is the best solution for the conservative Lagrange problem. The duality gap of the Algorithm 1 is bounded as*

$$\mathbb{E}[\mathcal{L}(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \boldsymbol{\lambda}_\kappa^*) - \mathcal{L}(\mathbf{u}^\dagger, \mathbf{v}^\dagger, \bar{\boldsymbol{\lambda}})]$$
$$\leq \mathcal{O}\left(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{1}{(1-\gamma)\varphi} \right). \tag{24}$$

The proof of Theorem 1 is provided in Appendix. The result in Theorem 1 describes a sublinear dependence of the duality gap onto the state-action space cardinality upto a logarithmic factor. In the next subsection we utilize the duality gap upper bound to derive a bound on the objective suboptimality and the constraint violation separately.

Dual Objective and Constraint Violation

Recall that the saddle point problem in Eq. (22) is an equivalent problem to Eq. (6) where the main difference arises due to the newly introduced conservativeness parameter κ. Thus, a convergence analysis for duality gap should imply the convergence in occupancy measure in Eq. (8). But before that, we need to characterize the gap between the original problem (6) and its conservative version in (8). The following Lemma 3 shows that the gap is of the order of parameter κ.

Lemma 3. *Under Assumption 1, and condition $\kappa \leq \min\{\frac{\varphi}{2}, 1\}$, it holds that the difference of optimal values between original problem and conservative problem is $\mathcal{O}(\kappa)$. Mathematically, it holds that $\langle \boldsymbol{\lambda}^*, \mathbf{r} \rangle - \langle \boldsymbol{\lambda}_\kappa^*, \mathbf{r} \rangle \leq \frac{\kappa}{\varphi}$.*

The proof of Lemma 3 is provided in Appendix. Using the statement of Lemma 3 and Theorem 1, we obtain the convergence result in terms of output occupancy measure in following Theorem 2.

Theorem 2. *For any $0 < \epsilon < 1$, there exists a constant $\tilde{c}_1$ such that if*

$$T \geq \max\left\{ 16, 4\varphi^2, \frac{1}{\epsilon^2} \right\} \cdot \tilde{c}_1^2 \frac{I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2 \varphi^2} \tag{25}$$

set $\kappa = \frac{2\tilde{c}_1}{1-\gamma} \sqrt{\frac{I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{T}}$ and $M = 4[\frac{1}{\varphi} + \frac{1}{1-\gamma} + \frac{2}{(1-\gamma)\varphi}]$, then the constraints of the original problem in (6) satisfy:

$$\mathbb{E}\left\langle \bar{\boldsymbol{\lambda}}, \mathbf{g}^i \right\rangle \geq \epsilon\varphi \quad \forall i \in [I], \tag{26a}$$
$$\mathbb{E}\left\| \sum_a (\gamma \mathbf{P}_a^T - \mathbf{I})\bar{\boldsymbol{\lambda}}_a + (1-\gamma)\boldsymbol{\rho} \right\|_1 \leq (1-\gamma)\epsilon\varphi. \tag{26b}$$

Additionally, the objective sub-optimality of (6) *is given by*

$$\mathbb{E}[\langle \boldsymbol{\lambda}^*, \mathbf{r} \rangle - \langle \bar{\boldsymbol{\lambda}}, \mathbf{r} \rangle] \leq 3\epsilon. \tag{27}$$

The proof of Theorem 2 is provided in Appendix. Next, we present the special case of Theorem 2 in the form of Corollary 1 (see proof in Appendix), which shows the equivalent results for the case without conservation parameter, $\kappa = 0$.

Corollary 1 (Non Zero-Violation Case). *Set* $\kappa = 0$. *For any* $\epsilon > 0$, *there exists a constant* $\tilde{c}_1$ *such that if* $T \geq \tilde{c}_1^2 \cdot \frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2\varphi^2\epsilon^2}$ *then* $\bar{\boldsymbol{\lambda}}$ *satisfies the constraint violation as*

$$\mathbb{E}\langle \bar{\boldsymbol{\lambda}}, \mathbf{g}^i \rangle \geq -\epsilon \quad \forall i \in [I] \tag{28a}$$

$$\mathbb{E}\left\| \sum_a (\gamma \mathbf{P}_a^T - \mathbf{I})\bar{\boldsymbol{\lambda}}_a + (1-\gamma)\boldsymbol{\rho} \right\|_1 \leq (1-\gamma)\epsilon\varphi, \tag{28b}$$

and the sub-optimality is given by $\mathbb{E}[\langle \boldsymbol{\lambda}^*, \mathbf{r} \rangle - \langle \bar{\boldsymbol{\lambda}}, \mathbf{r} \rangle] \leq \epsilon$.

