Journal of Artificial Intelligence Research 78 (2023) 975-1016 Submitted 08/2023; published 12/2023

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

Qinbo Bai BAI113@PURDUE.EDU
Purdue University, West Lafayette, IN 47907, USA

Amrit Singh Bedi AMRITBD@UMD.EDU
Institute of Systems Research,
University of Maryland, College Park, MD 20742, USA

Mridul Agarwal AGARW180@PURDUE.EDU
Purdue University, West Lafayette, IN 47907, USA

Alec Koppel AEKOPPEL314@GMAIL.COM
JP Morgan AI Research, New York, NY 10017, USA

Vaneet Aggarwal VANEET@PURDUE.EDU
Purdue University, West Lafayette, IN 47907, USA

Abstract

Reinforcement learning (RL) is widely used in applications where one needs to perform sequential decision-making while interacting with the environment. The standard RL problem with safety constraints is generally mathematically modeled by constrained Markov Decision Processes (CMDP), which is linear in objective and rules in occupancy measure space, where the problem becomes challenging in the case where the model is unknown apriori. The problem further becomes challenging when the decision requirement includes optimizing a concave utility while satisfying some nonlinear safety constraints. To solve such a nonlinear problem, we propose a conservative stochastic primal-dual algorithm (CSPDA) via a randomized primal-dual approach. By leveraging a generative model, we prove that CSPDA not only exhibits $\tilde{\mathcal{O}}\left(1/\epsilon^2\right)$ sample complexity, but also achieves zero constraint violations for the concave utility CMDP. Compared with the previous works, the best available sample complexity for CMDP 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

1. Introduction

Reinforcement learning (RL) is a machine learning framework that 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; Al-Abbasi et al., 2019; Geng et al., 2020; Chen et al., 2021a). The problem is mathematically formulated as a Markov Decision Process (MDP) which constitutes 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), and safe exploration (Moldovan & Abbeel, 2012), etc.). The standard MDP equipped with the cost function for the constraints is called constrained Markov Decision process (CMDP) (Altman, 1999). It is well-known that the CMDP problem can be equivalently written as a linear program (LP) in occupancy measure space (Altman, 1999), where objective and constraints are linear with respect to occupancy measure. But in many applications demand more general non-linear objectives and constraints in terms of occupancy measure, e.g., risk-sensitive constraints/objectives (Mihatsch & Neuneier, 2002), maximizing the entropy of state-action distribution (Hazan et al., 2019), imitation learning (Ho & Ermon, 2016), and fairness in multi-agent resource allocation (Margolies et al., 2014) etc. In this work, we consider a novel MDP with concave objective and convex constraints and call it CCMDP (concave CMDP). We remark here that CCMDP is still a constrained convex optimization problem. it can be efficiently solved by using any existing solution from constrained optimization literature. But the main issue here is that to do so, one would need to access the transition probabilities of the environment, which is not available in realistic model-free environment settings. Hence, efficient approaches to develop model-free algorithms for CCMDP are required. Before, moving forward, we provide a motivating example here. For more examples, one may refer to (Zhang et al., 2020).

Example 1. *(Maximaing Entropy)(Hazan et al., 2019) A fundamental problem in reinforcement learning is that of exploring the state space. How do we understand what is even possible in the context of a given environment in the absence of a reward signal? Such a problem is useful in a realistic setting since reward functions may be poorly specified or sparse. A possible quantity of interest is the entropy of the induced distribution since such an objective will encourage the agent to explore uniformly in the MDP. The maximizing entropy environment is formally defined as*

$$\max_{\pi} -\sum_{s} \bar{\boldsymbol{\lambda}}_s^{\pi} \log[\bar{\boldsymbol{\lambda}}_s^{\pi}] \tag{1}$$

where $\bar{\lambda}_s^{\pi}(s) = (1-\gamma)\sum_a \Big(\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a)\Big)$ *is the normalized occupancy measure.*

Remark 1. *It is well known that the entropy is a concave function, which satisfies the Assumption 1. However, to make the example also satisfy the Assumption 2, one may define a shifted function as* $f(\boldsymbol{\lambda}) = -\sum_s (\boldsymbol{\lambda}_s + \boldsymbol{c}) \log(\boldsymbol{\lambda}_s + \boldsymbol{c})$*, where* $\boldsymbol{c} > 0$ *is a positive shift parameter. Thus, the Lipschitz property can be guaranteed.*

To solve the CMDP problem without apriori knowledge (in a model free manner) of the transition probability, various algorithms are proposed in the literature (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., 2021) 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 CCMDP problem without apriori knowledge of the transition probabilities?*"

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

- To best of our knowledge, this work is the first attempt to provide model-free algorithm for CCMDPs that achieves optimal sample complexity with zero constraint violation. There exist one exceptions (for the special case of CMDP) 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, & Ying, 2021). In contrast, we are able to achieve zero constraint violation with $\tilde{\mathcal{O}}\left(1/\epsilon^2\right)$ sample complexity.

- This is the first attempt that provides a model-free algorithm for CCMDPs. The key challenge for solving CCMDP is the formulation of the unbiased estimator for the Lagrangian function. A trivial estimator following from previous work (Bai, Bedi, Agarwal, Koppel, & Aggarwal, 2022b) will lead to a biased estimator and make the analysis challenging (see Remark 3 for details).

- We utilized the idea of conservative constraints to derive the zero constraint violations. Such an idea was used recently for showing zero constraint violations in online constrained convex optimization in (Akhtar et al., 2021). However, the problem of CCMDP is more challenging than online constrained optimization because (1) How to achieve an unbiased estimator is unknown and (2) Following the same idea can only derive zero violation in the occupancy measure domain (see Theorem 2), while zero violation in the policy domain is required. Theorem 5.3 is then used to derive such results utilizing the novel analysis unique to this work.

- The adaptive state-action pair sampling in the proposed approach would lead to the high dependence of the number of state and action space if the standard stochastic optimization analysis is directly applied (See Remark 4 for details). To match the lower bound, we use KL divergence as the regularizer for the dual update, which is similar to (Zhang et al., 2021).

- To provide empirical evidence, we solve a problem of queuing systems in Sec. 6 and show the efficacy of the proposed algorithm.

2. Related Works

In this section, we list the related works in model-free constraint RL and Concave Utility RL fields. For the other works, please refer to Table 1.

	Algorithm	Sample Complexity	Constraint violation	Generative Model
Model-Based	OptDual-CMDP (Efroni et al., 2020) [1]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varphi^2(1-\gamma)^3\epsilon^2}\right)$	$\tilde{O}(\epsilon)$	No
	OptPrimalDual-CMDP (Efroni et al., 2020) [1]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varphi^2(1-\gamma)^3\epsilon^2}\right)$	$\tilde{O}(\epsilon)$	No
	UC-CFH (Kalagarla et al., 2021) [2]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\tilde{O}(\epsilon)$	No
	CONRL(Brantley et al., 2020)	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^6\epsilon^2}\right)$	$\tilde{O}(\epsilon)$	No
	OptPess-PrimalDual (Liu et al., 2021a)	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{\varphi^2(1-\gamma)^4\epsilon^2}\right)$	$\tilde{O}(\epsilon^2)$	No
	OPDOP (Ding et al., 2021)[Theorem 2]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varphi^2(1-\gamma)^6\epsilon^2}\right)$	$\tilde{O}(\epsilon)$	No
	UCBVI-γ (He et al., 2021)[Theorem 4.3]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	N/A	No
Model-Free	NPG-PD (Ding et al., 2020)[Theorem 4] [3]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varphi^2(1-\gamma)^5\epsilon^2}\right)$	$\tilde{O}(\epsilon)$	Yes
	CRPO (Xu et al., 2021) [4]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\epsilon^4}\right)$	$\tilde{O}(\epsilon)$	Yes
	PDSC (Chen et al., 2021b) [5]	$\tilde{O}\left(\frac{1}{\varphi^2(1-\gamma)^6\epsilon^2}\right)$	$\tilde{O}(\epsilon)$	Yes
	Triple-Q (Wei et al., 2021)	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert^{2.5}\lvert\mathcal{A}\rvert^{2.5}}{\varphi^2(1-\gamma)^{18.5}\epsilon^5}\right)$	Zero	No
	Randomized Primal–Dual (Wang, 2020)	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	N/A	Yes
	CSPDA (**This work**, Theorem 3) [6]	$\tilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varphi^2(1-\gamma)^6\epsilon^2}\right)$	**Zero**	**Yes**
Lower bound	(Lattimore & Hutter, 2012) and (Azar et al., 2013)	$\tilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	N/A	N/A
	(Vaswani et al., 2022)	$\tilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varphi^2(1-\gamma)^5\epsilon^2}\right)$	Zero	N/A

Table 1: This table summarizes the different model-based and mode-free state of the art algorithms available in the literature for CMDPs, where φ is the Slater variable in Assumption 3. It is worthy to notice that the lower bound for zero constraint violation and unconstrained problem are different. We note that the proposed algorithm achieves the best sample complexity compared with all other model-free approaches which requires generative model and achieves zero constraint violation at the same time. For the works considering different setting such as episodic setting, we provide a detailed method to convert the result to the form of sample complexity in infinite horizon setup in Appendix A.1.

Model-free CRL. As compared to the model-based algorithms, existing results for the model-free algorithms are fewer. The constrained policy optimization (CPO) algorithm is proposed in (Achiam et al., 2017) and reward constrained policy optimization (RCPO) al-

1. (Efroni et al., 2020) used $\mathcal{N}$, which is the maximum number of non-zero transition probabilities across the entire state-action pairs. We bound it by $\mathcal{S}$. Moreover, a factor of $\sqrt{|\mathcal{A}|}$ is missed in their result, which we believe is a typo in their work.
2. (Kalagarla et al., 2021) used C, which is the upper bound on the number of possible successor states for a state-action pair. We bound it by $\mathcal{S}$.
3. We use the result in Theorem 4 in (Ding et al., 2020). Notice that in the Algorithm 2 of their paper, $\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$ samples are necessary for each outer loop.
4. Notice that in line 4 of Algorithm 1 in (Xu et al., 2021), a inner loop with K_{in} iteration is needed for policy evaluation and $K_{in} = \tilde{O}(\frac{T}{(1-\gamma)|\mathcal{S}||\mathcal{A}|})$
5. The dependence on $\mathcal{S}, \mathcal{A}$ is not clear in (Chen et al., 2021b). An estimation for the Q-function is needed in the algorithm. However, the authors didn't include analysis for the estimation.
6. Notice that the value function defined in this paper is a normalized version. Thus, an extra $\frac{1}{(1-\gamma)^2}$ is needed for a fair comparison.

gorithm is proposed in (Tessler et al., 2018). Moreover, in (Gattami et al., 2021), it 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 et al., 2021) to achieve a sublinear convergence rate to the global optimal policy and a 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 Table 1. Recently, (Chen et al., 2021b) 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.

Concave Utility RL. Another major research area related to constrained RL is concave utility RL. A special case of maximizing the entropy is considered in (Hazan et al., 2019). (Kostrikov et al., 2019) considered a KL-divergence minimization for imitation learning. (Bai et al., 2022a; Brantley et al., 2020; Agarwal et al., 2022a; Agarwal & Aggarwal, 2023) considered a concave function of possibly vector rewards. Among these works, (Brantley et al., 2020; Agarwal et al., 2022a; Agarwal & Aggarwal, 2023) proposed a model-based approach and (Bai et al., 2022a) proposed a model-free policy gradient algorithm. (Zhang et al., 2020, 2021; Ying et al., 2023) and this work considered a more general setting, where the objective function is a concave function of the occupancy measure. However, all of the other works did not target zero-constraint violations. Recently, (Agarwal et al., 2022b) proposed model-based algorithms based on optimism and posterior sampling approaches that achieves zero constraint violations. In contrast, our work considers a model-free approach.

3. Problem Formulation

An infinite horizon discounted reward constrained Markov Decision Process (CMDP) 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 i^{th} 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.

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 et al., 2021b)

$$V_{\mathbf{r}}^{\pi}(s) = (1-\gamma)\mathbf{E}\Big[\sum\nolimits_{t=0}^{\infty}\gamma^t \mathbf{r}(s_t,a_t)\Big],$$
$$V_{\mathbf{g}^i}^{\pi}(s) = (1-\gamma)\mathbf{E}\Big[\sum\nolimits_{t=0}^{\infty}\gamma^t \mathbf{g}^i(s_t,a_t)\Big], \tag{2}$$

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 (2) is with respect to the transition dynamics of the environment and the stochastic policy π. The standard CMDP problem considers the problem maximizing value function for reward and satisfying some constraints on value function for cost function, given by

$$\begin{aligned} \max_{\pi} \quad & V_{\mathbf{r}}^{\pi}(s) \\ \text{s. t.} \quad & V_{\mathbf{g}^i}^{\pi}(s) \geq 0 \quad \forall i \in [I], \end{aligned} \tag{3}$$

Next, let us define $\boldsymbol{\lambda}^{\pi} : \mathcal{S}\times\mathcal{A} \to [0,1]$ is known as cumulative discounted occupancy measure under policy π given by

$$\lambda^{\pi}(s,a) = (1-\gamma)\Big(\sum\nolimits_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t=s, a_t=a)\Big), \tag{4}$$

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. Then, the problem in (3) which optimizes over policy space, can be equivalently written in the occupancy measure space (Zhang et al., 2021) (Altman, 1999)[Theorem 3.3] as

$$\begin{aligned} \max_{\boldsymbol{\lambda}\geq \mathbf{0}} \quad & \boldsymbol{\lambda}^T \mathbf{r} \\ s.t. \quad & \boldsymbol{\lambda}^T \mathbf{g}_i \geq 0 \quad \forall i \in [I], \\ & \sum\nolimits_{a\in\mathcal{A}} (\mathbf{I} - \gamma \mathbf{P}_a^T)\boldsymbol{\lambda}_a = (1-\gamma)\boldsymbol{\rho}. \end{aligned} \tag{5}$$

We note that in (5), the objective and constraints are linear with respect to $\boldsymbol{\lambda}$. In this work, we are interested in non-linear objective (concave) and non-linear constraints (convex) which arises frequently in the literature, for instance in maximizing the entropy of state-action distribution (Hazan et al., 2019), imitation learning (Ho & Ermon, 2016), and fairness in multi-agent resource allocation (Margolies et al., 2014). The concave utility constrained optimization problem can be formulated as

$$\begin{aligned} \max_{\boldsymbol{\lambda}\geq \mathbf{0}} \quad & f(\boldsymbol{\lambda}) \\ s.t. \quad & h^i(\boldsymbol{\lambda}) \geq 0 \quad \forall i \in [I], \\ & \sum\nolimits_{a\in\mathcal{A}} (\mathbf{I} - \gamma \mathbf{P}_a^T)\boldsymbol{\lambda}_a = (1-\gamma)\boldsymbol{\rho}, \end{aligned} \tag{6}$$

where f is a known concave objective, $h^i, i \in [I]$ are constraint functions.

In (6), we define $\boldsymbol{\lambda}_a = [\boldsymbol{\lambda}(1,a), \cdots, \boldsymbol{\lambda}(|\mathcal{S}|,a)] \in \mathbb{R}^{|\mathcal{S}|}$ as the a^{th} 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

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

Using Theorem 3.3(c) in (Altman, 1999), we have that if $\boldsymbol{\lambda}^*$ is the optimal solution for the problem in Eq. (6), then $\pi_{\boldsymbol{\lambda}^*}$ will be the corresponding optimal policy.

4. Algorithm Development

Before developing the algorithm, we first describe some assumptions and demonstrate some properties of the objective function and constraint functions in (6).

Assumption 1. *(Concavity) The objective function f and constraint functions $h^i, i \in [I]$ are concave functions with respect to the occupancy measure $\boldsymbol{\lambda}$ on the set Λ.*

Assumption 2. *(Lipschitz) The objective function f and constraint function $h^i, i \in [I]$ are Lipschitz functions with Lipschitz constant L_f and L_h with respect to the occupancy measure $\boldsymbol{\lambda}$ on the set Λ. For simplicity, we assume $L_f \geq 1$ and $L_h \geq 1$ (i.e. use $L'_f = \max\{L_f, 1\}$) Formally, for any $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$*

$$\|f(\boldsymbol{\lambda}) - f(\bar{\boldsymbol{\lambda}})\|_2 \leq L_f \|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2 \tag{8}$$
$$\|h(\boldsymbol{\lambda}) - h(\bar{\boldsymbol{\lambda}})\|_2 \leq L_h \|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2 \tag{9}$$

Under Assumption 1 and 2, we derive the following Lemmas.

Lemma 1. *(Shalev-Shwartz et al., 2011)[Lemma 2.6] The gradient of objective function and constraint function are bounded by their Lipschitz constants on the set Λ. Formally,*

$$\|\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda})\|_2 \leq L_f, \forall \boldsymbol{\lambda} \in \Lambda$$
$$\|\nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda})\|_2 \leq L_h, \forall \boldsymbol{\lambda} \in \Lambda, \forall i \in [I].$$

Lemma 2. *The objective function and constraint functions are bounded by a constant on the set Λ, respectively. Without loss of generality, we assume they are bounded by 1.*

Proof. Define $\bar{\boldsymbol{\lambda}} = \frac{1}{|\mathcal{S}||\mathcal{A}|}\mathbf{e}$, where $\mathbf{e}$ is one vector. By Assumption 2, we have for any $\boldsymbol{\lambda} \in \Lambda$

$$\|f(\boldsymbol{\lambda}) - f(\bar{\boldsymbol{\lambda}})\|_2 \leq L_f \|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2 \leq L_f \sqrt{|\mathcal{A}||\mathcal{S}|}.$$

Thus, we can write $\|f(\boldsymbol{\lambda})\|_2 \leq L_f \sqrt{|\mathcal{A}||\mathcal{S}|} + f(\bar{\boldsymbol{\lambda}})$. □

Assumption 3. *(Strict feasibility) There exists a strictly feasible occupancy measure $\hat{\boldsymbol{\lambda}} \geq 0$ to problem in (11) such that*

$$\begin{aligned} h^i(\hat{\boldsymbol{\lambda}}) - \varphi \geq 0 \quad \forall i \in [I] \\ \sum_a (\mathbf{I} - \gamma \mathbf{P}_a^T)\hat{\boldsymbol{\lambda}}_a = (1-\gamma)\boldsymbol{\rho} \end{aligned} \tag{10}$$

for some $0 < \varphi < 1$.