The positive lower bound of $\epsilon\varphi$ in (26a) hints that $\bar{\boldsymbol{\lambda}}$ is feasible (hence zero constraint violation). On the other hand, the lower bound in (28a) is negative $-\epsilon$ which states that the constraints in the dual space may not be satisfied for $\bar{\boldsymbol{\lambda}}$. Next, we show that how the result in Theorem 2 helps to achieve the zero constraint violation in the policy space.

Convergence Analysis in Policy Space

We have established the convergence in the occupancy measure space in Sec. and shown that $\bar{\boldsymbol{\lambda}}$ achieves an ϵ-optimal ϵ-feasible solution but the claim of zero constraint violation is still not clear. But a small violation in Eq. (26b) makes $\bar{\boldsymbol{\lambda}}$ to loose its physical meaning as discussed in (Zhang et al. 2021, Proposition 1). Thus, to make the idea clearer and explicitly show the benefit of the conservative idea utilized in this work, we further present the results in the policy space. The bound in Eq. (26b) provides an intuition that the output occupancy measure is close to the optimal one and therefore, the induced policy should also be close to the optimal policy. Such a result is mathematically presented next in Theorem 3.

Theorem 3 (Zero-Violation). *Under the condition in Theorem 2 the induced policy* $\bar{\pi}$ *by the output occupancy measure* $\bar{\boldsymbol{\lambda}}$ *is an* ϵ-*optimal policy and achieves 0 constraint violation. Mathematically, this implies that*

$$J_{\mathbf{r},\boldsymbol{\rho}}(\pi^*) - \mathbb{E}[J_{\mathbf{r},\boldsymbol{\rho}}(\bar{\pi})] \leq \epsilon \tag{29a}$$

$$\mathbb{E}[J_{\mathbf{g}^i,\boldsymbol{\rho}}(\bar{\pi})] \geq 0 \quad \forall i \in [I]. \tag{29b}$$

The proof of Theorem 3 is provided in Appendix. To get better idea about the importance of result in Theorem 3, we next present a Corollary 2 which is a special case of Theorem 3 for $\kappa = 0$.

Corollary 2 (Non Zero-Violation Case). *Under the condition in Corollary 1, the induced policy* $\bar{\pi}$ *by the output occupancy measure* $\bar{\boldsymbol{\lambda}}$ *is an* ϵ-*optimal policy w.r.t both objective and constraints. More formally,*

$$J_{\mathbf{r},\boldsymbol{\rho}}(\pi^*) - \mathbb{E}[J_{\mathbf{r},\boldsymbol{\rho}}(\bar{\pi})] \leq \epsilon \tag{30a}$$

$$\mathbb{E}[J_{\mathbf{g}^i,\boldsymbol{\rho}}(\bar{\pi})] \geq -\epsilon \quad \forall i \in [I]. \tag{30b}$$

The benefit of utilizing the conservation parameter κ becomes clear after comparing the results in (29b) and (30b).
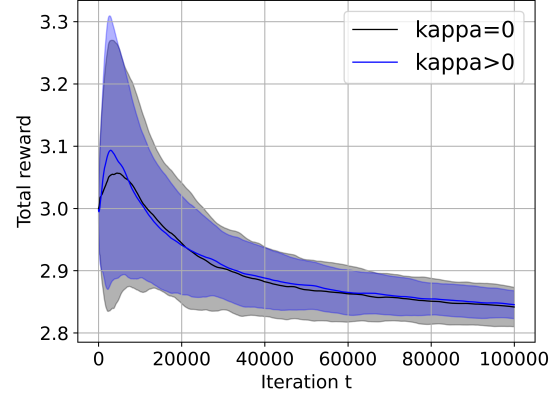

Figure 1: Learning Process of the proposed algorithm for objective with $\kappa = 0$ and $\kappa > 0$. The total reward is the objective in (31).