Remark 2. *Assumption 3 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 et al., 2012; Akhtar et al., 2021) and also helps to ensure the boundedness of dual variables (see Lemma 3).*

The problem in (6) is well studied in the literature for the linear objectives and constraints. In this work, we consider concave utilities and the aim 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 (Mahdavi et al., 2012; Akhtar et al., 2021) and utilize it to propose a conservative version of the constrained problem with general utility function in (6) as

$$\max_{\boldsymbol{\lambda}\geq \mathbf{0}} \quad f(\boldsymbol{\lambda}) \tag{11a}$$

$$\text{s.t.} \quad h^i(\boldsymbol{\lambda}) \geq \kappa \quad \forall i \in [I], \tag{11b}$$

$$\sum_{a\in\mathcal{A}} (\mathbf{I} - \gamma \mathbf{P}_a^T)\boldsymbol{\lambda}_a = (1-\gamma)\boldsymbol{\rho}, \tag{11c}$$

where κ is the tuning parameter that 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. It should be noticed that κ and φ are two different concepts. κ is an artificially added parameter, while φ is the intrinsic property of the original problem. Moreover, By the assumption 3, it is natural to see that $0 < \kappa < \varphi < 1$ and we will specify the specific value of the parameter κ later in the convergence analysis section (cf. Sec. 5).

With Assumption 1, note that the conservative version of the problem in Eq. (11) is still a convex programming and hence the strong duality holds under Slater condition in Assumption 3, which motivates us to develop the primal-dual based algorithms to solve the problem in (11). By the KKT theorem, the problem in Eq. (11) is equivalent to the following a saddle point problem which we obtain by writing the Lagrangian of (11) as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) =& f(\boldsymbol{\lambda}) + \sum\nolimits_{i\in[I]} u^i \left(h^i(\boldsymbol{\lambda}) - \kappa\right) \\ &+ (1-\gamma)\left\langle \boldsymbol{\rho}, \mathbf{v}\right\rangle + \sum_{a\in\mathcal{A}} \boldsymbol{\lambda}_a^T(\gamma \mathbf{P}_a - \mathbf{I})\mathbf{v} \\ =& f(\boldsymbol{\lambda}) + \left\langle \mathbf{u}, \mathbf{h}^T(\boldsymbol{\lambda}) - \kappa \mathbf{1}\right\rangle \\ &+ (1-\gamma)\left\langle \boldsymbol{\rho}, \mathbf{v}\right\rangle + \sum_{a\in\mathcal{A}} \boldsymbol{\lambda}_a^T(\gamma \mathbf{P}_a - \mathbf{I})\mathbf{v},\end{aligned} \tag{12}$$

where $\mathbf{u} := [u_1, u_2, \cdots, u^i]^T$ is a column vector of the dual variable corresponding to constraints in (11b), $\mathbf{v}$ is the dual variable corresponding to equality constraint in (11c) and $\mathbf{h} := [\mathbf{h}^1, \cdots, \mathbf{h}^I]$ collects all the h^i's corresponding to I constraints in (11b), and $\mathbf{1}$ is the all one column vector. From the Lagrangian in (12), the equivalent saddle point problem is given by

$$\max_{\boldsymbol{\lambda}\in\Lambda} \min_{\mathbf{u}\geq 0, \mathbf{v}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}). \tag{13}$$

Since the Lagrange function is concave w.r.t. primal and convex w.r.t dual variables, it is known that the saddle point can be solved by the primal-dual gradient descent (Nedić & Ozdaglar, 2009). However, since we assume that the transition dynamics P_a is unknown, then directly evaluating gradients of Lagrangian in (13) 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 (13) as

$$\begin{aligned}\mathcal{L}^{\boldsymbol{\zeta}}_{(s,a,s'),s_0}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v}) &= (1-\gamma)\mathbf{v}(s_0)+\mathbf{1}_{\boldsymbol{\zeta}(s,a)>0}\cdot\frac{\boldsymbol{\lambda}(s,a)[\gamma\mathbf{v}(s')-\mathbf{v}(s)-M_1]}{\boldsymbol{\zeta}(s,a)} \\ &\quad + f(\boldsymbol{\lambda})+\langle\mathbf{u},\mathbf{h}(\boldsymbol{\lambda})-\kappa\mathbf{1}\rangle-M_2\boldsymbol{\lambda},\end{aligned} \tag{14}$$

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 the stochastic approximation $\mathcal{L}^{\boldsymbol{\zeta}}_{(s,a,s'),s_0}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v})$ in (18) is an unbiased estimator for the Lagrangian function in Eq. (12) if we omit the constant M_1 and M_2, which implies that $\mathbb{E}_{\boldsymbol{\zeta}\times\mathbf{P}(\cdot|s,a),\boldsymbol{\rho}}[\mathcal{L}^{\boldsymbol{\zeta}}_{(s,a,s'),s_0}]=\mathcal{L}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v})+M_1+M_2$ 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

$$\begin{aligned}\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v}) &= \mathbf{1}_{\boldsymbol{\zeta}(s,a)>0}\cdot\frac{\gamma\mathbf{v}(s')-\mathbf{v}(s)-M_1}{\boldsymbol{\zeta}(s,a)}\cdot\mathbf{E}_{sa} \\ &\quad+\nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})+\sum_{i\in[I]}u^i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})-M_2\mathbf{1},\end{aligned} \tag{15}$$

$$\hat{\nabla}_{\mathbf{u}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v})=\mathbf{h}(\boldsymbol{\lambda})-\kappa\mathbf{1}, \tag{16}$$

$$\hat{\nabla}_{\mathbf{v}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v})=\mathbf{e}(s_0{}')+\mathbf{1}_{\boldsymbol{\zeta}(s,a)>0}\cdot\frac{\boldsymbol{\lambda}(s,a)(\gamma\mathbf{e}(s')-\mathbf{e}(s))}{\boldsymbol{\zeta}(s,a)}, \tag{17}$$

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 s_0^{th} 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. We remark that M_1 and M_2 in (15) is a shift parameter that is used in the convergence analysis.

Remark 3. *We note that the special case presented in (Bai et al., 2022a) for CMDP uses a similar primal-dual method as follow.*

$$\begin{aligned}&\mathcal{L}^{\boldsymbol{\zeta}}_{(s,a,s'),s_0}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v}) \\ &\quad=(1-\gamma)\mathbf{v}(s_0)+\mathbf{1}_{\boldsymbol{\zeta}(s,a)>0}\cdot\frac{\boldsymbol{\lambda}(s,a)(Z_{sa}-M)}{\boldsymbol{\zeta}(s,a)}-\textstyle\sum_{i\in[I]}\kappa u_i,\end{aligned} \tag{18}$$

where

$$Z_{sa}:=\mathbf{r}(s,a)+\gamma\mathbf{v}(s')-\mathbf{v}(s)+\sum_{i\in[I]}u_i\mathbf{g}^i(s,a), \tag{19}$$

However, the approximated Lagrange function defined in (18) *is different from the above equations. It can be noticed that the above approach extended to general functions leads to a biased estimator of the gradient of approximated Lagrange due to the nonlinear function f and g. This biased estimation will make the analysis much more challenging due to the analysis in Appendix C.8 and C.11 requiring unbiasedness. Thus, in this paper, we redefine the approximated Lagrange function, where we only sample for transition function but not together with objectives or constraints. The estimator in* (15) *is an unbiased estimator for the gradient with respect to $\boldsymbol{\lambda}$.*

Remark 4. *It should be noticed that despite the proposed estimator having a bounded second-order moment, the standard analysis of the stochastic optimization will lead to an extra factor of $\mathcal{O}(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{\delta}})$. This is because for a given state and action pair (s,a) with $\boldsymbol{\zeta}(s,a) \geq 0$*

$$\begin{aligned}
&\mathbf{E}\left[\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v})(s,a)\right]^2 \\
=&\mathbf{E}_{s,a,s'}\left[\mathbf{1}_{\boldsymbol{\zeta}(s,a)>0}\cdot\frac{\gamma\mathbf{v}(s')-\mathbf{v}(s)-M_1}{\boldsymbol{\zeta}(s,a)}+\nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})(s,a)+\sum_{i\in[I]}u^i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})(s,a)-M_2\right]^2
\end{aligned}$$

As for the first item,

$$\begin{aligned}
&\mathbf{E}_{s,a,s'}\left[\mathbf{1}_{\boldsymbol{\zeta}(s,a)>0}\cdot\frac{\gamma\mathbf{v}(s')-\mathbf{v}(s)-M_1}{\boldsymbol{\zeta}(s,a)}\right]^2 && (20)\\
=&\mathbf{E}_{s'}\left[\boldsymbol{\zeta}(s,a)\cdot\left(\frac{\gamma\mathbf{v}(s')-\mathbf{v}(s)-M_1}{\boldsymbol{\zeta}(s,a)}\right)^2\right] && (21)\\
=&\mathbf{E}_{s'}\left[\frac{[\gamma\mathbf{v}(s')-\mathbf{v}(s)-M_1]^2}{(1-\delta)\boldsymbol{\lambda}(s,a)+\delta\frac{\mathbf{1}}{|\mathcal{S}||\mathcal{A}|}}\right]\\
\leq&\frac{4}{\delta}|\mathcal{S}||\mathcal{A}|M_1^2=:\sigma^2
\end{aligned}$$

where we can find the bound of the second moment has a dependence on $\frac{|\mathcal{S}||\mathcal{A}|}{\delta}$. By the result of standard stochastic optimization analysis (Juditsky, Nemirovski, & Tauvel, 2011)[Corollary 1], the convergence rate has a dependence on σ, which finally leads to an extra order of $\mathcal{O}(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{\delta}})$. To solve this problem, we use the KL divergence to regularize the occupancy measure updates. By using KL divergence, we do not require to bound the second moment, but need to bound $\mathbf{E}\left[\sum_{sa}\boldsymbol{\lambda}(s,a)\Delta_{s,a}^2\right]$ where $\Delta_{s,a}$ is the $(s,a)^{th}$ element of $\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v})(s,a)$ (Lemma 6). Hence, unsampled (or less sampled) state-action pairs do not contribute to the update. However, one still needs to ensure that the initial distribution over state action pairs support all state action pairs (Lemma 7 and Appendix C.6).

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

Algorithm 1 **C**onservative **S**tochastic **P**rimal-**D**ual **A**lgorithm (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: $\quad \boldsymbol{\zeta}^t := (1-\delta)\boldsymbol{\lambda}^t + \frac{\delta}{|\mathcal{S}||\mathcal{A}|}\mathbf{1}$

4: $\quad$ Sample $(s_t, a_t) \sim \boldsymbol{\zeta}^t$ and $s_0 \sim \boldsymbol{\rho}$

5: $\quad$ Sample $s_t' \sim \mathcal{P}(\cdot|a_t, s_t)$ from the generative model and observe reward r_{sa}

6: $\quad$ Update value functions as $\mathbf{u}$ and $\mathbf{v}$ as

$$\mathbf{u}^{t+1} = \Pi_{\mathcal{U}}(\mathbf{u}^t - \alpha\hat{\nabla}_{\mathbf{u}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)) \tag{22}$$

$$\mathbf{v}^{t+1} = \Pi_{\mathcal{V}}(\mathbf{v}^t - \alpha\hat{\nabla}_{\mathbf{v}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)) \tag{23}$$

7: $\quad$ Update occupancy measure as

$$\boldsymbol{\lambda}^{t+\frac{1}{2}} = \arg\max_{\boldsymbol{\lambda}} \left\langle \hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t \right\rangle - \frac{1}{\beta} KL(\boldsymbol{\lambda}\|\boldsymbol{\lambda}^t) \tag{24}$$

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^{t+\frac{1}{2}} / \|\boldsymbol{\lambda}^{t+\frac{1}{2}}\|_1 \tag{25}$$

8: **end for**

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 3 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.

5. 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 (13). Then we show that the output of Algorithm 1 given by $\bar{\boldsymbol{\lambda}}$ is ϵ-optimal for the conservative version of the dual domain optimization problem in (11) 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 $\bar{\boldsymbol{\lambda}}$ is also ϵ-optimal and achieves zero constraint violation at the same time.

5.1 Convergence Analysis for Duality Gap

In order to bound the duality gap, we note that the standard analysis of saddle point algorithms (Nedić & Ozdaglar, 2009; Akhtar et al., 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. Define $(\boldsymbol{\lambda}_\kappa^*, \mathbf{u}_\kappa^*, \mathbf{v}_\kappa^*)$ as the solution of saddle-point problem in Eq. (13). Notice that the optimal primal and dual variables are the function of conservative variable κ. When $\kappa = 0$ which means we are considering the original problem in Eq. (6), we omit the subscript κ and denote optimal primal and dual variables as $(\boldsymbol{\lambda}^*, \mathbf{u}^*, \mathbf{v}^*)$. We start the analysis by consider the form of Slater's condition in Assumption 3, and show that the dual variables $\mathbf{u}$ and $\mathbf{v}$ are bounded.

Lemma 3 (Bounded dual variable $\mathbf{u}$ and $\mathbf{v}$)**.** *Under the Assumption 3, the optimal dual variables $\mathbf{u}_\kappa^*$ and $\mathbf{v}_\kappa^*$ are bounded. Formally, it holds that $\|\mathbf{u}_\kappa^*\|_1 \leq \frac{4L_f}{\varphi}$ and $\|\mathbf{v}_\kappa^*\|_\infty \leq \frac{L_f}{1-\gamma} + \frac{4L_f L_h}{(1-\gamma)\varphi}$.*

The proof of Lemma 3 is provided in Appendix C.1. As a result, we define $\mathcal{U} := \left\{\mathbf{u} \mid \|\mathbf{u}\|_1 \leq \frac{8L_f}{\varphi}\right\}$ and $\mathcal{V} := \left\{\mathbf{v} \mid \|\mathbf{v}\|_\infty \leq 2[\frac{L_f}{1-\gamma} + \frac{4L_f L_h}{(1-\gamma)\varphi}]\right\}$.

Since we have mathematically defined the set $\mathcal{U}$ and $\mathcal{V}$, now we rewrite the saddle point formulation in (13) as

$$\max_{\boldsymbol{\lambda}\in\Lambda} \min_{(\mathbf{u}\in\mathcal{U}, \mathbf{v}\in\mathcal{V})} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}). \tag{26}$$

In the analysis presented next, we will work with the problem in (26). First, we decompose the duality gap in Lemma 4 as follows.

Lemma 4 (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*

$$\mathcal{L}(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*) - \mathcal{L}(\mathbf{u}, \mathbf{v}, \bar{\boldsymbol{\lambda}}) \leq \frac{1}{T}\sum_{t=1}^{T}\Bigg[\underbrace{\left\langle \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{w}^t, \boldsymbol{\lambda}^t), \boldsymbol{\lambda}_{\boldsymbol{\kappa}}^* - \boldsymbol{\lambda}^t \right\rangle}_{(I)} + \underbrace{\left\langle \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}^t, \boldsymbol{\lambda}^t), \mathbf{w}^t - \mathbf{w} \right\rangle}_{(II)}\Bigg]. \tag{27}$$

The bound on terms (I) and (II) in the statement of Lemma 4 are provided in Lemma 6 and 7 in the Appendix C.3 (see proofs in Appendix C.4 and C.5, respectively). 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 $\boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*$ is the best solution for the conservative Lagrange problem. The duality gap of the Algorithm 1 is bounded as*

$$\mathbb{E}[\mathcal{L}(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*) - \mathcal{L}(\mathbf{u}^\dagger, \mathbf{v}^\dagger, \bar{\boldsymbol{\lambda}})] \leq \mathcal{O}\left(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{(1-\gamma)\varphi}\right). \tag{28}$$

The proof of Theorem 1 is provided in Appendix C.3. The result in Theorem 1 describes a sub-linear 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.

5.2 Dual Objective and Constraint Violation

Recall that the saddle point problem in Eq. (26) 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. (11). But before that, we need to characterize the gap between the original problem (6) and its conservative version in (11). The following Lemma 5 shows that the gap is of the order of parameter κ.

Lemma 5. *Under Assumption 3, 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 5 is provided in Appendix D.1. Using the statement of Lemma 5 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*

$$T \geq \max\left\{16, 4\varphi^2, \frac{1}{\epsilon^2}\right\} \cdot \tilde{c}_1^2 \frac{L_f^2 L_h^2 I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2\varphi^2}, \tag{29}$$

and we set

$$\kappa = \frac{2L_f L_h \tilde{c}_1}{1-\gamma}\sqrt{\frac{I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{T}}, M_1 = 4\left[\frac{L_f}{1-\gamma} + \frac{4L_f L_h}{(1-\gamma)\varphi}\right], M_2 = L_f + \frac{8L_f L_h}{\varphi},$$

then the constraints of the original problem in (6) satisfy:

$$\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq \epsilon\varphi \quad \forall i \in [I], \tag{30a}$$

$$\mathbb{E}\Big\| \sum_a (\gamma \mathbf{P}_a^T - \mathbf{I})\bar{\boldsymbol{\lambda}}_a + (1-\gamma)\boldsymbol{\rho}\Big\|_1 \leq \frac{(1-\gamma)\epsilon\varphi}{L_f L_h}. \tag{30b}$$

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

$$\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] \leq 3\epsilon. \tag{31}$$

The proof of Theorem 2 is provided in Appendix D.2. Next, we present the special case of Theorem 2 in the form of Corollary 1 (see proof in Appendix D.3), 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 $T \geq \tilde{c}_1^2 \cdot \frac{L_f^2 L_h^2 I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2\varphi^2\epsilon^2}$ then $\bar{\boldsymbol{\lambda}}$ satisfies the constraint violation as*

$$\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq -\epsilon \quad \forall i \in [I] \tag{32a}$$

$$\mathbb{E}\Big\| \sum_a (\gamma \mathbf{P}_a^T - \mathbf{I})\bar{\boldsymbol{\lambda}}_a + (1-\gamma)\boldsymbol{\rho}\Big\|_1 \leq \frac{(1-\gamma)\epsilon\varphi}{L_f L_h}, \tag{32b}$$

and the sub-optimality is given by $\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] \leq \epsilon$.