Evaluations on a Queuing System

In this section, we evaluate the proposed Algorithm 1 on a queuing system with a single server in discrete time (Altman 1999)[Chapter 5]. In this model, we assume a buffer of finite size L. A possible arrival is assumed to occur at the beginning of the time slot. The state of the system is the number of customers waiting in the queue at the beginning of time slot such that the size of state space is $|S| = L + 1$. We assume that there are two kinds of actions: service action and flow action. The service action is selected from a finite finite subset $\mathcal{A}$ of $[a_{min}, a_{max}]$ such that $0 < a_{min} \leq a_{max} < 1$. With a service action a, we assume that a service of a customer is successfully completed with probability a. If the service succeeds, the length of the queue will reduce by one, otherwise queue length remains the same. The flow action is a finite subset $\mathcal{B}$ of $[b_{min}, b_{max}]$ such that $0 \leq b_{min} \leq b_{max} < 1$. Given a flow action b, a customer arrives with probability b. Let the state at time t be x_t, and we assume that no customer arrives when state $x_t = L$. Finally, the overall action space is the product of service action space and flow action space, i.e., $\mathcal{A} \times \mathcal{B}$. Given an action pair (a, b) and current state x_t, the transition of this system $P(x_{t+1}|x_t, a_t = a, b_t = b)$ is shown in Table 1.

Assuming $\gamma = 0.5$, we want to maximize the total discounted cumulative reward while satisfying two constraints with respect to service and flow, simultaneously. Thus, the overall optimization problem is given as

$$\min_{\pi^a, \pi^b} \quad \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, \pi^a(s_t), \pi^b(s_t)) \right] \tag{31}$$

$$s.t. \quad \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c^i(s_t, \pi^a(s_t), \pi^b(s_t)) \right] \geq 0 \quad i = 1, 2$$

where $s_0 \sim \boldsymbol{\rho}$, π^a and π^b are the policies for the service and flow, respectively. We note that the expectation in (31) is with respect to both the stochastic policies and the transition probability. For simulations, we choose $L = 5$,

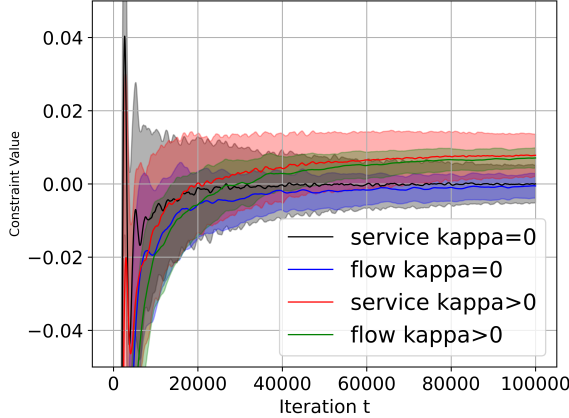| Current State	$P(x_{t+1} = x_t - 1)$	$P(x_{t+1} = x_t)$	$P(x_{t+1} = x_t + 1)$
$1 \le x_t \le L - 1$	$a(1-b)$	$ab + (1-a)(1-b)$	$(1-a)b$
$x_t = L$	a	$1-a$	0
$x_t = 0$	0	$1 - b(1-a)$	$b(1-a)$

Table 1: Transition probability of the queue system

Figure 2: Learning Process of the proposed algorithm for constraint value with $\kappa = 0$ and $\kappa > 0$. The constraint value is the L.H.S. of the constraint in (31).

$\mathcal{A} = [0.2, 0.4, 0.6, 0.8]$, and $\mathcal{B} = [0.4, 0.5, 0.6, 0.7]$ for all states besides the state $s = L$, Further, we select Slater variable $\varphi = 0.2$, number of iteration $T = 100000$, $\tilde{c}_1 = 0.02$ and conservative variable κ is selected as the statement of Theorem 2. The initial distribution ρ is set as uniform distribution. Moreover, the cost function is set to be $c(s, a, b) = -s + 5$, the constraint function for the service is defined as $c^1(s, a, b) = -10a + 3$, and the constraint function for the flow is $c^2(s, a, b) = -8(1-b)^2 + 1.2$. We run 200 independent simulations and collect the mean value and standard variance. In Fig. 1 and 2, we show the learning process of cumulative reward and constraint value for $\kappa = 0$ and $\kappa > 0$ respectively. Note that the y-axis in Fig. 1 and 2 are cumulative reward and constraint function defined in Eq. (31). It can be seen that when $\kappa > 0$, the constraint values are strictly larger than 0, which matches the result in theory. Further, the rewards are similar for both $\kappa = 0$ and $\kappa > 0$, while the case where $\kappa > 0$ helps to achieve zero constraint violation.