The positive lower bound of $\epsilon\varphi$ in (30a) hints that $\bar{\boldsymbol{\lambda}}$ is feasible (hence zero constraint violation). On the other hand, the lower bound in (32a) 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.

5.3 Convergence Analysis in Policy Space

We have established the convergence in the occupancy measure space in Sec. 5.2 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. (30b) makes $\bar{\boldsymbol{\lambda}}$ to loose its physical meaning as discussed in Proposition 1 in (Zhang et al., 2021). 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. (30b) 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*

$$f(\boldsymbol{\lambda}^*) - \mathbb{E}[f(\boldsymbol{\lambda}^{\bar{\pi}})] \leq \epsilon \tag{33a}$$

$$\mathbb{E}[h^i(\boldsymbol{\lambda}^{\bar{\pi}})] \geq 0 \quad \forall i \in [I]. \tag{33b}$$

The proof of Theorem 3 is provided in Appendix E.1. To get better idea about the importance of result in Theorem 3, we next present a Corollary 2 (see proof in E.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,*

$$f(\boldsymbol{\lambda}^*) - \mathbb{E}[f(\boldsymbol{\lambda}^{\bar{\pi}})] \leq \epsilon \tag{34a}$$

$$\mathbb{E}[h^i(\boldsymbol{\lambda}^{\bar{\pi}})] \geq -\epsilon \quad \forall i \in [I]. \tag{34b}$$

The benefit of utilizing the conservation parameter κ becomes clear after comparing the results in (33b) and (34b).

6. Empirical Evaluations

In this section, we evaluate the proposed CSPDA algorithm, on three different environments. For the first environment, as considered by (Liu et al., 2021b), we construct a random MDP. The second environment is a grid world environment where the agent needs to cross a border to reach a goal state and the fastest route is unsafe and there exists another route which is safe but longer (Paternain et al., 2019). The third environment is a queuing system with a single server in discrete time (Altman, 1999, Chapter 5) as considered in (Agarwal et al., 2022c; Gattami et al., 2021). We now provide the experimental details and simulation results separately for each of the environments.

6.1 Random MDPs

The random MDP has 100 states with transition probabilities sampled from a Dirichlet distribution. The rewards $r(s, a)$ are sampled from a uniform distribution over $[0, 1)$ and

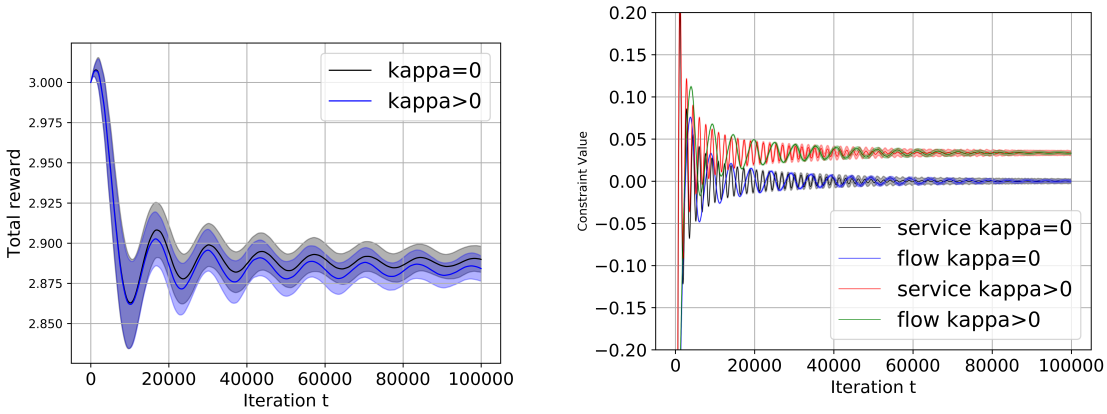

Figure 1: Learning Process of the proposed algorithm for **linear** objective and constraint value with $\kappa = 0$ and $\kappa > 0$. The total reward is the objective in (37) with $c = 0$ and the constraint value is the L.H.S. of the constraint in (37).

the costs $c(s, a)$ are sampled from a uniform distribution over $[-0.5, 0.5)$. The goal of the agent is to maximize the average reward $\lambda^T r$ while ensuring the average cost parameter $\lambda^T c$ is at least 0.

We sample a single MDP and run 100 independent runs of the CSPDA algorithm on the sampled MDP. For this example, we choose the value of $T = 10000$ and the step sizes α and β are set in accordance to Section C.3 as:

$$\alpha = \frac{L_f L_h \sqrt{|\mathcal{S}|}}{(1-\gamma)\phi\sqrt{TI}} \tag{35}$$

$$\beta = \frac{(1-\gamma)\phi}{L_f L_h} \sqrt{\frac{\log(|\mathcal{S}|\mathcal{A}|)}{T|\mathcal{S}|\mathcal{A}|}} \tag{36}$$

with value $|\mathcal{S}| = 100$, $|\mathcal{A}| = 4$, with $L_f = L_h = 1$ as we consider a linear setup of maximum reward and cost bounded by 1. Further, since we have only one cost function, we have $I = 1$. Finally, we set the value of $\phi = 0.48$ and the value of $\delta = 0.01$.

We present the simulation results for the proposed CSPDA algorithm on the random MDP in Figure 2. We note that the choice of κ plays a significant role in the performance of the algorithm. The objective value of the average rewards is higher, but not significantly, for $\kappa = 0$. However, when comparing the average cost values, the implementation with $\kappa > 0$ performs significantly better showing the role tuning κ can play in obtaining the performance of the learnt policy.

6.2 Gridworld Environment

We next evaluate the proposed algorithm on a 15×15 gridworld environment. The agent starts from a fixed position on the map and can move in 4 directions if permitted. The agent aims to cross the room and reach the goal state as soon as possible to obtain some reward. The map of the gridworld is presented in Figure 3. The agent does not receive any

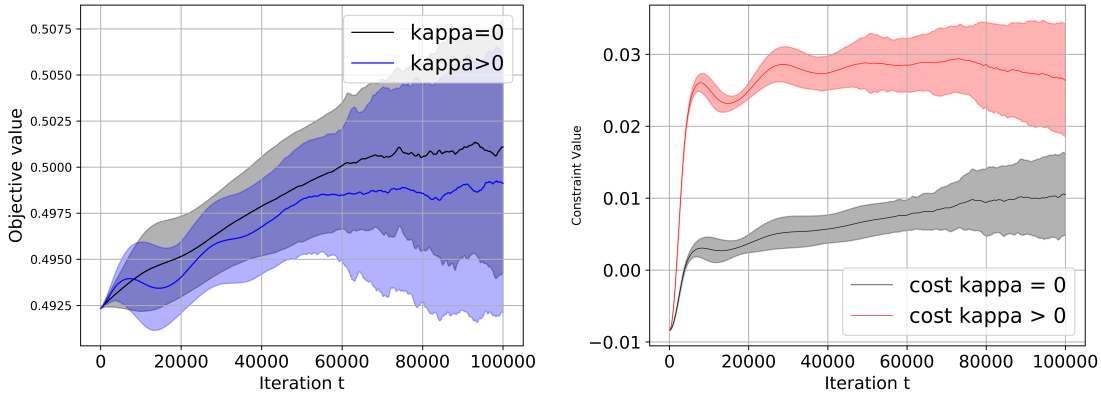

Figure 2: Learning Process of the proposed algorithm for objective and constraint value with $\kappa = 0$ and $\kappa > 0$ evaluated on random MDP with 100 states and 4 actions.

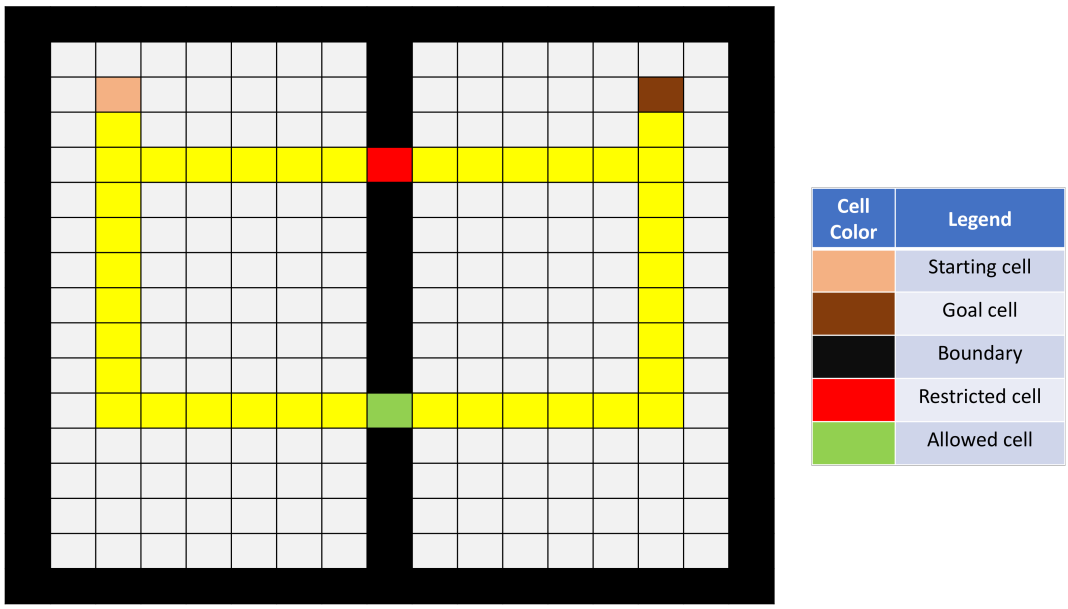

Figure 3: Map of the gridworld environment. The agent has to reach the goal state to obtain a reward of 1 unit. If the agent crosses the red cell, it incurs a cost of -1 unit whereas the agent can cross the green cell without incurring any cost.

reward till it reaches the goal state. After the agent reaches the goal state, the agent does not change the state and receives a reward of unit 1 till eternity. The room has a wall in the middle which separates the starting cell and the goal state. The wall has two openings, one of which is restricted. If the agent used the restricted cell, it can reach the goal faster. However, it received a penalty in terms of a cost of -1. We aim to not allow the agent pass through the restricted cell, and thus, the average cost should be non-negative.

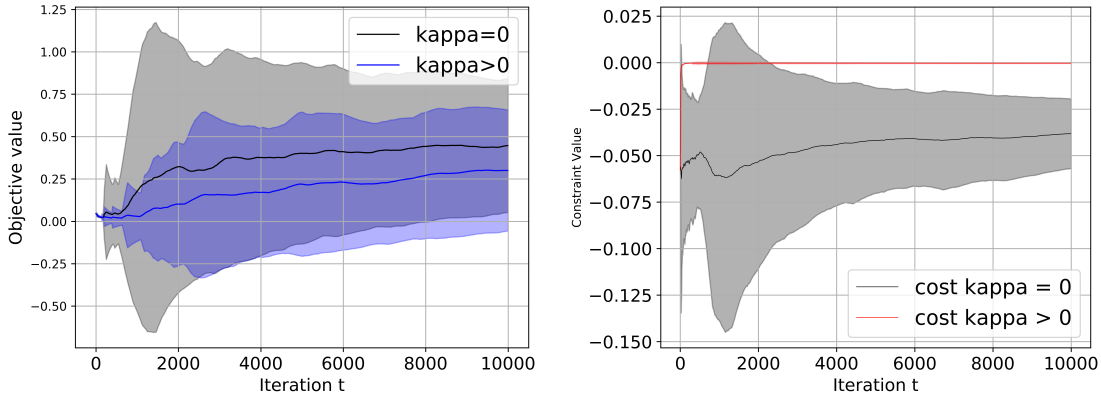

Figure 4: Learning Process of the proposed algorithm for objective and constraint value with $\kappa = 0$ and $\kappa > 0$ evaluated on random MDP with 225 states and 4 actions.

We again run 100 independent runs of the CSPDA algorithm on the sampled MDP. We present present the simulation results for the proposed CSPDA algorithm on the gridworld in Figure 4. We set the value of $\beta = 0.0001$ and $\alpha = 10$. We again note that the choice of κ plays a significant role in the performance of the algorithm. The objective value of the average rewards is higher, but not significantly, for $\kappa = 0$. However, when comparing the average cost values, the implementation with $\kappa > 0$ performs significantly better showing the role tuning κ can play in obtaining the performance of the learnt policy.

6.3 Evaluations on a Queuing System

In this section, we evaluate the proposed Algorithm 1 on a queuing system with a single queue. 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 2.

Assuming $\gamma = 0.5$, we define the objective function f as total discounted cumulative reward plus entropy regularization. And define two constraints function h^1, h^2 as standard total discounted constraint value with respect to service and flow. Thus, the overall optimization problem is given as

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

Table 2: Transition probability of the queue system

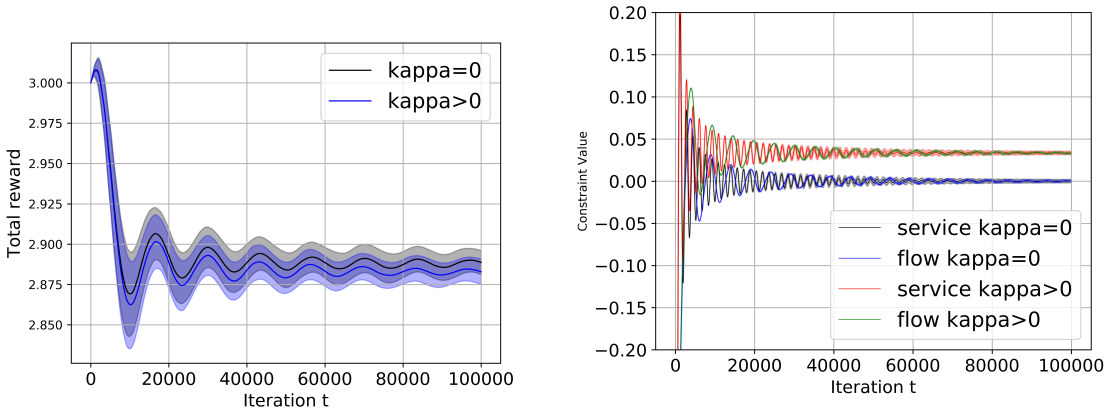

Figure 5: Learning Process of the proposed algorithm for **concave** objective and constraint value with $\kappa = 0$ and $\kappa > 0$. The total reward is the objective in (37) with $c = 1$ and the constraint value is the L.H.S. of the constraint in (37).

$$\max_{\pi} \quad \langle \boldsymbol{\lambda}^{\pi}, \mathbf{r} \rangle - c \sum_{s,a} \lambda^{\pi}_{s,a} \log(\lambda^{\pi}_{s,a}) \tag{37}$$
$$s.t. \quad \left\langle \boldsymbol{\lambda}^{\pi}, \mathbf{g}^{i} \right\rangle \geq 0 \quad i = 1, 2$$

where $s_0 \sim \boldsymbol{\rho}$, π^a and π^b are the policies for the service and flow, respectively. It is not hard to find that the above objective function is concave and Lipschitz. For simulations, we choose $L = 5$, $\mathcal{A} = [0.2, 0.4, 0.6, 0.8]$, 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 $\boldsymbol{\rho}$ is set as uniform distribution. Moreover, the cost function is set to be $r(s, a, b) = -s+5$, the constraint function for the service is defined as $g^1(s, a, b) = -10a+4$, and the constraint function for the flow is $g^2(s, a, b) = -8(1-b)^2 + 1.28$. We run 100 independent simulations and collect the mean value and standard variance. In Fig. 1 and Fig. 5, we set $c = 0$ and $c = 1$, which means they are the standard CMDP problem and concave utility problem, respectively. In each figure, we show the learning process of objective value and constraint value for $\kappa = 0$ and $\kappa > 0$ respectively (in the case of $\kappa > 0$, the value is chosen based on the value in Theorem 2.). Note that the y-axis in Figs. 1 and 5 is the objective function (on left) and the constraint function (on right) defined in Eq. (37). In both the cases, it can be seen that when $\kappa = 0$, the constraint values converge to a small negative number when T goes larger, while for $\kappa > 0$, the constraint values will

converge to a positive value, which matches the result in theory. Further, the objective value are similar for both $\kappa = 0$ and $\kappa > 0$, while the case where $\kappa > 0$ helps to achieve zero constraint violation. Having κ as a hyperparameter in practice can lead to optimal objectives where the constraint violations converge to zero.

7. Conclusion

In this work, we considered the problem of learning optimal policies for infinite-horizon concave constrained Markov Decision Processes (CCMDP) under finite state $\mathcal{S}$ and action $\mathcal{A}$ spaces with I number of constraints. Such constrained reinforcement learning (CRL) with concave utility hasn't been studied 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 $\Omega(\frac{L_f L_h I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2\varphi^2\epsilon^2})$ 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.

We note that while the results in parametrized setup have been studied for concave utility without constraints (Bai et al., 2022a) for linear utility without constraints (Mondal & Aggarwal, 2023), and for linear utility with constraints (Bai, Bedi, & Aggarwal, 2023), corresponding results with concave utility and constraints remain open.

Acknowledgment

The work of A. S. Bedi and A. Koppel was completed when they were with the US Army Research Laboratory, Adelphi, MD, USA. A special case of this paper was presented in part at AAAI, Feb 2022, where the utility and the constraints were linear (Bai et al., 2022b).

Appendix A. Preliminaries

A.1 Explanation of Comparison among References in Table 1

Step 1: From Regret to PAC result

Many references listed in the Table 1 are in the episodic setting and give the result in the form of regret, which is defined as

$$\sum_{k=1}^{K} V_{r,1}^{*}(s_1) - V_{r,1}^{\pi_k}(s_1) \leq f(H, |\mathcal{S}|, |\mathcal{A}|, T, \delta) \quad \textbf{with probability at least } 1-\delta \tag{38}$$

where $T = KH$. The following method provides a probably approximately correct (PAC) result from the regret. At the end of learning horizon K, a policy $\bar{\pi}$ can be defined as follow

$$\bar{\pi}(s) = \begin{cases} \pi_1(s) & \text{with probability } 1/K \\ \dots & \dots \\ \pi_k(s) & \text{with probability } 1/K \\ \dots & \dots \\ \pi_K(s) & \text{with probability } 1/K \end{cases} \tag{39}$$

Note that $\bar{\pi}$ chooses the different policies π^k for $k \in [K]$ uniformly at random. Thus, we know $\frac{1}{K}\sum_{k=1}^{K} V_{r,1}^{\pi_k}(s_1) = V_{r,1}^{\bar{\pi}}(s_1)$. Divide Eq. (38) by K on both side, we have

$$V_{r,1}^{*}(s_1) - V_{r,1}^{\bar{\pi}}(s_1) \leq \frac{f(H, |\mathcal{S}|, |\mathcal{A}|, T, \delta)}{K} \tag{40}$$

If the function f is sub-linear w.r.t. T, then for large enough K, we have $V_{r,1}^{*}(s_1) - V_{r,1}^{\bar{\pi}}(s_1) \leq \epsilon$ with probability at least $1-\delta$, which means that $\bar{\pi}$ is an ϵ-optimal policy.

Step 2: From episodic setting to infinite horizon discounted setting

As mentioned above, many references consider the problem in episodic setting. In order to make a comparison, it is necessary to have a fair conversion. Here, we use the method from (Jin et al., 2018)[footnote 3 in page 3]. Firstly, we check whether the MDP model in the given result assume a horizon dependent transition dynamics, i.e, whether $\mathbf{P}$ is a function of h. If so, then define $\mathcal{S}' = \mathcal{S}H$. If not, then define $\mathcal{S}' = \mathcal{S}$. This conversion is easy to understand and reasonable because an extra H times state space is needed if transition dynamics is different for each h. After this step, we change H to $\frac{1}{1-\gamma}$. This is because the infinite horizon discounted value function can be simulated by the following algorithm.

Algorithm 2 Unbiased estimator for Value Function

Input: Initial distribution $\boldsymbol{\rho}$. Discounted factor γ. Policy π
Output: Value function $V_{r,1}^{\pi}$

```
Sample s_1 ~ ρ, H ~ Geo(1 − γ)
for Each state s_1 in S do
    for h = 1, 2, ..., H do
        Take action a_h ~ π(·|s_h), observe next state s_{h+1} and reward r(s_h, a_h)
    end for
    V^π_{r,1}(s_1) = Σ_{h=1}^{H} r(s_h, a_h)
end for
```

The sample horizon is taken from the geometry distribution with parameter $(1-\gamma)$ and thus the expected length of horizon is $\frac{1}{1-\gamma}$, which explains why it is fair to change H to $\frac{1}{1-\gamma}$. Following these two steps, we convert the result in episodic setting into infinite horizon discounted setting.

Step3: From High Probability result to Expectation result

After converting the result from episodic setting to infinite horizon discounted setting, we get an ϵ-optimal result with probability at least $1-\delta$. However, the result in this paper is in the form of expectation. Thus, we can convert the result with the following method. Notice that the value function V_r is bounded by $\frac{1}{1-\gamma}$, we have

$$\mathbb{E}[V_r^*(s_1) - V_r^\pi(s_1)] \leq \epsilon * (1-\delta) + \delta * \frac{1}{1-\gamma} \tag{41}$$

If $\delta < \epsilon(1-\gamma)$, then, we have $\mathbb{E}[V_r^*(s_1) - V_r^\pi(s_1)] \leq 2\epsilon$.

An example for UC-CFH in (Kalagarla et al., 2021)

In the UC-CFH algorithm, the author proposed an ϵ-optimal result with at most $\tilde{O}(\frac{|\mathcal{S}||\mathcal{A}|C^2H^2}{\epsilon^2}\log(\frac{1}{\delta}))$ episodes, where C is the upper bound on the number of possible successor states for a state-action pair. Thus, $C < |\mathcal{S}|$ and the above equation can be bounded by $\tilde{O}(\frac{|\mathcal{S}|^3|\mathcal{A}|H^2}{\epsilon^2}\log(\frac{1}{\delta}))$. Notice that this is already a PAC result and we begin converting it into infinite horizon discounted setting.

- Firstly, we know $K = \tilde{O}(\frac{|\mathcal{S}|^3|\mathcal{A}|H^2}{\epsilon^2}\log(\frac{1}{\delta}))$ and thus the total sample complexity is $KH = \tilde{O}(\frac{|\mathcal{S}|^3|\mathcal{A}|H^3}{\epsilon^2}\log(\frac{1}{\delta}))$. Notice that UC-CFH algorithm doesn't assume horizon dependent transition dynamics (They assume in the model, however, not in the algorithm and theorem). Thus, by changing H to $\frac{1}{1-\gamma}$, we have sample complexity $\tilde{O}(\frac{|\mathcal{S}|^3|\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\log(\frac{1}{\delta}))$.
- Secondly, change δ to $\epsilon(1-\gamma)$, we get the sample complexity in the form of expectation, which means with $\tilde{O}(\frac{|\mathcal{S}|^3|\mathcal{A}|}{(1-\gamma)^3\epsilon^2})$ sample, we have

$$\mathbb{E}[V_1^*(s_1) - V_1^{\pi_k}(s_1)] \leq \epsilon \tag{42}$$

Appendix B. Notations

For the purpose of analysis in the appendix, we have used the shorthand notation $\boldsymbol{\lambda}_{sa}$ for $\boldsymbol{\lambda}(s,a)$.

Appendix C. Proofs for Section 5.1

C.1 Proof of Lemma 3

Proof. **Bound on $\|\mathbf{u}_\kappa^*\|_1$:** Let us denote the optimal value of optimization problem in (11) as p_κ^* and write the corresponding dual problem as

$$\mathcal{D}_\kappa(\mathbf{u},\mathbf{v}) := \max_{\boldsymbol{\lambda}\geq\mathbf{0}} \mathcal{L}(\boldsymbol{\lambda},\mathbf{u},\mathbf{v}) = \max_{\boldsymbol{\lambda}\geq\mathbf{0}} f(\boldsymbol{\lambda}) + \sum_{i\in[I]} u^i \left(h^i(\boldsymbol{\lambda}) - \kappa\right) + (1-\gamma)\langle\boldsymbol{\rho},\mathbf{v}\rangle + \sum_{a\in\mathcal{A}} \boldsymbol{\lambda}_a^T(\gamma\mathbf{P}_a - \mathbf{I})\mathbf{v} \tag{43}$$

The optimal dual variables are given by

$$(\mathbf{u}_\kappa^*, \mathbf{v}_\kappa^*) := \arg\min_{\mathbf{u}\geq\mathbf{0},\mathbf{v}} \mathcal{D}_\kappa(\mathbf{u},\mathbf{v}), \tag{44}$$

and let us denote the optimal dual value by $d_\kappa^* = \mathcal{D}_\kappa(\mathbf{u}_\kappa^*, \mathbf{v}_\kappa^*)$. We note that the problem in (11) is a convex programming problem. By the Slater condition in the Assumption 3, we know strong duality holds, i.e $p_\kappa^* = d_\kappa^*$. To proceed, let us consider a constant C and define a set $\mathcal{C} := \{(\mathbf{u},\mathbf{v}) \geq \mathbf{0} | \mathcal{D}_\kappa(\mathbf{u},\mathbf{v}) \leq C\}$. For any $(\mathbf{u},\mathbf{v}) \in \mathcal{C}$ and a feasible $\hat{\boldsymbol{\lambda}}$ which satisfies Assumption 3, we could write

$$\begin{aligned} C \geq \mathcal{D}_\kappa(\mathbf{u},\mathbf{v}) \overset{(a)}{\geq} & \mathcal{L}(\hat{\boldsymbol{\lambda}},\mathbf{u},\mathbf{v}) \\ =& f(\hat{\boldsymbol{\lambda}}) + \sum_{i\in[I]} u^i \left(h^i(\hat{\boldsymbol{\lambda}}) - \kappa\right) + (1-\gamma)\langle \boldsymbol{\rho}, \mathbf{v}\rangle + \sum_{a\in\mathcal{A}} \hat{\boldsymbol{\lambda}}_a^T(\gamma \mathbf{P}_a - \mathbf{I})\mathbf{v} \\ \overset{(b)}{\geq} & f(\hat{\boldsymbol{\lambda}}) + \left\langle \mathbf{u}, \frac{\varphi \mathbf{1}}{2} \right\rangle \\ =& f(\hat{\boldsymbol{\lambda}}) + \frac{\varphi}{2}\|\mathbf{u}\|_1, \end{aligned} \tag{45}$$

where step (a) holds by the definition of dual function and step (b) is true by Assumption 3 and $\kappa \leq \frac{\varphi}{2}$. From weak duality, we have

$$D_\kappa(\mathbf{u},\mathbf{v}) \geq d_\kappa^* \geq p_\kappa^* = \langle \boldsymbol{\lambda}_\kappa^*, \mathbf{r}\rangle \tag{46}$$

Now let $C = \langle \boldsymbol{\lambda}^*, \mathbf{r}\rangle$, all inequalities in Eq. (46) become equality for $(\mathbf{u},\mathbf{v}) \in \{(\mathbf{u},\mathbf{v}) \geq \mathbf{0} | \mathcal{D}_\kappa(\mathbf{u},\mathbf{v}) \leq \langle \boldsymbol{\lambda}^*, \mathbf{r}\rangle\}$. Thus, this set is the optimal dual variable set. We set $C = \langle \boldsymbol{\lambda}^*, \mathbf{r}\rangle$ and rearrange the Eq. (45) to obtain

$$\|\mathbf{u}_\kappa^*\|_1 \leq \frac{2[f(\boldsymbol{\lambda}_\kappa^*) - f(\hat{\boldsymbol{\lambda}})]}{\varphi} \overset{(a)}{\leq} \frac{2L_f\|\boldsymbol{\lambda}_\kappa^* - \hat{\boldsymbol{\lambda}}\|_2}{\varphi} \overset{(b)}{\leq} \frac{2L_f[\|\boldsymbol{\lambda}_\kappa^*\|_1 + \|\hat{\boldsymbol{\lambda}}\|_1]}{\varphi} \overset{(c)}{\leq} \frac{4L_f}{\varphi} \tag{47}$$

where the step (a) holds by the Lipschitz Assumption 2. The second step holds by triangle inequality and last step holds because occupancy measure sum up to 1.

Bound on $\|\mathbf{v}_\kappa^*\|_\infty$: To solve the convex programming in (11), the KKT conditions should be sufficient and necessary, which can be written as

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_\kappa^*, \mathbf{u}_\kappa^*, \mathbf{v}_\kappa^*) = 0 \tag{48a}$$

$$h^i(\boldsymbol{\lambda}_\kappa^*) \geq \kappa \quad \forall i \in [I] \tag{48b}$$

$$\sum_a (\mathbf{I} - \gamma \mathbf{P}_a^T)\boldsymbol{\lambda}_{\kappa,a}^* = (1-\gamma)\boldsymbol{\rho} \tag{48c}$$

$$\sum_{i\in[I]} \mathbf{u}_{\kappa,i}^*[h^i(\boldsymbol{\lambda}_\kappa^*) - \kappa] = 0 \tag{48d}$$

$$\mathbf{u}_\kappa^* \geq \mathbf{0} \tag{48e}$$

By Eq. (48a), we have for any state-action pair (s,a)

$$\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{s,a} + \sum_{i\in[I]} u_{\kappa,i}^* \nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda}_\kappa^*)_{s,a} - (\mathbf{e}_s - \gamma \mathbf{P}_{as})^T \mathbf{v}_\kappa^* = 0, \tag{49}$$

where $\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{s,a}$ is the (s,a) element of $\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)$ and $u_{\kappa,i}^*$ is the i^{th} elemnt of vector u_κ^*. $\mathbf{P}_{as}$ is a column vector and $P_{as}(s') = P(s'|a,s)$. Given a fixed action $\bar{a}$, denote $\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{\bar{a}} := [\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{1,\bar{a}}, \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{2,\bar{a}}, \cdots, \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{\mathcal{S},\bar{a}}]^T$, $\nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda}_\kappa^*)_{\bar{a}} := [\nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda}_\kappa^*)_{1,\bar{a}}, \nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda}_\kappa^*)_{2,\bar{a}}, \cdots, \nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda}_\kappa^*)_{\mathcal{S},\bar{a}}]^T$ and $\tilde{\mathbf{P}} := [P_{\bar{a},1}, \cdots, P_{\bar{a},|\mathcal{S}|}] \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$. By Eq. (49), we have

$$(\mathbf{I} - \gamma \tilde{\mathbf{P}}^T)\mathbf{v}_\kappa^* = \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{\bar{a}} + \sum_{i\in[I]} u_{\kappa,i}^* \nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda}_\kappa^*)_{\bar{a}} \tag{50}$$

As a result, we have

$$\begin{aligned} L_f + \frac{4L_f L_h}{\varphi} &\overset{(a)}{\geq} L_f + L_h\|\mathbf{u}_\kappa^*\|_1 \overset{(b)}{\geq} \|\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_\kappa^*)_{\bar{a}} + \sum_{i\in[I]} u_{\kappa,i}^* \nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda}_\kappa^*)_{\bar{a}}\|_\infty = \|(\mathbf{I} - \gamma\tilde{\mathbf{P}}^T)\mathbf{v}_\kappa^*\|_\infty \\ &\overset{(c)}{\geq} \|\mathbf{v}_\kappa^*\|_\infty - \|\gamma\tilde{\mathbf{P}}^T\mathbf{v}_\kappa^*\|_\infty \overset{(d)}{\geq} (1-\gamma)\|\mathbf{v}_\kappa^*\|_\infty, \end{aligned} \tag{51}$$

where the step (a) holds by the Lemma 3, step (b) holds by the definition of r, g_i, step (c) comes from the triangle inequality, and step (d) is true because each row in $\tilde{\mathbf{P}}^T$ adds up to 1. Finally, we have the bound $\|\mathbf{v}_\kappa^*\|_\infty \leq \frac{L_f}{1-\gamma} + \frac{4L_f L_h}{(1-\gamma)\varphi}$. □

C.2 Proof of Lemma 4

Proof. Consider the Lagrangian in (12) and note that it is convex w.r.t $\mathbf{u}$ as well as $\mathbf{v}$. w.r.t The gradient of the Lagrange function $\mathbf{u}$ and $\mathbf{v}$ are given by

$$\begin{aligned} \nabla_{\mathbf{u}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) &= \mathbf{h}(\boldsymbol{\lambda}) - \kappa \mathbf{1}, \\ \nabla_{\mathbf{v}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) &= (1-\gamma)\boldsymbol{\rho} + \sum_a (\gamma \mathbf{P}_a^T - \mathbf{I})\boldsymbol{\lambda}_a. \end{aligned} \tag{52}$$

It is obvious that $\nabla_{\mathbf{u}}^2 \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \nabla_{\mathbf{u},\mathbf{v}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \nabla_{\mathbf{v},\mathbf{u}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \nabla_{\mathbf{v}}^2 \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v}) = \mathbf{0}$, which means that the Hessian matrix $\nabla_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{v})$ is a zero matrix. Thus, Lagrange function is convex w.r.t $\mathbf{w}$. Then, let us define $\mathbf{w} = [\mathbf{u}^T, \mathbf{v}^T]^T$, $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^T \bar{\mathbf{w}}_t$, and decompose the duality gap as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}_\kappa^*, \bar{\mathbf{u}}, \bar{\mathbf{v}}) - \mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{u}, \mathbf{v}) &= \mathcal{L}(\boldsymbol{\lambda}_\kappa^*, \bar{\mathbf{w}}) - \mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{w}) \\ &\overset{(a)}{\leq} \frac{1}{T}\sum_{t=1}^T \big[\mathcal{L}(\boldsymbol{\lambda}_\kappa^*, \mathbf{w}^t) - \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w})\big] \\ &= \frac{1}{T}\sum_{t=1}^T \big[\mathcal{L}(\boldsymbol{\lambda}_\kappa^*, \mathbf{w}^t) - \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t) + \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t) - \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w})\big] \\ &\overset{(b)}{\leq} \frac{1}{T}\sum_{t=1}^T \big[\, \big\langle \nabla_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \boldsymbol{\lambda}_\kappa^* - \boldsymbol{\lambda}^t \big\rangle + \big\langle \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \mathbf{w}^t - \mathbf{w} \big\rangle \,\big], \end{aligned} \tag{53}$$

where step (a) holds by Jensen inequality and the step (b) utilizes the convexity of $\mathcal{L}(\boldsymbol{\lambda}, \cdot)$ and concavity of $\mathcal{L}(\cdot, \mathbf{w})$. □

C.3 Proof of Theorem 1

Proof. We collect the dual variables $\mathbf{u}$ and $\mathbf{v}$ in one variable $\mathbf{w}$ as defined in Lemma 4 for the ease of analysis. The next two Lemmas provide the bound on the terms I and II in Eq. (27).