Conclusion

In this work, we considered the problem of learning optimal policies for infinite-horizon constrained Markov Decision Processes (CMDP) under finite state $\mathcal{S}$ and action $\mathcal{A}$ spaces with I number of constraints. This problem is also called as the constrained reinforcement learning (CRL) in the literature. To solve the problem in a model-free manner, we proposed a novel Conservative Stochastic Primal-Dual Algorithm (CSDPA) based upon the randomized primal-dual saddle point approach proposed in (Wang 2020). We show

that to achieve an ϵ-optimal policy, it is sufficient to run the proposed Algorithm 1 for $\Omega(\frac{I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2 \varphi^2 \epsilon^2})$ steps. Additionally, we proved that the proposed Algorithm 1 does not violate any of the I constraints which is unique to this work in the CRL literature. The idea is to consider a conservative version (controlled by parameter κ) of the original constraints and then a suitable choice of κ enables us to make the constraint violation zero while still achieving the best sample complexity for the objective suboptimality.

References

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

Akhtar, Z.; Bedi, A. S.; and Rajawat, K. 2021. Conservative Stochastic Optimization With Expectation Constraints. *IEEE Transactions on Signal Processing*, 69: 3190–3205.

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

Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; and Bharath, A. A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 263–272. PMLR.

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

Brantley, K.; Dudik, M.; Lykouris, T.; Miryoosefi, S.; Simchowitz, M.; Slivkins, A.; and Sun, W. 2020. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *arXiv preprint arXiv:2006.05051*.

Buratti, C.; Conti, A.; Dardari, D.; and Verdone, R. 2009. An overview on wireless sensor networks technology and evolution. *Sensors*, 9(9): 6869–6896.

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

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

Ding, D.; Zhang, K.; Basar, T.; and Jovanovic, M. R. 2020. Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes. In *NeurIPS*.

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

Gattami, A.; Bai, Q.; and Aggarwal, V. 2021. Reinforcement Learning for Constrained Markov Decision Processes. In *International Conference on Artificial Intelligence and Statistics*, 2656–2664. PMLR.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(4).

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is Q-Learning Provably Efficient? In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kalagarla, K. C.; Jain, R.; and Nuzzo, P. 2021. A Sample-Efficient Algorithm for Episodic Finite-Horizon MDP with Constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 8030–8037.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*.

Lattimore, T.; and Hutter, M. 2012. PAC Bounds for Discounted MDPs. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, 320–334. Berlin, Heidelberg: Springer-Verlag. ISBN 9783642341052.

Li, G.; Cai, C.; Chen, Y.; Gu, Y.; Wei, Y.; and Chi, Y. 2021. Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 6296–6306. PMLR.

Mahdavi, M.; Jin, R.; and Yang, T. 2012. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1): 2503–2528.

Moldovan, T. M.; and Abbeel, P. 2012. Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*.

Nedić, A.; and Ozdaglar, A. 2009. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1): 205–228.

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*.

Vu, T. L.; Mukherjee, S.; Yin, T.; Huang, R.; Huang, Q.; et al. 2020. Safe reinforcement learning for emergency loadshedding of power systems. *arXiv preprint arXiv:2011.09664*.

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

Wei, H.; Liu, X.; and Ying, L. 2021. A Provably-Efficient Model-Free Algorithm for Constrained Markov Decision Processes. *arXiv preprint arXiv:2106.01577*.

Wen, L.; Duan, J.; Li, S. E.; Xu, S.; and Peng, H. 2020. Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–7. IEEE.

Xiang, Y.; Lan, T.; Aggarwal, V.; and Chen, Y.-F. R. 2015. Joint latency and cost optimization for erasure-coded data center storage. *IEEE/ACM Transactions On Networking*, 24(4): 2443–2457.

Xu, T.; Liang, Y.; and Lan, G. 2021. CRPO: A New Approach for Safe Reinforcement Learning with Convergence Guarantee. In *International Conference on Machine Learning*, 11480–11491. PMLR.

Zhang, J.; Bedi, A. S.; Wang, M.; and Koppel, A. 2021. Cautious Reinforcement Learning via Distributional Risk in the Dual Domain. *IEEE Journal on Selected Areas in Information Theory*, 2(2): 611–626.