Lemma 6. *Let the iterate sequence $\{\boldsymbol{\lambda}^t\}$ be updated as mentioned in the updates (24) and (25) of Algorithm 1, then for any t it holds that*

$$\begin{aligned}\left\langle \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t \right\rangle \quad &\leq \frac{1}{\beta}\big[KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^t) - KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^{t+1})\big] + \frac{\beta}{2}\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2 \\ &\quad + \left\langle \hat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \boldsymbol{\lambda}^t - \boldsymbol{\lambda} \right\rangle. \end{aligned} \tag{54}$$

Lemma 7. *Define $\mathcal{W} = \mathcal{U} \times \mathcal{V}$ and consider the iterate sequence $\{\mathbf{w}^t\}$ updated according to the rule Eq. (22) and (23) in Algorithm 1. For any t, it holds that*

$$\begin{aligned}\left\langle \nabla_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \mathbf{w}^t - \mathbf{w} \right\rangle \leq \frac{1}{2\alpha}\Bigg[&\|\mathbf{w}^t - \mathbf{w}\|^2 - \|\mathbf{w}^{t+1} - \mathbf{w}\|^2 + \alpha^2 \|\hat{\nabla}_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t)\|^2 \\ &+ 2\alpha \left\langle \nabla_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{w}) - \hat{\nabla}_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{w}), \mathbf{w}^t - \mathbf{w} \right\rangle \Bigg]. \end{aligned} \tag{55}$$

Next, utilizing the results of Lemma 6 and 7 (see proofs in Appendix C.4 and C.5) into Lemma 4, we prove the main result in Theorem 1, which establishes the final bound on the duality gap as follows. Let $\boldsymbol{\lambda} = \boldsymbol{\lambda}_\kappa^*$ in Eq. (54) and $(\mathbf{u}^\dagger, \mathbf{v}^\dagger) := \arg\min_{\mathbf{u},\mathbf{v}} \mathcal{L}(\mathbf{u}, \mathbf{v}, \bar{\boldsymbol{\lambda}})$ in Eq. (55). Then, sum up Eq. (54) and (55) from $t = 1$ to T , we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\left[\left\langle \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \boldsymbol{\lambda}_\kappa^* - \boldsymbol{\lambda}^t \right\rangle + \left\langle \nabla_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\dagger \right\rangle \right] \\ &\quad \leq \underbrace{\frac{KL(\boldsymbol{\lambda}_\kappa^*||\boldsymbol{\lambda}^1)}{T\beta}}_{T_1} + \underbrace{\frac{\beta}{2T}\sum_{t=1}^{T}\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2}_{T_2} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left\langle \hat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \boldsymbol{\lambda}^t - \boldsymbol{\lambda}_\kappa^* \right\rangle}_{T_3} \\ &\qquad + \underbrace{\frac{1}{2T\alpha}\|\mathbf{w}^1 - \mathbf{w}^\dagger\|^2}_{T_4} + \underbrace{\frac{\alpha}{2T}\sum_{t=1}^{T}\|\hat{\nabla}_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t)\|^2}_{T_5} + \underbrace{\sum_{t=1}^{T}\left\langle \nabla_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t) - \hat{\nabla}_{\mathbf{w}} \mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\dagger \right\rangle}_{T_6} \end{aligned} \tag{56}$$

Combine the above result with the statement of Lemma. 4 to write

$$\mathbb{E}[\mathcal{L}(\boldsymbol{\lambda}_\kappa^*, \bar{\mathbf{u}}, \bar{\mathbf{v}}) - \mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{u}^\dagger, \mathbf{v}^\dagger)] \leq \sum_{j=1}^{6}\mathbb{E}[T_j]. \tag{57}$$

We derive an upper bound on the right hand side of (57) in Appendix C.6-C.11. Following the results in Appendix C.6-C.11, we have

$$\begin{aligned}
&\mathbb{E}[T_1] \leq \frac{\log(|\mathcal{S}||\mathcal{A}|)}{T\beta}, \quad \mathbb{E}[T_2] \leq \frac{4000\beta L_f^2 L_h^2 |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2\varphi^2} \quad \mathbb{E}[T_3] = 0, \\
&\mathbb{E}[T_4] \leq \frac{400|\mathcal{S}|L_f^2 L_h^2}{(1-\gamma)^2 T\alpha\varphi^2}, \quad \mathbb{E}[T_5] \leq 16\alpha I, \quad \mathbb{E}[T_6] \leq \frac{200 L_f L_h \sqrt{I|\mathcal{S}|}}{\sqrt{T}(1-\gamma)\varphi}.
\end{aligned} \tag{58}$$

Let $\beta = \frac{(1-\gamma)\varphi}{L_f L_h}\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}|)}{T|\mathcal{S}||\mathcal{A}|}}$ and $\alpha = \frac{L_f L_h \sqrt{|\mathcal{S}|}}{(1-\gamma)\varphi\sqrt{TI}}$, the final bound for duality gap could be written as

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\boldsymbol{\lambda}_\kappa^*, \bar{\mathbf{u}}, \bar{\mathbf{v}}) - \mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{u}^\dagger, \mathbf{v}^\dagger)] &\leq \frac{L_f L_h \sqrt{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}}{\sqrt{T}(1-\gamma)\varphi} + \frac{4000 L_f L_h \sqrt{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}}{\sqrt{T}(1-\gamma)\varphi} \\
&\quad + \frac{400\sqrt{|\mathcal{S}|I}}{\sqrt{T}(1-\gamma)\varphi} + \frac{16 L_f L_h \sqrt{|\mathcal{S}|I}}{\sqrt{T}(1-\gamma)\varphi} + \frac{200 L_f L_h \sqrt{|\mathcal{S}|I}}{\sqrt{T}(1-\gamma)\varphi} \\
&\leq \mathcal{O}\left(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{(1-\gamma)\varphi}\right),
\end{aligned} \tag{59}$$

which is as stated in the statement of Theorem 1. □

C.4 Proof of Lemma 6

The Proof of Lemma 6 in this work follows similar logic to (Zhang et al., 2021)[Lemma C.2]. The main difference lies in the selection of shift parameters M and we provide the proof here for completeness.

Proof. Let us defined Δ_{sa} as the (s,a)-th component of $\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)$. Consider the update in Eq. (24) and note that the problem is separable for each component of $\boldsymbol{\lambda}$ and could be solved in closed form as follows.

$$\begin{aligned}
&\max_{\boldsymbol{\lambda}} \left\langle \hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t \right\rangle - \frac{1}{\beta} KL(\boldsymbol{\lambda}\|\boldsymbol{\lambda}^t) \\
&\quad = \left\langle \hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t), -\boldsymbol{\lambda}^t \right\rangle + \max_{\boldsymbol{\lambda}} \left\{ \sum_{s,a} \Delta_{sa}^t \lambda_{sa} - \frac{1}{\beta}\sum_{s,a} \lambda_{sa} \log\left(\frac{\lambda_{sa}}{\lambda_{sa}^t}\right) \right\} \\
&\quad = \quad \sum_{s,a} \max_{\lambda_{sa}} \left\{ \lambda_{sa} \left[\Delta_{sa}^t - \frac{1}{\beta}\log\left(\frac{\lambda_{sa}}{\lambda_{sa}^t}\right) \right] \right\},
\end{aligned} \tag{60}$$

where we drop the terms which does not depend upon the variable $\boldsymbol{\lambda}$ and Λ denotes the set of probability distributions. Next, we solve the unconstrained maximization in (60) by differentiating and equating it to zero as follows

$$\frac{d}{d\lambda_{sa}}\left(\lambda_{sa}\left[\Delta_{sa}^t - \frac{1}{\beta}\log\left(\frac{\lambda_{sa}}{\lambda_{sa}^t}\right) \right] \right)\Bigg|_{\lambda_{sa}=\lambda_{sa}^{t+\frac{1}{2}}} = \Delta_{sa}^t - \frac{1}{\beta}\log\left(\frac{\lambda_{sa}^{t+\frac{1}{2}}}{\lambda_{sa}^t}\right) - \frac{1}{\beta} = 0. \tag{61}$$

After rearranging the terms, we obtain

$$\lambda_{sa}^{t+\frac{1}{2}} = \lambda_{sa}^{t} \exp(\beta \Delta_{sa}^{t} - 1). \tag{62}$$

Now, we project back the solution on to the set of valid probability distribution and obtain the update as

$$\lambda_{sa}^{t+1} = \frac{\lambda_{sa}^{t} \cdot \exp(\beta \Delta_{sa}^{t})}{\sum_{s',a'} \lambda_{s'a'}^{t} \cdot \exp(\beta \Delta_{s'a'}^{t})}, \tag{63}$$

where we note that $\lambda_{sa}^{t+1} \in \Lambda$. Next, we analyze the one step KL divergence of $\boldsymbol{\lambda}^{t+1}$ to any $\boldsymbol{\lambda}$ as

$$\begin{aligned} KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^{t}) - KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^{t+1}) &= \sum_{s,a} \lambda_{sa} \log\left(\frac{\lambda_{sa}}{\lambda_{sa}^{t}}\right) - \sum_{s,a} \lambda_{sa} \log\left(\frac{\lambda_{sa}}{\lambda_{sa}^{t+1}}\right) \\ &= \sum_{s,a} \lambda_{sa} \log\left(\frac{\lambda_{sa}^{t+1}}{\lambda_{sa}^{t}}\right). \end{aligned} \tag{64}$$

Next, we substitute the definition of λ_{sa}^{t+1} to obtain

$$\begin{aligned} KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^{t}) - KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^{t+1}) &= \sum_{s,a} \lambda_{sa} \left[\beta \Delta_{sa}^{t} - \log\left(\sum_{s',a'} \lambda_{s'a'}^{t} \cdot \exp(\beta \Delta_{s'a'}^{t})\right)\right] \\ &= \beta \left\langle \boldsymbol{\lambda}, \hat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^{t}, \mathbf{u}^{t}, \mathbf{v}^{t}) \right\rangle - \log\left(\sum_{s',a'} \lambda_{s'a'}^{t} \cdot \exp(\beta \Delta_{s'a'}^{t})\right), \end{aligned} \tag{65}$$

where we utilize the fact that $\sum_{s,a} \lambda_{sa} = 1$. To proceed next, recall that we have

$$\Delta_{sa} = \frac{\gamma v_{s'} - v_s - M_1}{\zeta_{sa}} + \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda})_{s,a} + \sum_{i \in [I]} u_i \nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda})_{s,a} - M_2 \tag{66}$$

where $\nabla_f(\boldsymbol{\lambda})(s,a)$ and $\nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda})(s,a)$ are the (s,a) element of $\nabla_f(\boldsymbol{\lambda})$ and $\nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda})$, respectively. We note that

$$|\gamma v_{s'} - v_s| \leq |\gamma v_{s'}| + |v_s| \leq 4\left[\frac{L_f}{1-\gamma} + \frac{4 L_f L_h}{(1-\gamma)\varphi}\right]. \tag{67}$$

Moreover, by Lemma 1

$$|\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda})_{s,a}| \leq L_f, \quad \text{and} \quad \left|\sum_{i \in [I]} u_i \nabla_{\boldsymbol{\lambda}} h^i(\boldsymbol{\lambda})_{s,a}\right| \leq \frac{8 L_f L_h}{\varphi}. \tag{68}$$

Hence, with the selection $M_1 = 4\left[\frac{L_f}{1-\gamma} + \frac{4 L_f L_h}{(1-\gamma)\varphi}\right]$ and $M_2 = L_f + \frac{8 L_f L_h}{\varphi}$, we can conclude that $\Delta_{sa} \leq 0$. Since $\exp(x) \leq (1 + x + \frac{1}{2}x^2)$ for $x \leq 0$, we can upper bound the second term

on the right hand side of (65) as

$$\begin{aligned}\log\left(\sum_{s',a'}\lambda^t_{s'a'}\cdot\exp(\beta\Delta^t_{s'a'})\right) &\le \log\left(\sum_{s',a'}\lambda^t_{s'a'}\cdot(1+\beta\Delta^t_{s'a'}+\frac{1}{2}\beta^2(\Delta^t_{s'a'})^2)\right)\\ &= \log\left(1+\beta\sum_{s',a'}\lambda^t_{s'a'}\Delta^t_{s'a'}+\frac{\beta^2}{2}\sum_{s',a'}\lambda^t_{s'a'}(\Delta^t_{s'a'})^2\right)\\ &= \log\left(1+\beta\left\langle\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t),\boldsymbol{\lambda}^t\right\rangle+\frac{\beta^2}{2}\sum_{s',a'}\lambda^t_{s'a'}(\Delta^t_{s'a'})^2\right)\\ &\le \beta\left\langle\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t),\boldsymbol{\lambda}^t\right\rangle+\frac{\beta^2}{2}\sum_{s',a'}\lambda^t_{s'a'}(\Delta^t_{s'a'})^2,\end{aligned} \tag{69}$$

where the last inequality holds by $\log(1+x)\le x$ for all $x>-1$. By combining Eq. (65) and (69), we obtain

$$\begin{aligned}KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^t)-KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^{t+1}) \ge&\beta\left\langle\boldsymbol{\lambda},\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t)\right\rangle-\beta\left\langle\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t),\boldsymbol{\lambda}^t\right\rangle\\ &-\frac{\beta^2}{2}\sum_{s',a'}\lambda^t_{s'a'}(\Delta^t_{s'a'})^2.\end{aligned} \tag{70}$$

Rearrange the items and divide both sides by β, to obtain

$$0\le\frac{1}{\beta}[KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^t)-KL(\boldsymbol{\lambda}||\boldsymbol{\lambda}^{t+1})]+\left\langle\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t),\boldsymbol{\lambda}^t-\boldsymbol{\lambda}\right\rangle+\frac{\beta}{2}\sum_{s',a'}\lambda^t_{s'a'}(\Delta^t_{s'a'})^2. \tag{71}$$

Add $\left\langle\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t),\boldsymbol{\lambda}-\boldsymbol{\lambda}^t\right\rangle$ on both side to get the desired result. □

C.5 Proof of Lemma 7

Proof. We can combine the update rule in Eq. (22)-(23) to obtain an update for $\mathbf{w}\in\mathcal{W}:=\mathcal{U}\times\mathcal{V}$. For any $\mathbf{w}\in\mathcal{W}$, it holds that

$$\begin{aligned}\|\mathbf{w}^{t+1}-\mathbf{w}\|^2 &= \|\Pi_{\mathcal{W}}(\mathbf{w}^t-\alpha\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t))-\mathbf{w}\|^2\\ &\le \|\mathbf{w}^t-\alpha\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t)-\mathbf{w}\|^2\\ &= \|\mathbf{w}^t-\mathbf{w}\|^2+\alpha^2\|\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t)\|^2-2\alpha\left\langle\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{w}),\mathbf{w}^t-\mathbf{w}\right\rangle\\ &= \|\mathbf{w}^t-\mathbf{w}\|^2+\alpha^2\|\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t)\|^2\\ &\quad-2\alpha\left\langle\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{w})-\nabla_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{w})+\nabla_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{w}),\mathbf{w}^t-\mathbf{w}\right\rangle,\end{aligned}$$

where the first inequality holds by the non-expansiveness of the Projection operator. The following equalities holds by expanding the squares and by adding subtracting the term $2\alpha\left\langle\nabla_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{w}),\mathbf{w}^t-\mathbf{w}\right\rangle$. After rearranging the terms in the above expression, we obtain

$$\begin{aligned}2\alpha\left\langle\nabla_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda},\mathbf{w}),\mathbf{w}^t-\mathbf{w}\right\rangle \le&\|\mathbf{w}^t-\mathbf{w}\|^2-\|\mathbf{w}^{t+1}-\mathbf{w}\|^2+\alpha^2\|\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t)\|^2\\ &-2\alpha\left\langle\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t)-\nabla_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t),\mathbf{w}^t-\mathbf{w}\right\rangle.\end{aligned} \tag{72}$$

Next, divide the both sides by $2\alpha>0$ to obtain the statement of Lemma 7. □

C.6 Upper Bound for $\mathbb{E}[T_1]$

$$\begin{aligned}\mathbb{E}[T_1] &= \frac{KL(\boldsymbol{\lambda}^*_\kappa||\boldsymbol{\lambda}^1)}{T\beta} = \frac{1}{T\beta}\sum_{s,a}\lambda^*_{\kappa,sa}\log\left(\frac{\lambda^*_{\kappa,sa}}{\lambda^1_{sa}}\right) = \frac{1}{T\beta}\sum_{s,a}\lambda^*_{\kappa,sa}[\log\lambda^*_{\kappa,sa} - \log\lambda^1_{sa}] \\ &\leq \frac{1}{T\beta}\sum_{s,a}\lambda^*_{\kappa,sa}\log(|\mathcal{S}||\mathcal{A}|) = \frac{\log(|\mathcal{S}||\mathcal{A}|)}{T\beta}.\end{aligned} \tag{73}$$

C.7 Upper Bound for $\mathbb{E}[T_2]$

For any fixed $\mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t$, we have

$$\begin{aligned}&\mathbb{E}[\sum_{s,a}\lambda^t_{sa}(\Delta^t_{sa})^2|\mathbf{u}^t,\mathbf{v}^t,\boldsymbol{\lambda}^t] \\ &= \mathbb{E}_{s_t,a_t}\left[\sum_{s,a}\lambda^t_{sa}\left(\frac{\gamma v_{s'} - v_s - M_1}{\zeta_{sa}}\cdot\mathbf{1}_{(s,a)=(s_t,a_t)} + \nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})(s,a) + \sum_{i\in[I]}u_i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})(s,a) - M_2\right)^2\right] \\ &\overset{(a)}{\leq} \mathbb{E}_{s_t,a_t}\left\{\sum_{s,a}\lambda^t_{sa}\left[2\left(\frac{\gamma v_{s'} - v_s - M_1}{\zeta_{sa}}\cdot\mathbf{1}_{(s,a)=(s_t,a_t)}\right)^2 + 2\left(\nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})(s,a) + \sum_{i\in[I]}u_i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})(s,a) - M_2\right)^2\right]\right\},\end{aligned} \tag{74}$$

where in step (a), we use the inequality $(a+b)^2 \leq 2a^2 + 2b^2$. Next, we perform further simplifications as

$$\begin{aligned}&\mathbb{E}[\sum_{s,a}\lambda^t_{sa}(\Delta^t_{sa})^2|\mathbf{u}^t,\mathbf{v}^t,\boldsymbol{\lambda}^t] \\ &\leq 2\mathbb{E}_{s_t,a_t}\left[\lambda^t_{s_ta_t}\left(\frac{\gamma v_{s'_t} - v_{s_t} - M_1}{\zeta_{s_ta_t}}\right)^2\right] + 2\sum_{s,a}\lambda^t_{sa}\left(\nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})(s,a) + \sum_{i\in[I]}u_i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})(s,a) - M_2\right)^2 \\ &= 2\sum_{s_t,a_t}\left[\lambda^t_{s_ta_t}\zeta^t_{s_ta_t}\left(\frac{\gamma v_{s'_t} - v_{s_t} - M_1}{\zeta^t_{s_ta_t}}\right)^2\right] + 2\sum_{s,a}\lambda^t_{sa}\left(\nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})(s,a) + \sum_{i\in[I]}u_i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})(s,a) - M_2\right)^2 \\ &= 2\sum_{s_t,a_t}\frac{\lambda^t_{s_ta_t}\left(\gamma v_{s'_t} - v_{s_t} - M_1\right)^2}{(1-\delta)\lambda^t_{s_ta_t} + \frac{\delta}{|\mathcal{S}||\mathcal{A}|}} + 2\sum_{s,a}\lambda^t_{sa}\left(\nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})(s,a) + \sum_{i\in[I]}u_i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})(s,a) - M_2\right)^2\end{aligned} \tag{75}$$

Next, after omitting the positive term in the denominator, we get

$$
\begin{aligned}
&\mathbb{E}[\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2|\mathbf{u}^t,\mathbf{v}^t,\boldsymbol{\lambda}^t]\\
&\leq 2\sum_{s_t,a_t}\frac{\lambda_{s_ta_t}^t\Big(\gamma v_{s_t'}-v_{s_t}-M_1\Big)^2}{(1-\delta)\lambda_{s_ta_t}^t}+2\sum_{s,a}\lambda_{sa}^t\Big(\nabla_{\boldsymbol{\lambda}}f(\boldsymbol{\lambda})(s,a)+\sum_{i\in[I]}u_i\nabla_{\boldsymbol{\lambda}}h^i(\boldsymbol{\lambda})(s,a)-M_2\Big)^2\\
&\overset{(c)}{\leq} 2\sum_{s_t,a_t}\frac{4M_1^2}{1-\delta}+2\sum_{s,a}4\lambda_{sa}^tM_2^2=8\Big(\frac{|\mathcal{S}||\mathcal{A}|M_1^2}{1-\delta}+M_2^2\Big)\\
&=\frac{128|\mathcal{S}||\mathcal{A}|[\frac{L_f}{1-\gamma}+\frac{4L_fL_h}{(1-\gamma)\varphi}]^2}{1-\delta}+8\Big[L_f+\frac{8L_fL_h}{\varphi}\Big]^2\\
&\overset{(d)}{\leq}\frac{128L_f^2|\mathcal{S}||\mathcal{A}|(1+4L_h)^2}{(1-\delta)(1-\gamma)^2\varphi^2}+\frac{8L_f^2(1+8L_h)^2}{(1-\delta)(1-\gamma)^2\varphi^2}\\
&\leq\frac{4000L_f^2L_h^2|\mathcal{S}||\mathcal{A}|}{(1-\delta)(1-\gamma)^2\varphi^2}.
\end{aligned}
\tag{76}
$$

Step (c) holds because we use the boundness of dual variable and Lemma 1. Step (d) holds since $0<\varphi<1$. Next, we write down the term $\mathbb{E}[T_2]$ as

$$
\begin{aligned}
\mathbb{E}[T_2]=\mathbb{E}[\frac{\beta}{2T}\sum_{t=1}^T\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2]&\overset{(a)}{=}\frac{\beta}{2T}\sum_{t=1}^T\mathbb{E}[\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2]\\
&\overset{(b)}{=}\frac{\beta}{2T}\sum_{t=1}^T\mathbb{E}[\mathbb{E}[\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2|\mathbf{u}^t,\mathbf{v}^t,\boldsymbol{\lambda}^t]]\\
&\leq\frac{4000\beta L_f^2L_h^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2\varphi^2},
\end{aligned}
\tag{77}
$$

where step (a) holds by the linear of expectation and step (b) holds due to law of total expectation. The last inequality holds by $\delta\in(0,\frac{1}{2})$.

C.8 Expression for $\mathbb{E}[T_3]$

For any fixed $\mathbf{u}^t,\mathbf{v}^t,\boldsymbol{\lambda}^t$, we have

$$
\mathbb{E}[\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t)|\mathbf{u}^t,\mathbf{v}^t,\boldsymbol{\lambda}^t]=\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{u}^t,\mathbf{v}^t)-M_1\cdot\mathbf{1}-M_2\cdot\mathbf{1}. \tag{78}
$$

Thus,

$$
\mathbb{E}[T_3]=\frac{1}{T}\sum_{t=1}^T\mathbb{E}[\Big\langle\hat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t)-\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}^t,\mathbf{w}^t),\boldsymbol{\lambda}^t-\boldsymbol{\lambda}\Big\rangle]=\frac{1}{T}\sum_{t=1}^T\mathbb{E}[\langle-(M_1+M_2)\cdot\mathbf{1},\boldsymbol{\lambda}^t-\boldsymbol{\lambda}\rangle]=0 \tag{79}
$$

where the last step is true because $\langle\boldsymbol{\lambda}^t\cdot\mathbf{1}\rangle=\langle\boldsymbol{\lambda}^*\cdot\mathbf{1}\rangle=1$

C.9 Upper Bound for $\mathbb{E}[T_4]$

For any $\mathbf{u} \in \mathcal{U}$

$$\|\mathbf{u}^1 - \mathbf{u}\|^2 \leq \|\mathbf{u}^1\|^2 + \|\mathbf{u}\|^2 + 2|\left\langle \mathbf{u}^1, \mathbf{u} \right\rangle| \leq \|\mathbf{u}^1\|^2 + \|\mathbf{u}\|^2 + 2\|\mathbf{u}^1\|\|\mathbf{u}\| \leq \frac{256L_f^2}{\varphi^2} \tag{80}$$

where the last inequality holds by $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$ for any $\mathbf{x}$ and the definition of $\mathcal{U}$. Similarly, for any $\mathbf{v} \in \mathcal{V}$

$$\begin{aligned}\|\mathbf{v}^1 - \mathbf{v}\|^2 &\leq \|\mathbf{v}^1\|^2 + \|\mathbf{v}\|^2 + 2\|\mathbf{v}^1\|\|\mathbf{v}^t\| \leq |\mathcal{S}|(\|\mathbf{v}^1\|_\infty^2 + \|\mathbf{v}\|_\infty^2 + 2\|\mathbf{v}^1\|_\infty\|\mathbf{v}^t\|_\infty) \\ &\leq 16|\mathcal{S}|[\frac{L_f}{1-\gamma} + \frac{4L_fL_h}{(1-\gamma)\varphi}]^2 \leq \frac{400|\mathcal{S}|L_f^2L_h^2}{(1-\gamma)^2\varphi^2}\end{aligned} \tag{81}$$

Finally, combine above two inequalities,

$$\mathbb{E}[T_4] = \frac{1}{2T\alpha}\|\mathbf{w}^1 - \mathbf{w}^\dagger\|^2 = \frac{1}{2T\alpha}[\|\mathbf{u}^1 - \mathbf{u}^\dagger\|^2 + \|\mathbf{v}^1 - \mathbf{v}^\dagger\|^2] \leq \frac{400|\mathcal{S}|L_f^2L_h^2}{(1-\gamma)^2T\alpha\varphi^2} \tag{82}$$

C.10 Upper Bound for $\mathbb{E}[T_5]$

For any fixed $\mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t$, we have

$$\mathbb{E}\left[\|\hat{\nabla}_{\mathbf{u}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)\|^2 \middle| \mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t\right] = \|\mathbf{h}(\boldsymbol{\lambda}^t) - \kappa\mathbf{1}\|^2 \leq 2\|\mathbf{h}(\boldsymbol{\lambda}^t)\|^2 + 2\kappa^2 I \leq 4I \tag{83}$$

where the last step holds because $|h^i(\boldsymbol{\lambda})| \leq 1, \forall i \in [I]$ by the Lemma 2 and the fact $0 < \kappa \leq 1$.

$$\begin{aligned}\mathbb{E}\left[\|\hat{\nabla}_{\mathbf{v}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)\|^2 \middle| \mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t\right] &= \mathbb{E}_{s_t,a_t,s_t',s_0}\left[\|(1-\gamma)\mathbf{e}_{s_0} + \frac{\lambda_{s_ta_t}(\gamma\mathbf{e}_{s_t'} - \mathbf{e}_{s_t})}{\zeta_{s_ta_t}}\|^2 \middle| \mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t\right] \\ &\overset{(a)}{=} \mathbb{E}_{s_t,a_t,s_t',s_0}\left[\|(1-\gamma)\mathbf{e}_{s_0} + \frac{\lambda_{s_ta_t}(\gamma\mathbf{e}_{s_t'} - \mathbf{e}_{s_t})}{(1-\delta)\lambda_{s_ta_t}^t + \frac{\delta}{|\mathcal{S}||\mathcal{A}|}}\|^2 \middle| \mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t\right] \\ &\overset{(b)}{\leq} \mathbb{E}_{s_t,a_t,s_t',s_0}\left[3\|(1-\gamma)\mathbf{e}_{s_0}\|^2 + 3\|\frac{\lambda_{s_ta_t}(\gamma\mathbf{e}_{s_t'})}{(1-\delta)\lambda_{s_ta_t}^t}\|^2 + 3\|\frac{\lambda_{s_ta_t}(\mathbf{e}_{s_t})}{(1-\delta)\lambda_{s_ta_t}^t}\|^2 \middle| \mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t\right] \\ &\leq \mathbb{E}_{s_t,a_t,s_t',s_0}\left[3(1-\gamma)^2 + \frac{3\gamma^2+3}{(1-\delta)^2}\right] \leq 3 + \frac{6}{(1-\delta)^2}\end{aligned} \tag{84}$$

where step (a) holds by using the definition of $\zeta_{s_ta_t}$ in the algorithm. Step (b) comes from the Cauchy-Schwartz inequality. Combined Eq. (83), (84) with the definition of $\mathbf{w}$,

$$\begin{aligned}\mathbb{E}[T_5] &= \frac{\alpha}{2T}\sum_{t=1}^{T}\mathbb{E}\|\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t)\|^2 = \frac{\alpha}{2T}\sum_{t=1}^{T}\left[\mathbb{E}\|\hat{\nabla}_{\mathbf{u}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)\|^2 + \mathbb{E}\|\hat{\nabla}_{\mathbf{u}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)\|^2\right] \\ &\leq \frac{\alpha}{2}\left[3 + \frac{6}{(1-\delta)^2} + 4I\right] \leq 16\alpha I\end{aligned} \tag{85}$$

where the last step holds by $\delta \in (0, \frac{1}{2})$

C.11 Upper Bound for $\mathbb{E}[T_6]$

Firstly, notice that T_6 is different from T_3 because $\mathbf{w}^\dagger$ depends on $\bar{\boldsymbol{\lambda}}$, which is a random variable. However $\boldsymbol{\lambda}_\kappa^*$ depends only on κ, which is a constant. Thus, in order to bound T_6, we need following Lemma.

Lemma 8 ((Beck, 2017)). *Let $\mathcal{Z} \subset \mathbb{R}^d$ be a convex set and $\omega : \mathcal{Z} \to \mathbb{R}$ be a 1-strongly convex function with respect to norm $\|\cdot\|$ over $\mathcal{Z}$. With the assumption that for all $x \in \mathcal{Z}$ we have $\omega(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{Z}} \omega(\mathbf{x}) \le \frac{1}{2}D^2$, then for any martingale difference sequence $\{\mathbf{Z}_k\}_{k=1}^K \in \mathbb{R}^d$ and any random vector $\mathbf{x} \in \mathcal{Z}$, it holds that*

$$\mathbb{E}\Big[\sum_{k=1}^K \langle \mathbf{Z}_k, \mathbf{x}\rangle\Big] \le \frac{D}{2}\sqrt{\sum_{k=1}^K \mathbb{E}[\|\mathbf{Z}_k\|_*^2]} \tag{86}$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$

For any fixed $\mathbf{u}^t, \mathbf{v}^t, \boldsymbol{\lambda}^t$, the gradient estimation is unbiased.

$$\mathbb{E}[\hat{\nabla}_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t)] = \nabla_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{u}^t, \mathbf{v}^t) \tag{87}$$

where $\boldsymbol{\phi} = \mathbf{u}$ or $\mathbf{v}$. Thus,

$$\begin{aligned}\mathbb{E}[T_6] &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\Big[\Big\langle \nabla_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t) - \hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\dagger\Big\rangle\Big] \\ &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\Big[\Big\langle \hat{\nabla}_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t) - \nabla_{\mathbf{w}}\mathcal{L}(\boldsymbol{\lambda}^t, \mathbf{w}^t), \mathbf{w}^\dagger\Big\rangle\Big].\end{aligned} \tag{88}$$

To apply Lemma 8, let $\mathcal{Z} = \mathcal{W}$, $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, $\mathbf{x} = \mathbf{w}^\dagger$ and $\mathbf{Z}_k = \hat{\nabla}_{\mathbf{w}}\mathcal{L}(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}^k, \boldsymbol{\lambda}^k)$, which is a martingale difference. Then, $\omega(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{Z}}\omega(\mathbf{x}) = \omega(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \le \frac{1}{2}D^2$ and thus $D \ge \|\mathbf{w}\|$. The norm of $\mathbf{w}$ can be bounded as

$$\begin{aligned}\|\mathbf{w}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 &\le \|\mathbf{u}\|_1^2 + |\mathcal{S}|\|\mathbf{v}\|_\infty^2 = (\frac{8L_f}{\varphi})^2 + 2|\mathcal{S}|\left[\frac{L_f}{1-\gamma} + \frac{4L_fL_h}{(1-\gamma)\varphi}\right]^2 \\ &\le \frac{256L_f^2}{\varphi^2} + \frac{2|\mathcal{S}|L_f^2}{(1-\gamma)^2} + \frac{16|\mathcal{S}|L_f^2L_h}{(1-\gamma)^2\varphi} + \frac{32|\mathcal{S}|L_f^2L_h^2}{(1-\gamma)^2\varphi^2} \le \frac{324|\mathcal{S}|L_f^2L_h^2}{(1-\gamma)^2\varphi^2}.\end{aligned} \tag{89}$$

Thus, $\|\mathbf{w}\| \le \frac{18L_fL_h\sqrt{|\mathcal{S}|}}{(1-\gamma)\varphi} =: D$. Apply Lemma 8 to Eq. (88),

$$\begin{aligned}\mathbb{E}[T_6] &\le \frac{18L_fL_h\sqrt{|\mathcal{S}|}}{T(1-\gamma)\varphi}\sqrt{\sum_{t=1}^T \mathbb{E}[\|\hat{\nabla}_{\mathbf{w}}\hat{\mathcal{L}}(\mathbf{w}^t, \boldsymbol{\lambda}^t) - \nabla_{\mathbf{w}}\hat{\mathcal{L}}(\mathbf{w}^t, \boldsymbol{\lambda}^t)\|^2]} \\ &\le \frac{18L_fL_h\sqrt{|\mathcal{S}|}}{T(1-\gamma)\varphi}\sqrt{\sum_{t=1}^T \mathbb{E}[\|\hat{\nabla}_{\mathbf{w}}\hat{\mathcal{L}}(\mathbf{w}^t, \boldsymbol{\lambda}^t)\|^2]} \\ &= \frac{18L_fL_h\sqrt{|\mathcal{S}|}}{T(1-\gamma)\varphi}\sqrt{\frac{2T}{\alpha}\mathbb{E}[T_5']} \\ &\le \frac{18L_fL_h\sqrt{|\mathcal{S}|}}{\sqrt{T}(1-\gamma)\varphi}\sqrt{32I} = \frac{200L_fL_h\sqrt{I|\mathcal{S}|}}{\sqrt{T}(1-\gamma)\varphi}.\end{aligned} \tag{90}$$

Appendix D. Proofs for Section 5.2

D.1 Proof of Lemma 5

Proof. Recall $\boldsymbol{\lambda}^*$ is the optimal occupancy measure to the original problem, which gives

$$h^i(\boldsymbol{\lambda}^*) \geq 0 \tag{91}$$

Further, under the Slater Condition Assumption 3, there exists at least one occupancy measure $\tilde{\boldsymbol{\lambda}}$ such that

$$h^i(\tilde{\boldsymbol{\lambda}}) \geq \varphi \tag{92}$$

Define a new occupancy measure $\hat{\boldsymbol{\lambda}} = (1-\frac{\kappa}{\varphi})\boldsymbol{\lambda}^* + \frac{\kappa}{\varphi}\tilde{\boldsymbol{\lambda}}$. By the concavity of the cost function, it can be shown a feasible occupancy measure to the conservative problem.

$$h^i(\hat{\boldsymbol{\lambda}}) = h^i\left((1-\frac{\kappa}{\varphi})\boldsymbol{\lambda}^* + \frac{\kappa}{\varphi}\tilde{\boldsymbol{\lambda}}\right) \geq (1-\frac{\kappa}{\varphi})h^i\left(\boldsymbol{\lambda}^*\right) + \frac{\kappa}{\varphi}h^i\left(\tilde{\boldsymbol{\lambda}}\right) \geq \frac{\kappa}{\varphi}\varphi = \kappa \tag{93}$$

$$\sum_a (\mathbf{I} - \gamma \mathbf{P}_a^T)\hat{\boldsymbol{\lambda}}_a = (1-\frac{\kappa}{\varphi})\sum_a (\mathbf{I} - \gamma \mathbf{P}_a^T)\boldsymbol{\lambda}_a^* + \frac{\kappa}{\varphi}\sum_a (\mathbf{I} - \gamma \mathbf{P}_a^T)\tilde{\boldsymbol{\lambda}}_a = (1-\gamma)\tilde{\boldsymbol{\rho}} \tag{94}$$

Then, we can bound the difference

$$\begin{aligned} f(\boldsymbol{\lambda}^*) - f(\boldsymbol{\lambda}^*_{\boldsymbol{\kappa}}) &\overset{(a)}{\leq} f(\boldsymbol{\lambda}^*) - f(\hat{\boldsymbol{\lambda}}) = f(\boldsymbol{\lambda}^*) - f\left((1-\frac{\kappa}{\varphi})\boldsymbol{\lambda}^* + \frac{\kappa}{\varphi}\tilde{\boldsymbol{\lambda}}\right) \\ &\leq \frac{\kappa}{\varphi}f(\boldsymbol{\lambda}^*) - \frac{\kappa}{\varphi}f(\tilde{\boldsymbol{\lambda}}) \overset{(b)}{\leq} \frac{\kappa}{\varphi}f(\boldsymbol{\lambda}^*) \overset{(c)}{\leq} \frac{\kappa}{\varphi} \end{aligned} \tag{95}$$

The first step (a) holds because $\boldsymbol{\lambda}^*_{\boldsymbol{\kappa}}$ is the optimal solution of the conservative problem, which gives larger value function than any other feasible occupancy measure. We drop the negative term in the step (b) and the last step (c) is true because $f(\boldsymbol{\lambda}^*) \leq 1$ by the Lemma 2. □

D.2 Proof of Theorem 2

Proof. In order to construct the relation between duality gap and result in occupancy measure space, let us consider the expression for the Lagrangian function. By the feasibility of $\boldsymbol{\lambda}^*_{\boldsymbol{\kappa}}$, we can write

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}^*_{\boldsymbol{\kappa}}, \mathbf{u}^t, \mathbf{v}^t) =& f(\boldsymbol{\lambda}^*_{\boldsymbol{\kappa}}) + \left\langle \mathbf{u}^t, \mathbf{h}(\boldsymbol{\lambda}^*_{\boldsymbol{\kappa}}) - \boldsymbol{\kappa} \right\rangle + \left[\sum_a (\boldsymbol{\lambda}^*_{\boldsymbol{\kappa},a})^T(\gamma \mathbf{P}_a - \mathbf{I}) - (1-\gamma)\boldsymbol{\rho}\right]\mathbf{v}^t \\ \geq& f(\boldsymbol{\lambda}^*_{\boldsymbol{\kappa}}). \end{aligned} \tag{96}$$

Define the set $\mathcal{I} = \{i | h^i(\bar{\boldsymbol{\lambda}}) < 0\}$. Denote $\mathbf{u}' = [u'_1, u'_2, \cdots, u'_I]^T$, where $u'_i = u_i$ if $i \in \mathcal{I}$ and $u'_i = 0$ otherwise. Define $C_1 := \frac{4L_f}{\varphi}$ and $C_2 = \frac{L_f}{1-\gamma} + \frac{4L_f L_h}{(1-\gamma)\varphi}$ for simplicity, which is the

bound for $\|\mathbf{u}_\kappa^*\|_1$ and $\|\mathbf{v}_\kappa^*\|_\infty$, respectively. By the definition of $\mathbf{u}^\dagger, \mathbf{v}^\dagger$

$$\begin{aligned}\mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{u}^\dagger, \mathbf{v}^\dagger) &= \min_{\mathbf{u}\in\mathcal{U},\mathbf{v}\in\mathcal{V}} f(\bar{\boldsymbol{\lambda}}) + \left\langle \mathbf{u}, \mathbf{h}(\bar{\boldsymbol{\lambda}}) - \boldsymbol{\kappa}\right\rangle + \left[\sum_a (\bar{\boldsymbol{\lambda}}_a)^T(\gamma\mathbf{P}_a - \mathbf{I}) - (1-\gamma)\boldsymbol{\rho}\right]\mathbf{v} \\ &\overset{(a)}{=} \min_{\mathbf{u}'\in\mathcal{U},\mathbf{v}\in\mathcal{V}} f(\bar{\boldsymbol{\lambda}}) + \left\langle \mathbf{u}', [\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \boldsymbol{\kappa}]_-\right\rangle + \left[\sum_a (\bar{\boldsymbol{\lambda}}_a)^T(\gamma\mathbf{P}_a - \mathbf{I}) - (1-\gamma)\boldsymbol{\rho}\right]\mathbf{v},\end{aligned} \tag{97}$$

where the notation $x_- := \min\{x, 0\}$ and the equality holds because $u_i = 0, i \in \mathcal{I}^c$ for those constraints which are satisfied. Let us consider the second term on the right hand side of the above expression as follows

$$\begin{aligned}\left\langle \mathbf{u}', [\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \boldsymbol{\kappa}]_-\right\rangle \leq& \|\mathbf{u}'\|_1 \|[\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \kappa\mathbf{1}]_-\|_\infty \\ \leq& 2C_1\|[\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \kappa\mathbf{1}]_-\|_\infty.\end{aligned} \tag{98}$$

Notice that equality in the above inequality is achievable by selecting $u_j^\dagger = 2C_1$ for $j = \operatorname{argmax}_i |h^i(\bar{\boldsymbol{\lambda}}) - \kappa|$ and $u_k^\dagger = 0$ for $k \neq j$. Such $\mathbf{u}^\dagger$ gives the minimum of $\left\langle \mathbf{u}', [\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \boldsymbol{\kappa}]_-\right\rangle = 2C_1\|[\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \kappa\mathbf{1}]_-\|_\infty$. Similarly, $\mathbf{v}^\dagger = 2C_2\mathbf{1}$ gives the minimum of $\left[\sum_a (\bar{\boldsymbol{\lambda}}_a)^T(\gamma\mathbf{P}_a - \mathbf{I}) - (1-\gamma)\rho\right]\mathbf{v} = 2C_2\|\sum_a (\bar{\boldsymbol{\lambda}}_a)^T(\gamma\mathbf{P}_a - \mathbf{I}) - (1-\gamma)\rho\|_1$ by Holder inequality. Hence, we could write the expression in (97) as

$$\mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{u}^\dagger, \mathbf{v}^\dagger) = \left\langle \bar{\boldsymbol{\lambda}}, \mathbf{r}\right\rangle - \|[\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \kappa\mathbf{1}]_-\|_\infty - 2C_2\|\sum_a (\bar{\boldsymbol{\lambda}}_a)^T(\gamma\mathbf{P}_a - \mathbf{I}) - (1-\gamma)\boldsymbol{\rho}\|_1. \tag{99}$$

Combining Eq. (99) with (96) and then taking expectation, we obtain

$$\mathbb{E}[\mathcal{L}(\boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*, \mathbf{u}^t, \mathbf{v}^t) - \mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{u}^\dagger, \mathbf{v}^\dagger)] \geq \mathbb{E}\left[f(\boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*) - f(\bar{\boldsymbol{\lambda}}) + \|[\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \kappa\mathbf{1}]_-\|_\infty + 2C_2\|\sum_a (\boldsymbol{\lambda}_a^t)^T(\gamma\mathbf{P}_a - \mathbf{I}) + (1-\gamma)\boldsymbol{\rho}\|_1\right]. \tag{100}$$

Combining with the result in Theorem 1, there exists a constant $\tilde{c}_1$ such that

$$\begin{aligned}&\mathbb{E}\left[f(\boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*) - f(\bar{\boldsymbol{\lambda}}) + \|[\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \kappa\mathbf{1}]_-\|_\infty + 2C_2\|\sum_a (\boldsymbol{\lambda}_a^t)^T(\gamma\mathbf{P}_a - \mathbf{I}) + (1-\gamma)\boldsymbol{\rho}\|_1\right] \\ &\leq \tilde{c}_1\left(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{(1-\gamma)\varphi}\right).\end{aligned} \tag{101}$$

Denote $L := \tilde{c}_1\left(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{(1-\gamma)\varphi}\right)$. By the Theorem 4 (see Appendix F for reference), we directly get

$$\mathbb{E}[f(\boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*) - f(\bar{\boldsymbol{\lambda}})] \leq L, \tag{102a}$$

$$\mathbb{E}\|[\mathbf{h}(\bar{\boldsymbol{\lambda}}) - \kappa\mathbf{1}]_-\|_\infty \leq \frac{2L}{C_1} = \frac{L\varphi}{2L_f} \leq L\varphi, \tag{102b}$$

$$\mathbb{E}\Big\|\sum_a (\gamma\mathbf{P}_a^T - \mathbf{I})\bar{\boldsymbol{\lambda}}_a + (1-\gamma)\boldsymbol{\rho}\Big\|_1 \leq \frac{2L}{C_2} = \frac{2L}{\frac{L_f}{1-\gamma} + \frac{4L_f L_h}{(1-\gamma)\varphi}} \leq \frac{(1-\gamma)L\varphi}{L_f L_h}. \tag{102c}$$

Note that the result in (102a) is at $\boldsymbol{\lambda}_\kappa^*$ and in order to obtain the result for $\boldsymbol{\lambda}^*$, let us consider and by the statement of Lemma 5, we could write

$$\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] = \mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*)] + \mathbb{E}[f(\boldsymbol{\lambda}_{\boldsymbol{\kappa}}^*) - f(\bar{\boldsymbol{\lambda}})] \leq \frac{\kappa}{\varphi} + L, \tag{103}$$

where we have utilized the upper bound developed in Lemma 5. Next, recall that

$$\kappa = 2\tilde{c}_1 \bigg(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{1-\gamma} \bigg),$$

and from the definition of L, we can write

$$\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] \leq 3\tilde{c}_1 \bigg(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{(1-\gamma)\varphi} \bigg), \tag{104}$$

which establishes the upper bound for the optimally gap for the original optimization problem. Further, from the result in (102b), we have for all $i \in [I]$

$$\mathbb{E}|[h^i(\bar{\boldsymbol{\lambda}}) - \kappa]_-| \leq L\varphi. \tag{105}$$

Note that by the definition of $[x]_- := \min\{x, 0\}$, it holds that $|[x]_-| = -\min\{x, 0\}$ which holds due to the fact that $\min\{x, 0\}$ is either zero or negative. Therefore, it holds that $|h^i(\bar{\boldsymbol{\lambda}}) - \kappa| = -[h^i(\bar{\boldsymbol{\lambda}}) - \kappa]_-$ and thus

$$\mathbb{E}\big([h^i(\bar{\boldsymbol{\lambda}}) - \kappa]_-\big) \geq -L\varphi. \tag{106}$$

Further, since $[x]_-$ is a concave function with respect to x, via Jensen's inequality, we can write

$$[\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}}) - \kappa]]_- \geq \mathbb{E}\big([h^i(\bar{\boldsymbol{\lambda}}) - \kappa]_-\big) \geq -L\varphi. \tag{107}$$

Again, by the definition of $[x]_-$, we simplifies (107) to

$$\min\{\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] - \kappa, 0\} \geq -L\varphi. \tag{108}$$

Thus, we obtain either $\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq \kappa > 0$ or $\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq \kappa - L\varphi$. The first case is trivial and for the second case, recall $\kappa = 2\tilde{c}_1 \bigg(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{1-\gamma} \bigg)$

$$\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq \kappa - L\varphi = \tilde{c}_1 \bigg(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{1-\gamma} \bigg) \tag{109}$$

Let $T = \tilde{c}_1^2 \frac{L_f^2 L_h^2 I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2\varphi^2\epsilon^2}$. By Eq. (102), we have the final result

$$\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] \leq 3\epsilon \tag{110a}$$

$$\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq \epsilon\varphi \quad \forall i \in [I] \tag{110b}$$

$$\mathbb{E}\| \sum_a (\gamma \mathbf{P}_a^T - \mathbf{I})\bar{\boldsymbol{\lambda}}_a + (1-\gamma)\boldsymbol{\rho}\|_1 \leq \frac{(1-\gamma)\epsilon\varphi}{L_f} \tag{110c}$$

Recall that it is required $\kappa \leq \min\{\frac{\varphi}{2}, 1\}$, which gives

$$T \geq 4\tilde{c}_1^2 \frac{L_f^2 L_h^2 I|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2\varphi^2} \max\{4, \varphi^2\} \tag{111}$$

□

D.3 Proof of Corollary 1

Proof. Under the condition that $\kappa = 0$, it is obvious that $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^*_\kappa$. Thus, we have

$$\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] \leq L = \tilde{c}_1 \left(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{(1-\gamma)\varphi} \right) \tag{112}$$

Furthermore, similar to Eq. (109)

$$\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq \kappa - L\varphi = -\tilde{c}_1 \left(\sqrt{\frac{I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{L_f L_h}{1-\gamma} \right) \tag{113}$$

Let $T = \tilde{c}_1^2 \frac{L_f^2 L_h^2 I|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^2 \varphi^2 \epsilon^2}$, we derive the following result

$$\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] \leq \epsilon \tag{114a}$$

$$\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq -\epsilon \quad \forall i \in [I] \tag{114b}$$

$$\mathbb{E}\| \sum_a (\gamma \mathbf{P}_a^T - \mathbf{I}) \bar{\boldsymbol{\lambda}}_a + (1-\gamma)\boldsymbol{\rho} \|_1 \leq \frac{(1-\gamma)\epsilon\varphi}{L_f L_h} \tag{114c}$$

□

Appendix E. Proofs for Section 5.3

E.1 Proof of Theorem 3

Proof. By the result in Eq. (30b) and the definition of $\|\cdot\|_1$, we have

$$\mathbb{E}\left[\sum_s \Big| \sum_{a'} \bar{\lambda}_{sa'} - \gamma \sum_{a'} \sum_{s'} P_{a'}(s', s) \bar{\boldsymbol{\lambda}}_{s'a'} - (1-\gamma)\rho_s \Big| \right] \leq \frac{(1-\gamma)\epsilon\varphi.}{L_f L_h} \tag{115}$$

For each $s \in \mathcal{S}$, let us define

$$\Big| \sum_{a'} \bar{\lambda}_{sa'} - \gamma \sum_{a'} \sum_{s'} P_{a'}(s', s) \bar{\boldsymbol{\lambda}}_{s'a'} - (1-\gamma)\rho_s \Big| = (1-\gamma)\epsilon_s. \tag{116}$$

We notice that the left hand side of Eq. (116) gives the physical meaning of occupancy measure, which can be seen in the following Eq. (117)-(121). Furthermore, Notice that ϵ_s is a random variable. It is obvious that $\epsilon_s \geq 0$ and $\mathbb{E}[\sum_s \epsilon_s] \leq \frac{\epsilon\varphi}{L_f l_H}$ by Eq. (115). Then, define the policy induced by $\bar{\lambda}$ as $\bar{\pi}(a|s) = \frac{\bar{\lambda}_{sa}}{\sum_{a'} \bar{\lambda}_{sa'}} \geq 0$. Multiply the both sides of Eq. (116) by $\bar{\pi}(a|s)$ to obtain

$$\Big| \bar{\lambda}_{sa} - \gamma \sum_{a'} \sum_{s'} P_{a'}(s', s) \bar{\pi}(a|s) \bar{\lambda}_{s'a'} - (1-\gamma)\rho_s \bar{\pi}(a|s) \Big| = (1-\gamma)\epsilon_s \bar{\pi}(a|s), \quad \forall a \in \mathcal{A}, s \in \mathcal{S}. \tag{117}$$

Now define $\rho_{sa} = \rho_s \bar{\pi}(a|s)$ which can be considered as the initial distribution for state and action following policy $\bar{\pi}$. Define $P_{\bar{\pi}}(s, a, s', a') = P_a(s, s') \cdot \bar{\pi}(a'|s')$, which can be considered as the transition matrix from current state and action pair (s, a) to next state and action

pair (s', a'). Furthermore, define $\epsilon_{sa} = \epsilon_s \bar{\pi}(a|s)$ and it is obvious that $\sum_a \epsilon_{sa} = \epsilon_s$. Then, Eq. (117) can be simplified as

$$\left|\bar{\lambda}_{sa} - \gamma \sum_{a'} \sum_{s'} P_{\bar{\pi}}(s', a', s, a) \bar{\lambda}_{s'a'} - (1-\gamma)\rho_{sa}\right| = (1-\gamma)\epsilon_{sa}, \quad \forall a \in \mathcal{A}, s \in \mathcal{S}. \tag{118}$$

With a little abuse of notation $\pm$, we can write

$$\bar{\lambda}_{sa} - \gamma \sum_{a'} \sum_{s'} P_{\bar{\pi}}(s', a', s, a) \bar{\lambda}_{s'a'} = (1-\gamma)(\rho_{sa} \pm \epsilon_{sa}), \quad \forall a \in \mathcal{A}, s \in \mathcal{S}, \tag{119}$$

where $\pm$ means the left hand side can be equal to $(1-\gamma)(\rho_{sa} + \epsilon_{sa})$ or $(1-\gamma)(\rho_{sa} - \epsilon_{sa})$. Next, define $\tilde{\boldsymbol{\rho}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} = [\rho_{s_1 a_1}, \rho_{s_1 a_2}, \cdots, \rho_{s_{|\mathcal{S}|} a_1}, \rho_{s_2 a_1}, \cdots, \rho_{s_{|\mathcal{S}|} a_{|\mathcal{A}|}}]^T$ as a vector, define $\tilde{\boldsymbol{\epsilon}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} = [\epsilon_{s_1 a_1}, \epsilon_{s_1 a_2}, \cdots, \epsilon_{s_2 a_1}, \cdots, \epsilon_{s_{|\mathcal{S}|} a_{|\mathcal{A}|}}]^T$ as a vector, and define $\mathbf{P}_{\bar{\pi}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ as a matrix. Then, we could write the expression in Eq. (119) in the following compact form as

$$\bar{\boldsymbol{\lambda}} - \gamma \mathbf{P}_{\bar{\pi}}^T \bar{\boldsymbol{\lambda}} = (1-\gamma)(\tilde{\boldsymbol{\rho}} \pm \tilde{\boldsymbol{\epsilon}}) \tag{120}$$

Notice that $\|\mathbf{P}_{\bar{\pi}}^T\|_1 = \max_j \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} |\mathbf{P}_{\bar{\pi}}^T(i,j)| = 1$ and thus $\|\gamma \mathbf{P}_{\bar{\pi}}^T\| \leq \gamma$. This means $(I - \gamma \mathbf{P}_{\bar{\pi}}^T)$ is invertable and $(I - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1} = \sum_{i=0}^{\infty} \gamma^i (\mathbf{P}_{\bar{\pi}}^T)^i$. Thus, we have

$$\bar{\boldsymbol{\lambda}} = (1-\gamma)(\mathbf{I} - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1}(\tilde{\boldsymbol{\rho}} \pm \tilde{\boldsymbol{\epsilon}}). \tag{121}$$

Rearrange items, take inner-product with $\mathbf{r}$ and take absolute value, we have

$$\bar{\boldsymbol{\lambda}} - (1-\gamma)(\mathbf{I} - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1} \tilde{\boldsymbol{\rho}} = (1-\gamma)(\mathbf{I} - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1} \tilde{\boldsymbol{\epsilon}}. \tag{122}$$

Notice that

$$(1-\gamma)(\mathbf{I} - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1} \tilde{\boldsymbol{\rho}} = (1-\gamma)\left[\tilde{\boldsymbol{\rho}}^T + \gamma \tilde{\boldsymbol{\rho}}^T \mathbf{P}_{\bar{\pi}} + \gamma^2 \tilde{\boldsymbol{\rho}}^T (\mathbf{P}_{\bar{\pi}})^2 + \cdots\right] = \boldsymbol{\lambda}^{\bar{\pi}} \tag{123}$$

The above equation can be bounded by

$$\begin{aligned}
\mathbb{E}|f(\bar{\boldsymbol{\lambda}}) - f(\boldsymbol{\lambda}^{\bar{\pi}})| &\overset{(a)}{\leq} L_f \mathbb{E}\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^{\bar{\pi}}\| \\
&= L_f (1-\gamma) \mathbb{E}\|(\mathbf{I} - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1} \tilde{\boldsymbol{\epsilon}}\|_2 \\
&\overset{(b)}{\leq} L_f (1-\gamma) \mathbb{E}\|(\mathbf{I} - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1} \tilde{\boldsymbol{\epsilon}}\|_1 \\
&\overset{(c)}{\leq} L_f (1-\gamma) \|(\mathbf{I} - \gamma \mathbf{P}_{\bar{\pi}}^T)^{-1}\|_1 \mathbb{E}\|\tilde{\boldsymbol{\epsilon}}\|_1 \\
&\overset{(d)}{\leq} \frac{1-\gamma}{L_h} \sum_{i=0}^{\infty} \|\gamma^i (\mathbf{P}_{\bar{\pi}}^T)^i\|_1 \epsilon \varphi \\
&\overset{(e)}{\leq} (1-\gamma) \sum_{i=0}^{\infty} \gamma^i \epsilon \varphi = \epsilon \varphi,
\end{aligned} \tag{124}$$

where step (a) holds by the Lipschitz assumption 2, step (b) holds by norm inequality, step (c) holds by definition of matrix norm, step (d) holds by triangle inequality and $\mathbb{E}\|\tilde{\boldsymbol{\epsilon}}\|_1 = \mathbb{E}[\sum_s \epsilon_s] \leq \frac{\epsilon \varphi}{L_f L_h}$. The last step (e) is true because $\|\mathbf{P}_{\bar{\pi}}^T\|_1 = 1$. Finally, we get the result

$$\mathbb{E}|f(\bar{\boldsymbol{\lambda}}) - f(\boldsymbol{\lambda}^{\bar{\pi}})| \overset{(a)}{\leq} \epsilon \varphi. \tag{125}$$

Recall $\mathbb{E}[f(\boldsymbol{\lambda}^*) - f(\bar{\boldsymbol{\lambda}})] \leq 3\epsilon$ in Eq. (31), hence we can write

$$\begin{aligned} f(\boldsymbol{\lambda}^*) - \mathbb{E}[f(\boldsymbol{\lambda}^{\bar{\pi}})] =& \big(f(\boldsymbol{\lambda}^*) - \mathbb{E}[f(\bar{\boldsymbol{\lambda}})]\big) + \mathbb{E}\big[f(\bar{\boldsymbol{\lambda}}) - f(\boldsymbol{\lambda}^{\bar{\pi}})\big] \\ \leq& 4\epsilon, \end{aligned} \tag{126}$$

which is for the objective suboptimality gap in the primal domain. Rescaling ϵ to $\frac{\epsilon}{4}$ finishes the proof. Similarly, for the constraints in the primal domain, we could write

$$\mathbb{E}[h^i(\boldsymbol{\lambda}^{\bar{\pi}}) - h^i(\bar{\boldsymbol{\lambda}})] \geq -\epsilon\varphi. \tag{127}$$

From the result in Eq. (30a), note that we have $\mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \geq \epsilon\varphi$. Hence, after rearranging the terms in (127), we obtain

$$\begin{aligned} \mathbb{E}[h^i(\boldsymbol{\lambda}^{\bar{\pi}})] \geq& -\epsilon\varphi + \mathbb{E}[h^i(\bar{\boldsymbol{\lambda}})] \\ =& -\epsilon\varphi + \epsilon\varphi \\ =& 0. \end{aligned} \tag{128}$$

Hence proved. □

E.2 Proof of Corollary 2

Proof. Recall the result in Eq. (31) and (125), we directly have

$$f(\boldsymbol{\lambda}^*) - \mathbb{E}[f(\boldsymbol{\lambda}^{\bar{\pi}})] \leq 2\epsilon \tag{129}$$

Similarly, combine Eq. (30a) and (127), we have

$$\mathbb{E}[h^i(\boldsymbol{\lambda}^{\bar{\pi}})] \geq -2\epsilon \tag{130}$$

Re-scaling ϵ to $\frac{\epsilon}{2}$ finishes the proof. □

Appendix F. Optimization Theory

Consider the standard optimization problem

$$f_{opt} = \min_{x \in \mathcal{X}} \{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{A}\mathbf{x} + \mathbf{b} = 0\} \tag{131}$$

where $\mathbf{A} \in \mathbb{R}^{d*n}$, $\mathbf{b} \in \mathbb{R}^d$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$. Define the value function as

$$p(\mathbf{u}, \mathbf{t}) = \min_{\mathbf{x} \in \mathcal{X}} \{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq \mathbf{u}, \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{t}\} \tag{132}$$

and the dual function as

$$q(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{x} \in \mathcal{X}} \{f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}) + \mathbf{z}^T(\mathbf{A}\mathbf{x} + \mathbf{b})\}, \mathbf{y} \in \mathbb{R}^m_+, \mathbf{z} \in \mathbb{R}^d \tag{133}$$

Then the dual problem can be written as

$$q_{opt} = \max_{\mathbf{y} \in \mathbb{R}^m_+, \mathbf{z} \in \mathbb{R}^d} q(\mathbf{y}, \mathbf{z}) \tag{134}$$

Lemma 9. *(Theorem 3.59 in (Beck, 2017))* ($\mathbf{y}$,$\mathbf{z}$) *is an optimal solution of problem Eq.* (134) *if and only if* $-(\mathbf{y}, \mathbf{z}) \in \partial p(\mathbf{0}, \mathbf{0})$

Theorem 4. *(Theorem 3.60 in (Beck, 2017)) Let* $f, \mathbf{g}$ *be convex functions,* $\mathcal{X}$ *a nonempty convex set,* $\mathbf{A} \in \mathbb{R}^{d*n}$ *and* $\mathbf{b} \in \mathbb{R}^d$*. Let* f_{opt}*,* q_{opt} *be the optimal values of the primal and dual problems Eq.* (131) *and* (134)*, respectively. Suppose that* $f_{opt} = q_{opt}$ *and that the optimal set of the dual problem is nonempty. Let* $(\mathbf{y}^*, \mathbf{z}^*)$ *be the optimal solution of the dual problem, Assume that* $\tilde{\mathbf{x}} \in \mathcal{X}$ *satisfies*

$$f(\tilde{\mathbf{x}}) - f_{opt} + C_1\|\mathbf{g}(\tilde{\mathbf{x}})_+\|_\infty + C_2\|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_1 \leq \delta \tag{135}$$

where $\delta > 0$ *and* C_1, C_2 *are constants satisfying* $C_1 \geq 2\|\mathbf{y}^*\|_1$*,* $C_2 \geq 2\|\mathbf{z}^*\|_\infty$*, then*

$$\begin{aligned} f(\tilde{\mathbf{x}}) - f_{opt} &\leq \delta \\ \|\mathbf{g}(\tilde{\mathbf{x}})_+\|_\infty &\leq \frac{2\delta}{C_1} \\ \|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_1 &\leq \frac{2\delta}{C_2} \end{aligned} \tag{136}$$

Proof. It is trivial that $f(\tilde{\mathbf{x}}) - f_{opt} \leq \delta$ due to the fact that $C_1\|\mathbf{g}(\tilde{\mathbf{x}})_+\|_\infty$ and $C_2\|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_1$ are both non-negative. Since $(\mathbf{y}^*, \mathbf{z}^*)$ is the optimal solution for the dual problem, it follows by Lemma 9 that $-(\mathbf{y}^*, \mathbf{z}^*) \in (\mathbf{0}, \mathbf{0})$. Therefore, for any $(\mathbf{u}, \mathbf{t}) \in dom(p)$

$$p(\mathbf{u}, \mathbf{t}) - p(\mathbf{0}, \mathbf{0}) \geq \langle -\mathbf{y}^*, \mathbf{u}\rangle + \langle -\mathbf{z}^*, \mathbf{t}\rangle \tag{137}$$

Plugging $\mathbf{u} = \tilde{\mathbf{u}} := [\mathbf{g}(\tilde{\mathbf{x}})]_+$ and $\mathbf{t} = \tilde{\mathbf{t}} := \mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}$ into Eq. (137), while using the inequality $p(\tilde{\mathbf{u}}, \tilde{\mathbf{t}}) \leq f(\tilde{\mathbf{x}})$ and the equality $p(\mathbf{0}, \mathbf{0}) = f_{opt}$, we obatin

$$\begin{aligned} (C_1 - \|\mathbf{y}^*\|_1)\|\tilde{\mathbf{u}}\|_\infty + (C_2 - \|\mathbf{z}^*\|_\infty)\|\tilde{\mathbf{t}}\|_1 &= -\|\mathbf{y}^*\|_1\|\tilde{\mathbf{u}}\|_\infty - \|\mathbf{z}^*\|_\infty\|\tilde{\mathbf{t}}\|_1 + C_1\|\tilde{\mathbf{u}}\|_\infty + C_2\|\tilde{\mathbf{t}}\|_1 \\ &\leq \langle -\mathbf{y}^*, \tilde{\mathbf{u}}\rangle + \left\langle -\mathbf{z}^*, \tilde{\mathbf{t}}\right\rangle + C_1\|\tilde{\mathbf{u}}\|_\infty + C_2\|\tilde{\mathbf{t}}\|_1 \\ &\leq p(\tilde{\mathbf{u}}, \tilde{\mathbf{t}}) - p(\mathbf{0}, \mathbf{0}) + C_1\|\tilde{\mathbf{u}}\|_\infty + C_2\|\tilde{\mathbf{t}}\|_1 \\ &\leq f(\tilde{\mathbf{x}}) - f_{opt} + C_1\|\tilde{\mathbf{u}}\|_\infty + C_2\|\tilde{\mathbf{t}}\|_1 \\ &\leq \delta \end{aligned} \tag{138}$$

It is clear that $C_1 - \|\mathbf{y}^*\|_1$ and $C_2 - \|\mathbf{z}^*\|_\infty$ are both non-negative. Thus,

$$\begin{aligned} (C_1 - \|\mathbf{y}^*\|_1)\|\tilde{\mathbf{u}}\|_\infty &\leq \delta \\ (C_2 - \|\mathbf{z}^*\|_\infty)\|\tilde{\mathbf{t}}\|_1 &\leq \delta \end{aligned} \tag{139}$$

Finally, using the assumption $C_1 \geq 2\|\mathbf{y}^*\|_1$, $C_2 \geq 2\|\mathbf{z}^*\|_\infty$

$$\begin{aligned} \|[g(\tilde{\mathbf{x}})_+]\|_\infty = \|\tilde{\mathbf{u}}\|_\infty &\leq \frac{\delta}{C_1 - \|\mathbf{y}\|_1} \leq \frac{2\delta}{C_1} \\ \|[\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}]\|_1 = \|\tilde{\mathbf{t}}\|_1 &\leq \frac{\delta}{C_2 - \|\mathbf{z}\|_\infty} \leq \frac{2\delta}{C_2} \end{aligned} \tag{140}$$

□

References

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

Agarwal, M., & Aggarwal, V. (2023). Reinforcement learning for joint optimization of multiple rewards. *Journal of Machine Learning Research*, *24*(49), 1–41.

Agarwal, M., Aggarwal, V., & Lan, T. (2022a). Multi-objective reinforcement learning with non-linear scalarization. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 9–17.

Agarwal, M., Bai, Q., & Aggarwal, V. (2022b). Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*.

Agarwal, M., Bai, Q., & Aggarwal, V. (2022c). Regret guarantees for model-based reinforcement learning with long-term average constraints. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Akhtar, Z., Bedi, A. S., & Rajawat, K. (2021). Conservative stochastic optimization with expectation constraints. *IEEE Transactions on Signal Processing*, *69*, 3190–3205.

Al-Abbasi, A. O., Ghosh, A., & Aggarwal, V. (2019). Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, *20*(12), 4714–4727.

Altman, E. (1999). *Constrained Markov decision processes*, Vol. 7. CRC Press.

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

Azar, M. G., Munos, R., & Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, *91*(3), 325–349.

Bai, Q., Agarwal, M., & Aggarwal, V. (2022a). Joint optimization of concave scalarized multi-objective reinforcement learning with policy gradient based algorithm. *Journal of Artificial Intelligence Research*, *74*, 1565–1597.

Bai, Q., Bedi, A. S., Agarwal, M., Koppel, A., & Aggarwal, V. (2022b). Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 3682–3689.

Bai, Q., Bedi, A. S., & Aggarwal, V. (2023). Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, pp. 6737–6744.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Brantley, K., Dudik, M., Lykouris, T., Miryoosefi, S., Simchowitz, M., Slivkins, A., & Sun, W. (2020). Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, *33*, 16315–16326.

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

Chen, J., Umrawal, A. K., Lan, T., & Aggarwal, V. (2021a). Deepfreight: A model-free deep-reinforcement-learning-based algorithm for multi-transfer freight delivery. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 31, pp. 510–518.

Chen, Y., Dong, J., & Wang, Z. (2021b). A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*.

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

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

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

Gattami, A., Bai, Q., & Aggarwal, V. (2021). Reinforcement learning for constrained markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 2656–2664. PMLR.

Geng, N., Lan, T., Aggarwal, V., Yang, Y., & Xu, M. (2020). A multi-agent reinforcement learning perspective on distributed traffic engineering. In *2020 IEEE 28th International Conference on Network Protocols (ICNP)*, pp. 1–11. IEEE.

Hazan, E., Kakade, S., Singh, K., & Van Soest, A. (2019). Provably efficient maximum entropy exploration. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 2681–2691. PMLR.

He, J., Zhou, D., & Gu, Q. (2021). Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, *34*, 22288–22300.

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

Jin, C., Allen-Zhu, Z., Bubeck, S., & Jordan, M. I. (2018). Is q-learning provably efficient?. In *Advances in Neural Information Processing Systems*, Vol. 31.

Juditsky, A., Nemirovski, A., & Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, *1*(1), 17–58.

Kalagarla, K. C., Jain, R., & 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., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*.

Kostrikov, I., Nachum, O., & Tompson, J. (2019). Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*.

Lattimore, T., & Hutter, M. (2012). Pac bounds for discounted mdps. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, p. 320–334, Berlin, Heidelberg. Springer-Verlag.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., & Tian, C. (2021a). Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, *34*, 17183–17193.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., & Tian, C. (2021b). Fast global convergence of policy optimization for constrained mdps. *arXiv preprint arXiv:2111.00552*.

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

Margolies, R., Sridharan, A., Aggarwal, V., Jana, R., Shankaranarayanan, N., Vaishampayan, V. A., & Zussman, G. (2014). Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms. *IEEE/ACM Transactions on Networking*, *24*(1), 355–367.

Mihatsch, O., & Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine learning*, *49*(2), 267–290.

Moldovan, T. M., & Abbeel, P. (2012). Safe exploration in markov decision processes. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1451–1458.

Mondal, W. U., & Aggarwal, V. (2023). Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes. *arXiv preprint arXiv:2310.11677*.

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

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

Shalev-Shwartz, S., et al. (2011). Online learning and online convex optimization. *Foundations and trends in Machine Learning*, *4*(2), 107–194.

Tessler, C., Mankowitz, D. J., & Mannor, S. (2018). Reward constrained policy optimization. In *International Conference on Learning Representations*.

Vaswani, S., Yang, L., & Szepesvári, C. (2022). Near-optimal sample complexity bounds for constrained mdps. *Advances in Neural Information Processing Systems*, *35*, 3110–3122.

Vu, T. L., Mukherjee, S., Yin, T., Huang, R., Tan, J., & Huang, Q. (2021). Safe reinforcement learning for emergency load shedding of power systems. In *2021 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5. IEEE.

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., & 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., & 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)*, pp. 1–7. IEEE.

Xiang, Y., Lan, T., Aggarwal, V., & 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., & Lan, G. (2021). Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pp. 11480–11491. PMLR.

Ying, D., Guo, M. A., Ding, Y., Lavaei, J., & Shen, Z.-J. (2023). Policy-based primal-dual methods for convex constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, pp. 10963–10971.

Zhang, J., Bedi, A. S., Wang, M., & 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.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., & Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 4572–4583.

Zhang, J., Ni, C., Szepesvari, C., Wang, M., et al. (2021). On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, *34*, 2228–2240.